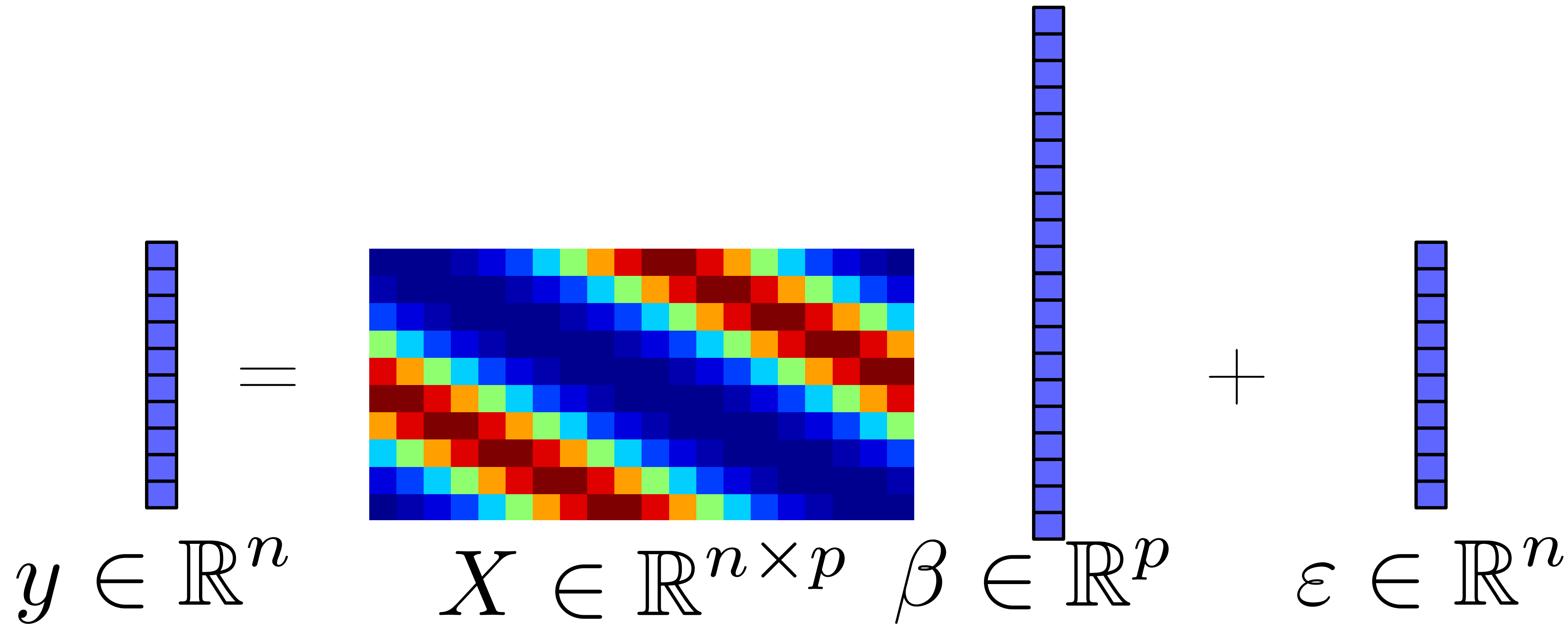
Bayesian Inference

Jalal Fadili
Normandie Université-ENSICAEN, GREYC

Mathematical coffees 2017

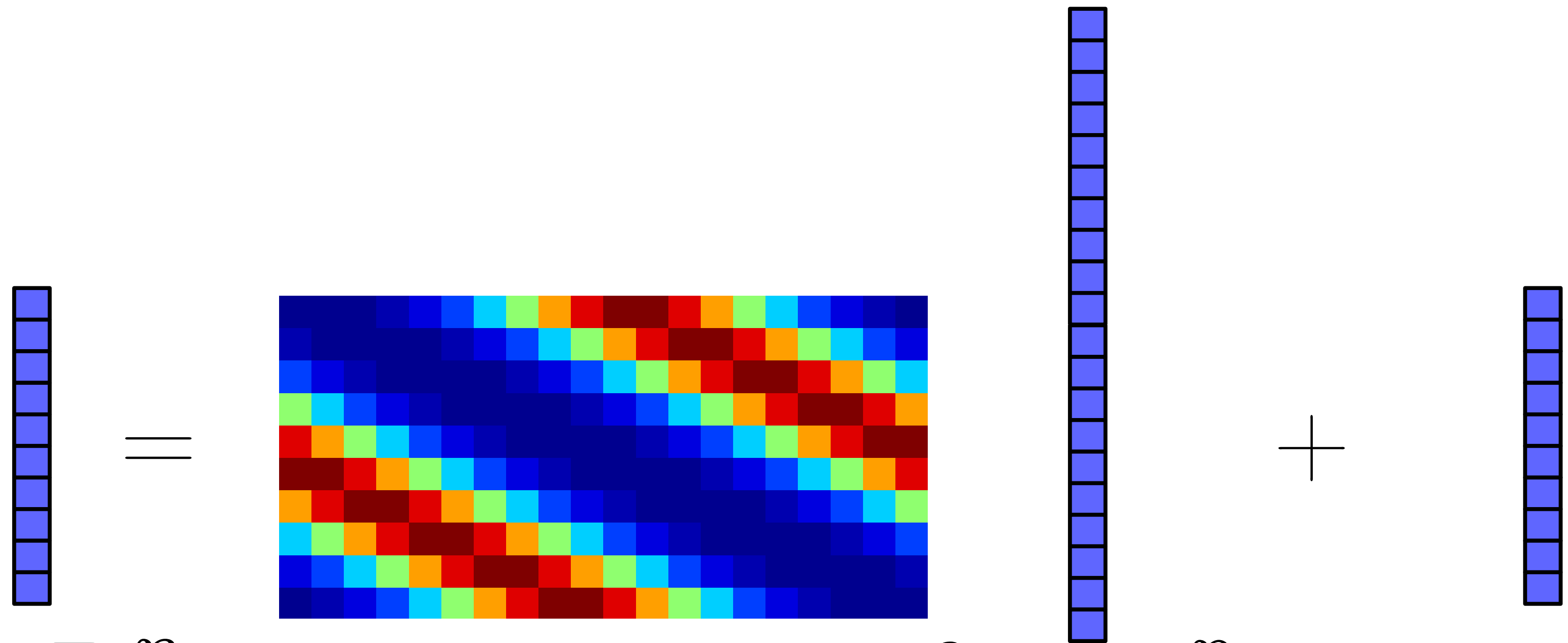
Normandie Université
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

Linear regression



- X is the design matrix (i.e. its columns are the predictors) :
- β unknown regression vector : Has some **prior** structure.
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$ independent of β.
- Several prior MCs on this model.
- In the Bayesian paradigm : random model on β.

Linear regression



$y \in \mathbb{R}^n$ $\quad X \in \mathbb{R}^{n \times p}$ $\quad \beta \in \mathbb{R}^p$ $\quad \varepsilon \in \mathbb{R}^n$

Inference : estimation and testing of (β, σ^2) from data (X, y)

- X is the design matrix (i.e. its columns are the predictors) :
- β unknown regression vector : Has some **prior** structure.
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$ independent of β.
- Several prior MCs on this model.
- In the Bayesian paradigm : random model on β.

Likelihood function

$$y = X\beta + \varepsilon$$ For simplicity, $\operatorname{rank}(X) = p$.

- Additive White Gaussian noise :

$$Y|(\beta, \sigma^2) \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{I}_n)$$

$$\|z\|^2 = z^\top z$$
$$\beta_{\text{ols}} = X^+ y$$
$$y_{\text{ols}} = X\beta_{\text{ols}}$$
$$= \operatorname{Proj}_{\operatorname{Im}(X)} y$$

$$\begin{aligned} p(y|\beta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y - X\beta\|^2}{2\sigma^2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y - y_{\text{ols}}\|^2 + \|X(\beta - \beta_{\text{ols}})\|^2}{2\sigma^2} \right) \\ &= \phi(y; y_{\text{ols}}, \sigma^2)\phi(\beta; \beta_{\text{ols}}, \sigma^2(X^\top X)^{-1}). \end{aligned}$$

Likelihood function

$$y = X\beta + \varepsilon$$ For simplicity, $\mathrm{rank}(X) = p$.

- Additive White Gaussian noise :

$$Y|(\beta, \sigma^2) \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{I}_n)$$

$$\|z\|^2 = z^\top z$$
$$\beta_{\mathrm{ols}} = X^+ y$$
$$y_{\mathrm{ols}} = X\beta_{\mathrm{ols}} = \mathrm{Proj}_{\mathrm{Im}(X)}\, y$$

$$\begin{aligned} p(y|\beta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y - X\beta\|^2}{2\sigma^2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|y - y_{\mathrm{ols}}\|^2 + \|X(\beta - \beta_{\mathrm{ols}})\|^2}{2\sigma^2} \right) \\ &= \phi(y; y_{\mathrm{ols}}, \sigma^2)\phi(\beta; \beta_{\mathrm{ols}}, \sigma^2 (X^\top X)^{-1}). \end{aligned}$$

- $(\beta_{\mathrm{ols}}, y - y_{\mathrm{ols}})$ are jointly sufficient statistics for (β, σ^2).
- Moreover,

$$\beta_{\mathrm{ols}}|(\beta, \sigma^2) \sim \mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1})$$

independent of $\|y - y_{\mathrm{ols}}\|^2 \, |\sigma^2 \sim \sigma^2 \chi^2_{n-p}$.

Jeffrey's prior

$$y = X\beta + \varepsilon$$ For simplicity, $\mathrm{rank}(X) = p$.

- Take the (Jeffrey's) prior :

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- The joint posterior is

$$\beta_{\text{ols}} = X^{+}y$$
$$y_{\text{ols}} = X\beta_{\text{ols}} = \mathrm{Proj}_{\mathrm{Im}(X)}\, y$$

$$\begin{aligned} p(\beta, \sigma^2|y) &= \frac{p(y|\beta, \sigma^2)\pi(\beta, \sigma^2)}{p(y)} \\ &= \frac{\phi(y; y_{\text{ols}}, \sigma^2)\pi(\beta, \sigma^2)}{p(y)}\phi(\beta; \beta_{\text{ols}}, \sigma^2(X^\top X)^{-1}) \\ &= p(\sigma^2|\, \|y - y_{\text{ols}}\|^2)\phi(\beta; \beta_{\text{ols}}, \sigma^2(X^\top X)^{-1}). \end{aligned}$$

Jeffrey's prior

$$y = X\beta + \varepsilon$$ For simplicity, $\mathrm{rank}(X) = p$.

- Take the (Jeffrey's) prior :

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- The joint posterior is

$$\beta_{\mathrm{ols}} = X^{+}y$$
$$y_{\mathrm{ols}} = X\beta_{\mathrm{ols}}$$

$$p(\beta, \sigma^2 | y) = p(\sigma^2 | \, \|y - y_{\mathrm{ols}}\|^2)\phi(\beta; \beta_{\mathrm{ols}}, \sigma^2(X^\top X)^{-1}). \quad = \mathrm{Proj}_{\mathrm{Im}(X)}\, y$$

- Integrating out β, the marginal posterior of σ^2 is $\propto$ inverse-χ^2_{n-p}, i.e.

$$\begin{aligned} p(\sigma^2|y) &= p(\sigma^2|\widehat{r}^2) \qquad (\widehat{r}^2 \stackrel{\mathrm{def}}{=} \|y - y_{\mathrm{ols}}\|^2) \\ &\propto p(\widehat{r}^2|\sigma^2)\pi(\sigma^2) \propto \frac{1}{\sigma^2} p_{\sigma^2\chi^2_{n-p}}(\widehat{r}^2) \\ &\propto \frac{1}{\sigma^4} p_{\chi^2_{n-p}}(\widehat{r}^2/\sigma^2) \propto \frac{\widehat{r}^4}{\sigma^4} p_{1/\chi^2_{n-p}}(\sigma^2/\widehat{r}^2). \end{aligned}$$

Jeffrey's prior

$$y = X\beta + \varepsilon$$ For simplicity, $\mathrm{rank}(X) = p$.

- Take the (Jeffrey's) prior :

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- The joint posterior is

$$\beta_{\text{ols}} = X^{+}y$$
$$y_{\text{ols}} = X\beta_{\text{ols}} = \mathrm{Proj}_{\mathrm{Im}(X)}\, y$$

$$p(\beta, \sigma^2 | y) = p(\sigma^2 |\, \|y - y_{\text{ols}}\|^2)\phi(\beta; \beta_{\text{ols}}, \sigma^2 (X^\top X)^{-1}).$$

- Integrating out σ^2, the marginal posterior of β is multivariate t_{n-p}, i.e.

$$p(\beta|y) = \frac{\Gamma(n/2)\det(X^\top X)^{1/2}\widehat{\sigma}^{-p}}{\pi^{p/2}\Gamma((n-p)/2)(n-p)^{p/2}} \left(1 + \frac{\|X(\beta - \beta_{\text{ols}})\|^2}{(n-p)\widehat{\sigma}^2}\right)^{-n/2}$$

$\widehat{\sigma}^2 \stackrel{\text{def}}{=} \|y - y_{\text{ols}}\|^2 / (n-p)$ (unbiased estimator of the variance).

Jeffrey's prior

$$y = X\beta + \varepsilon$$ For simplicity, $\mathrm{rank}(X) = p$.

- Take the (Jeffrey's) prior :

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- The joint posterior is

$\beta_{\mathrm{ols}} = X^+ y$

$y_{\mathrm{ols}} = X\beta_{\mathrm{ols}}$

$$p(\beta, \sigma^2 | y) = p(\sigma^2 | \left\| y - y_{\mathrm{ols}} \right\|^2) \phi(\beta; \beta_{\mathrm{ols}}, \sigma^2 (X^\top X)^{-1}).$$

$= \mathrm{Proj}_{\mathrm{Im}(X)}\, y$

- Integrating out σ^2, the marginal posterior of β is multivariate t_{n-p}, i.e.

$$p(\beta | y) = \frac{\Gamma(n/2) \det(X^\top X)^{1/2} \widehat{\sigma}^{-p}}{\pi^{p/2} \Gamma((n-p)/2)(n-p)^{p/2}} \left(1 + \frac{\left\| X(\beta - \beta_{\mathrm{ols}}) \right\|^2}{(n-p)\widehat{\sigma}^2} \right)^{-n/2}$$

$\widehat{\sigma}^2 \overset{\mathrm{def}}{=} \left\| y - y_{\mathrm{ols}} \right\|^2 / (n-p)$ (unbiased estimator of the variance).

- If $n \geq p + 2$, the posterior mean of β is β_{ols}, i.e. the MMSE is

$$\mathbb{E}\left[\beta | y\right] = \beta_{\mathrm{ols}}.$$

- The posterior mode is the same.

Gaussian prior: Wiener filter

$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \Sigma_e)$, $\Sigma_e \succ 0$.
- $\beta \sim \mathcal{N}(0, \Sigma_b)$, $\Sigma_b \succ 0$.
- ε and β uncorrelated (hence independent by normality).
- Σ_e and Σ_b fixed and known.

Gaussian prior: Wiener filter

$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \Sigma_e)$, $\Sigma_e \succ 0$.
- $\beta \sim \mathcal{N}(0, \Sigma_b)$, $\Sigma_b \succ 0$.
- ε and β uncorrelated (hence independent by normality).
- Σ_e and Σ_b fixed and known.

$$p(y|\beta, \Sigma_e) = \phi(y; X\beta, \Sigma_e) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_e)}} \exp\left(-\frac{\|y - X\beta\|^2_{\Sigma_e^{-1}}}{2}\right) \qquad \|z\|^2_{\mathrm{A}} = z^\top \mathrm{A} z$$

$$\pi(\beta|\Sigma_b) = \phi(\beta; 0, \Sigma_b) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^p \det(\Sigma_b)}} \exp\left(-\frac{\|\beta\|^2_{\Sigma_b^{-1}}}{2}\right)$$

Gaussian prior: Wiener filter

$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \Sigma_e)$, $\Sigma_e \succ 0$.
- $\beta \sim \mathcal{N}(0, \Sigma_b)$, $\Sigma_b \succ 0$.
- ε and β uncorrelated (hence independent by normality).
- Σ_e and Σ_b fixed and known.

$$p(y|\beta, \Sigma_e) = \phi(y; X\beta, \Sigma_e) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_e)}} \exp\left(-\frac{\|y - X\beta\|^2_{\Sigma_e^{-1}}}{2}\right) \qquad \|z\|^2_{\mathrm{A}} = z^\top \mathrm{A} z$$

$$\pi(\beta|\Sigma_b) = \phi(\beta; 0, \Sigma_b) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^p \det(\Sigma_b)}} \exp\left(-\frac{\|\beta\|^2_{\Sigma_b^{-1}}}{2}\right)$$

- The posterior of β is

$$p(\beta|y) \propto \phi(y; X\beta, \Sigma_e)\phi(\beta; 0, \Sigma_b).$$

Gaussian prior: Wiener filter

$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \Sigma_e)$, $\Sigma_e \succ 0$.
- $\beta \sim \mathcal{N}(0, \Sigma_b)$, $\Sigma_b \succ 0$.
- ε and β uncorrelated (hence independent by normality).
- Σ_e and Σ_b fixed and known.

$$p(y|\beta, \Sigma_e) = \phi(y; X\beta, \Sigma_e) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma_e)}} \exp\left(-\frac{\|y - X\beta\|^2_{\Sigma_e^{-1}}}{2}\right) \qquad \|z\|^2_{\mathrm{A}} = z^\top \mathrm{A} z$$

$$\pi(\beta|\Sigma_b) = \phi(\beta; 0, \Sigma_b) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^p \det(\Sigma_b)}} \exp\left(-\frac{\|\beta\|^2_{\Sigma_b^{-1}}}{2}\right)$$

- The posterior of β is

$$p(\beta|y) \propto \phi(y; X\beta, \Sigma_e)\phi(\beta; 0, \Sigma_b).$$

Proposition *Under the above Bayesian setting, The MAP and MMSE are given by*

$$\left(X^\top \Sigma_e^{-1} X + \Sigma_b^{-1}\right)^{-1} X^\top \Sigma_e^{-1} y = \left(\mathrm{I}_p - \Sigma_b X^\top (\Sigma_e + X \Sigma_b X^\top)^{-1} X\right) \Sigma_b X^\top \Sigma_e^{-1} y.$$

This also coincides with the Wiener estimator, i.e. the best linear estimator minimizing the quadratic risk.

Gaussian prior: diagonal estimation

$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \Sigma_e)$, $\Sigma_e \succ 0$.
- $\beta \sim \mathcal{N}(0, \Sigma_b)$, $\Sigma_b \succ 0$.
- ε and β uncorrelated (hence independent by normality).
- Σ_e and Σ_b fixed and known.

Proposition *Suppose also that X is circular convolution by a kernel h, and ε and β are wide-sense stationary zero-mean Gaussian vectors. Then, the MAP, the MMSE and the Wiener estimator of β are given by the following expression, which is coordinate-wise separable in the DFT domain :*

$$\frac{\mathcal{F}(h)_i^*}{|\mathcal{F}(h)_i|^2 + \frac{(\sigma_e^2)_i}{(\sigma_b^2)_i}} \mathcal{F}(y)_i,$$

where $\mathcal{F}$ is the DFT operator, and σ_e^2 and σ_b^2 are the vectors of eigenvalues of Σ_e and Σ_b respectively.

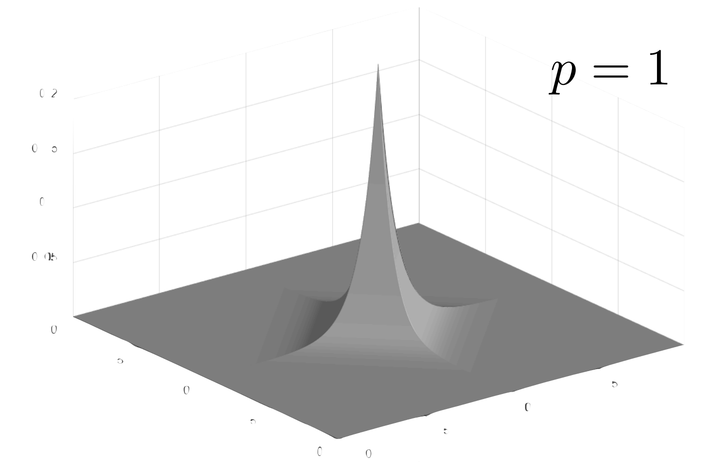
Generalized Gaussian prior

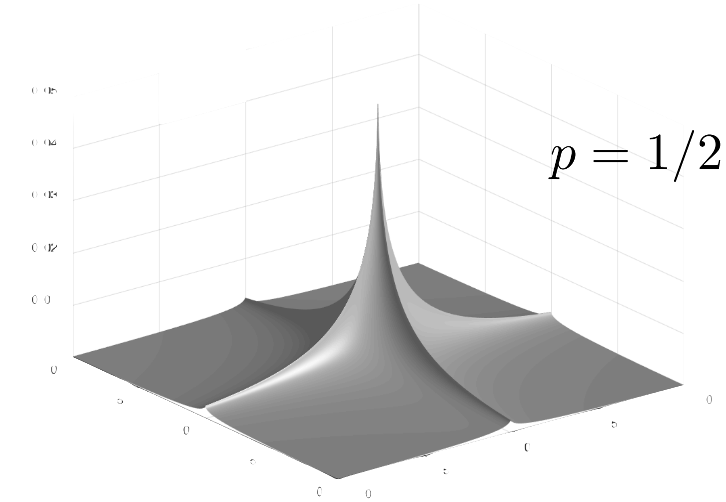
$$y = X\beta + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$.
- $\beta_i \sim_{iid} \mathrm{GGD}(p, \lambda)$, $\lambda > 0$, $p > 0$.

$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|y - X\beta\|^2}{2\sigma^2}}$$

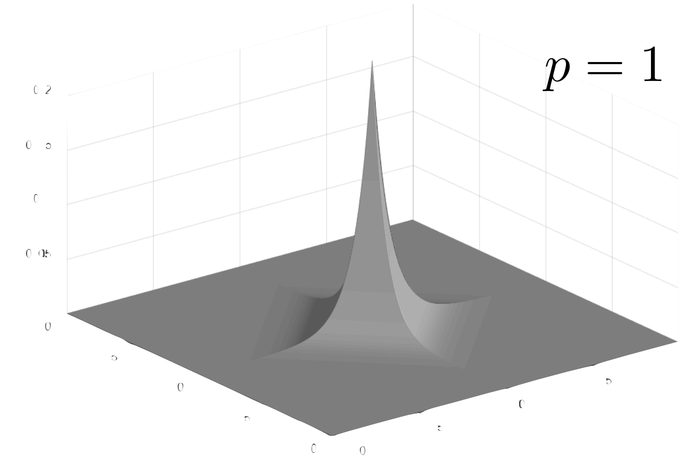
$$\pi(\beta|p, \lambda) = \prod_{i=1}^{p} \frac{p\sqrt[p]{\lambda}}{2\Gamma(1/p)} e^{-\lambda|\beta_i|^p}.$$

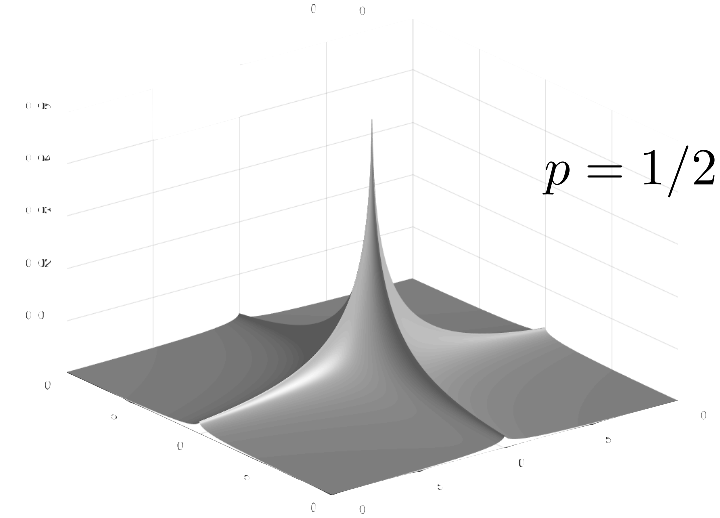




Generalized Gaussian prior

$$y = X\beta + \varepsilon$$





- $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$.
- $\beta_i \sim_{iid} \mathrm{GGD}(p, \lambda)$, $\lambda > 0$, $p > 0$.

$$p(y|\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|y - X\beta\|^2}{2\sigma^2}}$$

$$\pi(\beta|p, \lambda) = \prod_{i=1}^{p} \frac{p\sqrt[p]{\lambda}}{2\Gamma(1/p)} e^{-\lambda|\beta_i|^p}.$$

- Hyperparameters (σ, p, λ) known, the MAP reads :

$$\underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}} \; \frac{1}{2\sigma^2} \|y - X\beta\|^2 + \lambda \|\beta\|_p^p . \qquad \|\beta\|_p^p = \sum_{i=1} |\beta_i|^p$$

  - For $p = 1$, we recover the Lasso (see several previous MCs).
  - For X unitary, the MAP corresponds to computing $\mathrm{prox}_{\lambda\sigma^2|\cdot|^p}(y_i)$, which has a closed form or can be computed efficiently.
- Except for $p = 2$, the MMSE does not have a closed-form even when X is unitary.

Is any PMLE a MAP ?

MAP $$\operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \; -\log p(y|\beta, \theta_e) - \log \pi(\beta)$$

PMLE $$\operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \; -\log p(y|\beta, \theta_e) + \psi(\beta)$$

Is any PMLE a MAP ?

MAP $\quad \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \; -\log p(y|\beta, \theta_e) - \log \pi(\beta)$

PMLE $\quad \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \; -\log p(y|\beta, \theta_e) + \psi(\beta)$

Is any PMLE a MAP ?

Is any PMLE a MAP ?

MAP $$\underset{\beta \in \mathbb{R}^p}{\operatorname{Argmin}} \; -\log p(y|\beta, \theta_e) - \log \pi(\beta)$$

PMLE $$\underset{\beta \in \mathbb{R}^p}{\operatorname{Argmin}} \; -\log p(y|\beta, \theta_e) + \psi(\beta)$$

Is any PMLE a MAP ?

- PMLE with penalty ψ is MAP with prior density $\exp(-\psi(\beta))/Z$ if β is assumed Gibbsian.
- But this is only one possible Bayesian interpretation.
- There are other possible Bayesian interpretations.

Is any PMLE a MAP ?

$$y = \beta + \varepsilon \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MAP}} \in \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \tfrac{1}{2\sigma^2} \left\| y - \beta \right\|^2 - \log \pi(\beta)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MMSE}} = \mathbb{E}\left[\beta | y\right]$$

$$\widehat{\beta}^{\psi}_{\mathrm{PMLE}} \in \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \tfrac{1}{2\sigma^2} \left\| y - \beta \right\|^2 + \psi(\beta)$$

Is any PMLE a MAP ?

$$y = \beta + \varepsilon \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MAP}} \in \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \tfrac{1}{2\sigma^2} \|y - \beta\|^2 - \log \pi(\beta)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MMSE}} = \mathbb{E}\left[\beta | y\right]$$

$$\widehat{\beta}^{\psi}_{\mathrm{PMLE}} \in \operatorname*{Argmin}_{\beta \in \mathbb{R}^p} \tfrac{1}{2\sigma^2} \|y - \beta\|^2 + \psi(\beta)$$

$\exists \psi \quad \text{s.t.} \quad \widehat{\beta}^{\psi}_{\mathrm{PMLE}} = \widehat{\beta}^{\pi}_{\mathrm{MMSE}}$ for some $\pi(\beta) \neq \exp(-\psi(\beta))/Z$ in general.

Is any PMLE a MAP ?

$$y = \beta + \varepsilon \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MAP}} \in \underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}} \ \frac{1}{2\sigma^2} \left\| y - \beta \right\|^2 - \log \pi(\beta)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MMSE}} = \mathbb{E}\left[\beta | y\right]$$

$$\widehat{\beta}^{\psi}_{\mathrm{PMLE}} \in \underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}} \ \frac{1}{2\sigma^2} \left\| y - \beta \right\|^2 + \psi(\beta)$$

$\exists \psi \quad \text{s.t.} \quad \widehat{\beta}^{\psi}_{\mathrm{PMLE}} = \widehat{\beta}^{\pi}_{\mathrm{MMSE}}$ for some $\pi(\beta) \neq \exp(-\psi(\beta))/Z$ in general.

$\forall \pi, \ \exists \psi \quad \text{s.t.} \quad \widehat{\beta}^{\pi}_{\mathrm{MMSE}} = \widehat{\beta}^{\nu}_{\mathrm{MAP}}, \ \nu(\beta) = \exp(-\psi(\beta))/Z.$

Is any PMLE a MAP ?

$$y = \beta + \varepsilon \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MAP}} \in \underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}} \ \frac{1}{2\sigma^2} \left\| y - \beta \right\|^2 - \log \pi(\beta)$$

$$\widehat{\beta}^{\pi}_{\mathrm{MMSE}} = \mathbb{E}\left[\beta | y\right]$$

$$\widehat{\beta}^{\psi}_{\mathrm{PMLE}} \in \underset{\beta \in \mathbb{R}^p}{\mathrm{Argmin}} \ \frac{1}{2\sigma^2} \left\| y - \beta \right\|^2 + \psi(\beta)$$

- Actually, the Bayesian interpretation may lead to an apparent paradox as in, e.g., Lasso :
  - The Laplacian prior is not heavy-tailed, hence not a wise prior to promote sparsity.
  - Yet we have strong theoretical guarantees that Lasso has excellent performance to recover sparse vectors (reason lies in blessings of high-dimensional geometry as seen in the last MC).
- A variety of Bayesian priors promoting sparsity have been developed in the sparse representation literature, though they are not log-concave and enjoy guarantees only for specific settings.

GLM

- n independent observations $y_i \sim \mathcal{B}(k_i, p_i)$.
- $p_i = h(X^i\beta)$, $h : \mathbb{R} \to [0,1]$ is the link function (a cdf). X^i : $i-$th row of X
  - Logit : logistic cdf $h(t) = \frac{1}{1+e^t}$.
  - Probit : standard normal cdf $h = \Phi$.

GLM

- n independent observations $y_i \sim \mathcal{B}(k_i, p_i)$.
- $p_i = h(X^i\beta)$, $h : \mathbb{R} \to [0,1]$ is the link function (a cdf). X^i : $i-$th row of X
  - Logit : logistic cdf $h(t) = \frac{1}{1+e^t}$.
  - Probit : standard normal cdf $h = \Phi$.
- Estimate β from y.

GLM

- n independent observations $y_i \sim \mathcal{B}(k_i, p_i)$.
- $p_i = h(X^i\beta)$, $h : \mathbb{R} \to [0,1]$ is the link function (a cdf). X^i : $i-$th row of X
  - Logit : logistic cdf $h(t) = \frac{1}{1+e^t}$.
  - Probit : standard normal cdf $h = \Phi$.
- Estimate β from y.
- Likelihood :

$$p(y|\beta) = \prod_{i=1}^{n} h(X^i\beta)^{y_i}(1 - h(X^i\beta))^{k_i - y_i}.$$

- Posterior of β :

$$p(\beta|y) \propto \prod_{i=1}^{n} h(X^i\beta)^{y_i}(1 - h(X^i\beta))^{k_i - y_i}\pi(\beta).$$

- Largely intractable : no closed form even with a flat prior.

Logistic regression

- n independent observations $y_i \sim \mathcal{B}(k_i, h(X^i\beta))$.

  X^i : $i-$th row of X
- Logit : $h(t) = \frac{1}{1+e^t}$, hence $X^i\beta = -\log(p_i/(1-p_i))$.
- The likelihood is

$$p(y|\beta) = e^{-\left(\sum_{i=1}^n y_i X^i\right)\beta} \prod_{i=1}^n (1+\exp(-X^i\beta))^{-k_i}.$$

Logistic regression

- n independent observations $y_i \sim \mathcal{B}(k_i, h(X^i\beta))$.

  X^i : $i-$th row of X

- Logit : $h(t) = \frac{1}{1+e^t}$, hence $X^i\beta = -\log(p_i/(1-p_i))$.
- The likelihood is

$$p(y|\beta) = e^{-\left(\sum_{i=1}^{n} y_i X^i\right)\beta} \prod_{i=1}^{n} (1 + \exp(-X^i\beta))^{-k_i}.$$

- Posterior largely intractable.
- But $X^i\beta = \log(p_i/(1-p_i)) \Rightarrow$ large k_i normal approximation to the binomial.

Logistic regression

- n independent observations $y_i \sim \mathcal{B}(k_i, h(X^i\beta))$. X^i : $i-$th row of X
- Logit : $h(t) = \frac{1}{1+e^t}$, hence $X^i\beta = -\log(p_i/(1-p_i))$.
- The likelihood is

$$p(y|\beta) = e^{-\left(\sum_{i=1}^n y_i X^i\right)\beta} \prod_{i=1}^n (1+\exp(-X^i\beta))^{-k_i}.$$

- Posterior largely intractable.
- But $X^i\beta = \log(p_i/(1-p_i) \Rightarrow$ large k_i normal approximation to the binomial.
- $\widehat{p}_i \stackrel{\text{def}}{=} y_i/k_i$ are independent and $\widehat{p}_i \underset{d}{\rightarrow} \mathcal{N}(p_i, p_i(1-p_i)/k_i)$.
- By the Delta theorem, $(\widehat{\theta}_i - \theta_i)\sqrt{k_i\widehat{p}_i(1-\widehat{p}_i)}$ are independent and

$$(\widehat{\theta}_i - \theta_i)\sqrt{k_i\widehat{p}_i(1-\widehat{p}_i)} \underset{d}{\rightarrow} \mathcal{N}(0,1)$$

$\theta_i \stackrel{\text{def}}{=} -\log(p_i/(1-p_i))$

$\widehat{\theta}_i \stackrel{\text{def}}{=} -\log(\widehat{p}_i/(1-\widehat{p}_i))$

Logistic regression

- n independent observations $y_i \sim \mathcal{B}(k_i, h(X^i\beta))$. X^i : $i-$th row of X
- Logit : $h(t) = \frac{1}{1+e^t}$, hence $X^i\beta = -\log(p_i/(1-p_i))$.
- The likelihood is

$$p(y|\beta) = e^{-\left(\sum_{i=1}^n y_i X^i\right)\beta} \prod_{i=1}^n (1 + \exp(-X^i\beta))^{-k_i}.$$

- Posterior largely intractable.
- But $X^i\beta = \log(p_i/(1-p_i) \Rightarrow$ large k_i normal approximation to the binomial.
- $\widehat{p}_i \stackrel{\text{def}}{=} y_i/k_i$ are independent and $\widehat{p}_i \underset{d}{\rightarrow} \mathcal{N}(p_i, p_i(1-p_i)/k_i)$.
- By the Delta theorem, $(\widehat{\theta}_i - \theta_i)\sqrt{k_i\widehat{p}_i(1-\widehat{p}_i)}$ are independent and

$$(\widehat{\theta}_i - \theta_i)\sqrt{k_i\widehat{p}_i(1-\widehat{p}_i)} \underset{d}{\rightarrow} \mathcal{N}(0,1)$$

$\theta_i \stackrel{\text{def}}{=} -\log(p_i/(1-p_i))$

$\widehat{\theta}_i \stackrel{\text{def}}{=} -\log(\widehat{p}_i/(1-\widehat{p}_i))$

- Approximate large sample likelihood is a weighted least-square

$$p(y|\beta) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\sum_{i=1}^n \sqrt{k_i\widehat{p}_i(1-\widehat{p}_i)}(\widehat{p}_i - X^i\beta)^2}{2}}.$$

- Back to Gaussian (weighted) linear regression.

Probit model

- n independent observations $y_i \sim \mathcal{B}(k_i, h(X^i\beta))$. X^i : $i-$th row of X
- Logit : $h = \Phi$, standard normal cdf.
- The posterior of β

$$p(\beta|y) \propto \prod_{i=1}^{n} \Phi(X^i\beta)^{y_i}(1 - \Phi(X^i\beta))^{k_i - y_i}\pi(\beta).$$

- The likelihood and posterior even less tractable that for the logistic.
- One can also use the Delta theorem to get a normal approximation, though less precise that for the logistic.
- Otherwise MC sampling through latent variables.

Bayesian computations

- Bayesian inference requires computation of moments (e.g. mean, variance), modes and quantiles (e.g. medians) of the posterior distribution.
- MAP :
  - Involves an solving an optimization problem.
  - Closed-form : for some (interesting cases).
- MMSE :
  - Involves an integration problem.
  - Closed-form : rather an exception than a rule.
  - Analytical approximations (Laplace, saddlepoint, etc) : requires smoothness.
  - Numerical quadrature : unrealistic in high-dimensional settings.
  - Monte-Carlo methods.

MAP

$$\underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\ -\log p(y|\beta,\theta_e) - \log \pi(\beta)$$

- A structured composite optimization problem.
- A whole area in its own :
  - The key is to exploit the properties of each term individually and separately.
  - A rich literature including proximal splitting for large-scale data.
  - Previous MCs on the subject.

MAP

$$\underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \; -\log p(y|\beta, \theta_e) - \log \pi(\beta)$$

- A structured composite optimization problem.
- A whole area in its own :
  - The key is to exploit the properties of each term individually and separately.
  - A rich literature including proximal splitting for large-scale data.
  - Previous MCs on the subject.

Example (Linear regression with GGD prior)

- *Hyperparameters* (σ, p, λ) *known, the MAP reads :*

$$\underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \; \frac{1}{2\sigma^2} \left\| y - X\beta \right\|^2 + \lambda \left\| \beta \right\|_p^p, \quad p \geq 0$$

- *Forward-Backward splitting :*

$$\beta_{k+1} \in \text{prox}_{\lambda\sigma^2\gamma\|\cdot\|_p^p} \left(\beta_k + \gamma X^\top (y - X\beta_k)\right), \gamma \in]0, 1/\left\|X\right\|^2].$$

- *Convergence guarantees :*
  - $p \geq 1$ *: to a global minimizer (*γ *even to* $< 2/\left\|X\right\|^2$*.*
  - $p \in [0, 1[$ *: in general to a critical point (*o*-minimal geometry arguments), and a global minimizer if started sufficiently close to it.*

Laplace approximation

- Goal is to compute

$$\mathbb{E}\left[g(\beta)|y\right] = \frac{\int_{\mathbb{R}^p} g(\beta)p(y|\beta)\pi(\beta)d\beta}{\int_{\mathbb{R}^p} p(y|\beta)\pi(\beta)d\beta}$$

  where g, p and π are smooth enough functions of β.

- Approximately evaluate the following integral for large n

$$I \stackrel{\text{def}}{=} \int_{\mathbb{R}^p} q(\beta)\exp(-nh(\beta))d\beta,$$

  h and q are smooth enough around $\widehat{\beta}$, the unique minimizer of h at $\widehat{\beta}$.

Laplace approximation

- Goal is to compute

$$\mathbb{E}\left[g(\beta)|y\right] = \frac{\int_{\mathbb{R}^p} g(\beta)p(y|\beta)\pi(\beta)d\beta}{\int_{\mathbb{R}^p} p(y|\beta)\pi(\beta)d\beta}$$

  where g, p and π are smooth enough functions of β.

- Approximately evaluate the following integral for large n

$$I \stackrel{\text{def}}{=} \int_{\mathbb{R}^p} q(\beta)\exp(-nh(\beta))d\beta,$$

  h and q are smooth enough around $\widehat{\beta}$, the unique minimizer of h at $\widehat{\beta}$.

- The Laplace method involves a Taylor expansion of q and h around $\widehat{\beta}$:

$$I = \exp(-nh(\widehat{\beta}))\int_{\mathbb{R}^p}(q(\widehat{\beta}) + (\beta-\widehat{\beta})\nabla q(\widehat{\beta}) + \tfrac{1}{2}(\beta-\widehat{\beta})^\top\nabla^2 q(\widehat{\beta})(\beta-\widehat{\beta}) + \cdots)e^{-\frac{n(\beta-\widehat{\beta})^\top\nabla^2 h(\widehat{\beta})(\beta-\widehat{\beta})}{2}}d\beta$$

$$= \exp(-nh(\widehat{\beta}))\left((2\pi)^{p/2}n^{-p/2}\det(\nabla^2 h(\widehat{\beta}))^{-1/2}\right)\left(q(\widehat{\beta}) + O(n^{-1})\right)$$

- Apply to the numerator (resp. denominator) of $\mathbb{E}\left[g(\beta)|y\right]$, with $q = g$ (resp. $q = 1$) and $h(\beta) = -\log(p(y|\beta)) - \log(\pi(\beta))$:

$$\mathbb{E}\left[g(\beta)|y\right] = g(\widehat{\beta})\left(1 + O(n^{-1})\right).$$

- MMSE necessitates to solve the MAP supposed to be unique.

Tierney-Kanade approximation

- Goal is to compute
$$\mathbb{E}\left[g(\beta)|y\right] = \frac{\int_{\mathbb{R}^p} g(\beta)p(y|\beta)\pi(\beta)d\beta}{\int_{\mathbb{R}^p} p(y|\beta)\pi(\beta)d\beta}$$
where g, p and π are smooth enough functions of β, g positive.
$$\begin{aligned} I &\stackrel{\text{def}}{=} \int_{\mathbb{R}^p} q(\beta)\exp(-nh(\beta))d\beta \\ &= \exp(-nh(\widehat{\beta}))\left((2\pi)^{p/2}n^{-p/2}\det(\nabla^2 h(\widehat{\beta}))^{-1/2}\right)\left(q(\widehat{\beta}) + O(n^{-1})\right) \end{aligned}$$

- Suppose $\widehat{\beta}$ is the unique minimizer of $nh(\beta) = -\log(p(y|\beta)) - \log(\pi(\beta))$, and $\widehat{\beta^*}$ is the unique minimizer of $nh^* = -\log(p(y|\beta)) - \log(\pi(\beta)) - \log(g(\beta))$.
- Apply to the numerator (resp. denominator) of $\mathbb{E}\left[g(\beta)|y\right]$, with h^* (resp. h) and $q = 1$:
$$\mathbb{E}\left[g(\beta)|y\right] = \sqrt{\frac{\det(\nabla^2 h(\widehat{\beta}))}{\det(\nabla^2 h^*(\widehat{\beta^*}))}}\exp\left(n(h(\widehat{\beta}) - h^*(\widehat{\beta^*}))\right)\left(1 + O(n^{-2})\right).$$

- 2nd-order approximation, but necessitates to solve 2 non-degenerate optimization.
- Availability of all-purpose MC simulation approaches have rendered these methods less used.

Monte-Carlo sampling

- Consider the expectation wrt to measure μ on $\mathcal{Y}$:

$$\mathbb{E}\left[g(Y)\right] = \int_{\mathcal{Y}} g(y)\mu(dy).$$

- Statistical sampling is a natural way to evaluate this integral :
  - Generate m iid observations $y_1, y_2, \cdots, y_m$ from μ and compute

$$\bar{g}_m = \frac{1}{m}\sum_{i=1}^{m} g(y_i).$$

  - By the LLN, $\bar{g}_m$ converges in probability (or even a.s.) to $\mathbb{E}\left[g(Y)\right]$.
  - This justifies $\bar{g}_m$ as an approximation for $\mathbb{E}\left[g(Y)\right]$ for large m.
- This suggests to use MC sampling to approximate the MMSE $\mathbb{E}\left[\beta|y\right]$.
- One has to sample from the posterior distribution.
- Bayesian posterior distributions are generally non-standard which may not easily allow sampling from them.

Monte-Carlo sampling

- Consider the MMSE :

$$\mathbb{E}\left[\beta|y\right] = \frac{\int_{\mathbb{R}} \theta\phi(y;\beta,\sigma^2)\pi(\beta)d\beta}{\int_{\mathbb{R}} \phi(y;\beta,\sigma^2)\pi(\beta)d\beta},$$

  π is heavy-tailed and easy to sample from.

- Two alternatives :

  1. Ratio of expectations of $\theta\pi(\beta)$ and $\pi(\beta)$ wrt to $\mathcal{N}(y,\sigma^2)$.
     - Sample from $\mathcal{N}(y,\sigma^2)$ and approximate these expectations to get an approximation of $\mathbb{E}\left[\beta|y\right]$.
     - Unwise as π is heavy-tailed while the Gaussian concentrates around its mean, hence missing the contribution from the tails.
  2. Ratio of expectations of $\theta\phi(y;\beta,\sigma^2)$ and $\phi(y;\beta,\sigma^2)$ wrt to π.
     - Sample from π and approximate these expectations to get an approximation of $\mathbb{E}\left[\beta|y\right]$.
     - Not satisfactory either as $p(\beta|y)$ is not as heavy-tailed as π.

- Both alternatives would lead to slow convergence of the sample mean.
- Rather sample directly from $p(\beta|y)$ itself.

Importance sampling

- Consider the expectation wrt to measure μ on $\mathcal{Y}$:

$$\mathbb{E}_\mu \left[g(Y)\right] = \int_{\mathcal{Y}} g(y)\mu(dy).$$

- Suppose that it is difficult/expensive to sample from μ, but there exists a probability measure ν very close to μ from which it is easy to sample.
- Then

$$\mathbb{E}_\mu \left[g(Y)\right] = \mathbb{E}_\nu \left[g(Y)w(Y)\right], \qquad w = \mu/\nu$$

  where $w = \mu/\nu$ (beware of support issues).
- Sample from ν and compute sample mean of gw.
- ν is the importance sampling measure.

Importance sampling

- Consider the expectation wrt to measure μ on $\mathcal{Y}$:

$$\mathbb{E}_\mu \left[g(Y)\right] = \int_{\mathcal{Y}} g(y)\mu(dy).$$

- Suppose that it is difficult/expensive to sample from μ, but there exists a probability measure ν very close to μ from which it is easy to sample.
- Then

$$\mathbb{E}_\mu \left[g(Y)\right] = \mathbb{E}_\nu \left[g(Y)w(Y)\right], \qquad w = \mu/\nu$$

where $w = \mu/\nu$ (beware of support issues).
- Sample from ν and compute sample mean of gw.
- ν is the importance sampling measure.

Example *y_i are n i.i.d. $\mathcal{N}(\theta, \sigma^2)$, where θ and σ^2 are independent, θ has a double exponential distribution with density $e^{-|\theta|}/2$, and σ^2 has the prior density of $(1+\sigma^2)^{-2}$. One can show that*

$$p(\theta, \sigma^2|y) \propto f_1(\sigma^2|\theta)f_2(\theta)e^{-|\theta|}\left(\frac{\sigma^2}{1+\sigma^2}\right)^2,$$

where f_1 is the density of inverse Gamma with parameters $(n/2+1, tfracn2((\theta-\bar{y})^2+s^2)$, f_2 is the $n+1$ t-density, with location $\bar{y}$ (sample mean) and scale $\propto s$ (sample std).
The tails are mostly captured in $f_1(\sigma^2|\theta)f_2(\theta)$, which can serve as an importance sampling density.

MCMC: Iterative MC sampling

- MC methods necessitate complete determination of the sampling measure.
- Situations where posterior distributions are incompletely specified or are specified indirectly cannot be handled, e.g., only in terms of several conditional and marginal distributions.
- It turns out that it is indeed possible in such cases to adopt an iterative MC sampling scheme.
- These iterative MC procedures typically generate a random sequence with the Markov property such that this Markov chain is ergodic with the limiting distribution being the target posterior distribution.
- A whole class of such iterative procedures are dubbed Markov chain Monte Carlo (MCMC) procedures.

A glimpse of Markov chains

Definition *A sequence of random variables* $\{X_n\}_{n\geq 0}$ *is a Markov chain if for any* n*, given* X_n*, the past* $\{X_j : j \leq n-1\}$ *and the future* $\{X_j : j \geq n+1\}$ *are independent, i.e. for any two events* A *and* B *defined respectively in terms of the past and the future,*

$$P(A \cap B|X_n) = P(A|X_n)P(B|X_n),$$

Definition *A Markov chain has a time homogeneous or stationary transition probability iff the probability distribution of* $X_{n+1}|X_n = x$*, and the past,* $\{X_j : j \leq n-1\}$ *depends only on* x*. This is specified in terms of the transition kernel* P*, where* $P(x, A) = \Pr(X_{n+1} \in A|X_n = x)$*. If the state-space (set of values* X_n *can take), is countable, this reduces to specifying the transition probability matrix* P*,* $P_{ij} = \Pr(X_{n+1} = j|X_n = i)$*.*

Lemma *Suppose that* $\{X_n\}_{n\geq 0}$ *is a Markov chain on a countable state-space with stationary transition probabilities. Then the joint probability distribution of* $\{X_n\}_{n\geq 0}$ *is*

$$\Pr\left(X_i = j_i : i = 0, \cdots, n\right) = \Pr\left(X_0 = j_0\right) \prod_{i=1}^{n} P_{j_{i-1} j_i}.$$

A glimpse of Markov chains

Definition *A probability distribution μ is called stationary or invariant for a transition probability P or the associated Markov chain $\{X_n\}$ iff : when the probability distribution of X_0 is μ then the same is true for X_n for all $n \geq 1$.*

Lemma *Suppose that $\{X_n\}_{n\geq 0}$ is a Markov chain on a state-space S with kernel P. Then a probability distribution μ with density p is a stationary for P if*

$$\int_A p(x)dx = \int_S P(x, A)p(x)dx, \forall A \subset S.$$

In the countable case : μ is a left eigenvector of P.

Definition *A Markov chain $\{X_n\}_{n\geq 0}$ with a countable state space S and transition probability matrix P is said to be irreducible if for any two states i and j the probability of the Markov chain visiting j starting from i is positive. A similar notion of irreducibility can be stated for general state spaces.*

LLN for Markov chains

Theorem *Let* $\{X_n\}_{n\geq 0}$ *be a Markov chain with state-space* S *and kernel* P*. Further, suppose it is (Harris) irreducible and has a stationary distribution* μ*. Then, for any bounded function* $g : S \to \mathbb{R}$ *and for any initial distribution of* X_0

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_i) \underset{\mathcal{P}}{\to} \int_S g(x)\mu(dx).$$

LLN for Markov chains

Theorem *Let $\{X_n\}_{n\geq 0}$ be a Markov chain with state-space S and kernel P. Further, suppose it is (Harris) irreducible and has a stationary distribution μ. Then, for any bounded function $g : S \to \mathbb{R}$ and for any initial distribution of X_0*

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_i) \underset{\mathcal{P}}{\to} \int_S g(x)\mu(dx).$$

Remark *Under additional conditions (e.g. aperiodicity for countable state-space), one can also assert that the distribution of X_n converges (in an appropriate topology) to μ.*

LLN for Markov chains

Theorem *Let $\{X_n\}_{n\geq 0}$ be a Markov chain with state-space S and kernel P. Further, suppose it is (Harris) irreducible and has a stationary distribution μ. Then, for any bounded function $g : S \to \mathbb{R}$ and for any initial distribution of X_0*

$$\frac{1}{n}\sum_{i=0}^{n-1} g(X_i) \underset{\mathcal{P}}{\to} \int_S g(x)\mu(dx).$$

Remark *Under additional conditions (e.g. aperiodicity for countable state-space), one can also assert that the distribution of X_n converges (in an appropriate topology) to μ.*

This result is the backbone of Monte-Carlo Markov Chain (MCMC) methods.

Metropolis-Hastings algorithm

- A very general-purpose MCMC method.
- Idea : not sample from the target density, but simulate a Markov chain whose stationary/invariant distribution is the target density.

Metropolis-Hastings algorithm

- A very general-purpose MCMC method.
- Idea : not sample from the target density, but simulate a Markov chain whose stationary/invariant distribution is the target density.

Inputs : State-space S, μ a probability measure with density p on S. X_0.
Proposal transition kernel Q with density $q : \forall x \in S$, easy to sample from $q(x, \cdot)$.

Compute: Acceptance probability $\rho : \rho(x, y) = \min\left(1, \frac{p(y)q(y,x)}{p(x)q(x,y)}\right), \forall (x, y) \quad \text{s.t.} \quad p(x)q(x, y) > 0$.

repeat

  if $X_n = x$ **then**
    draw a sample Y_n from $q(x, \cdot)$;

  Set

$$X_{n+1} = \begin{cases} Y_n & \text{with prob. } \rho(X_n, Y_n) \\ X_n & \text{with prob. } 1 - \rho(X_n, Y_n). \end{cases}$$

until *convergence*;

Metropolis-Hastings algorithm

- A very general-purpose MCMC method.
- Idea : not sample from the target density, but simulate a Markov chain whose stationary/invariant distribution is the target density.

Inputs : State-space S, μ a probability measure with density p on S. X_0.
Proposal transition kernel Q with density q : $\forall x \in S$, easy to sample from $q(x, \cdot)$.

Compute: Acceptance probability ρ : $\rho(x,y) = \min\left(1, \frac{p(y)q(y,x)}{p(x)q(x,y)}\right), \forall (x,y) \quad \text{s.t.} \quad p(x)q(x,y) > 0.$

repeat

  if $X_n = x$ **then**
    draw a sample Y_n from $q(x, \cdot)$;
  Set

$$X_{n+1} = \begin{cases} Y_n & \text{with prob. } \rho(X_n, Y_n) \\ X_n & \text{with prob. } 1 - \rho(X_n, Y_n). \end{cases}$$

until *convergence*;

Proposition *(i)* $\{X_n\}_{n\geq 0}$ *is a Markov chain on* S.

(ii) μ *is a stationary/invariant probability distribution for* $\{X_n\}_{n\geq 0}$.

(iii) If Q *is irreducible on* S*, then so is* $\{X_n\}_{n\geq 0}$ *and the LLN on Markov chains applies.*

Metropolis-Hastings algorithm

- A very general-purpose MCMC method.
- Idea : not sample from the target density, but simulate a Markov chain whose stationary/invariant distribution is the target density.

Inputs : State-space S, μ a probability measure with density p on S. X_0.
Proposal transition kernel Q with density q : $\forall x \in S$, easy to sample from $q(x, \cdot)$.

Compute: Acceptance probability ρ : $\rho(x, y) = \min\left(1, \frac{p(y)q(y,x)}{p(x)q(x,y)}\right), \forall (x, y) \quad \text{s.t.} \quad p(x)q(x, y) > 0.$

repeat

if $X_n = x$ **then**

draw a sample Y_n from $q(x, \cdot)$;

Set

$$X_{n+1} = \begin{cases} Y_n & \text{with prob. } \rho(X_n, Y_n) \\ X_n & \text{with prob. } 1 - \rho(X_n, Y_n). \end{cases}$$

until *convergence*;

- A distinctive feature of MH for Bayesian inference is that it is enough to know p up to a multiplicative constant : acceptance probability depends on ratios.
- Conclusion : the normalization constant in the posterior density is of no importance at all in the MH algorithm.

Gibbs sampling

- The Gibbs sampler is especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high- dimensional space but having some special structure.
- The most interesting aspect of this approach is that it only draw samples from univariate distributions through the course of iterations.

Gibbs sampling

- The Gibbs sampler is especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high- dimensional space but having some special structure.
- The most interesting aspect of this approach is that it only draw samples from univariate distributions through the course of iterations.

Inputs : State-space $S \subset \mathbb{R}^p$, μ a probability measure on S.
Initial configuration X_0.

for $n = 1 \cdots$ **do**
  Draw sample $X_{n,2}$ from the univariate distribution $\mu(\cdot | x_{n-1,2}, \cdots, x_{n-1,p})$;
  for $i = 2$ **to** p **do**
    Draw sample $X_{n,i}$ from the univariate distribution $\mu(\cdot | X_{n,1}, \cdots, X_{n,i-1}, x_{n-1,i+1}, \cdots, x_{n-1,p})$.

Gibbs sampling

- The Gibbs sampler is especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high- dimensional space but having some special structure.
- The most interesting aspect of this approach is that it only draw samples from univariate distributions through the course of iterations.

Inputs : State-space $S \subset \mathbb{R}^p$, μ a probability measure on S.
Initial configuration X_0.

for $n = 1 \cdots$ **do**
  Draw sample $X_{n,2}$ from the univariate distribution $\mu(\cdot|x_{n-1,2}, \cdots, x_{n-1,p})$;
  for $i = 2$ **to** p **do**
    Draw sample $X_{n,i}$ from the univariate distribution $\mu(\cdot|X_{n,1}, \cdots, X_{n,i-1}, x_{n-1,i+1}, \cdots, x_{n-1,p})$.

Proposition *(i) The Gibbs sampler is a special case of MH.*

(ii) $\{X_n\}_{n\geq 0}$ is an irreducible Markov chain on S.

(iii) μ is a stationary/invariant probability distribution for $\{X_n\}_{n\geq 0}$.

(iv) The LLN on Markov chains applies.

Gibbs sampling

- The Gibbs sampler is especially suitable for generating an irreducible aperiodic Markov chain that has as its stationary distribution a target distribution in a high- dimensional space but having some special structure.
- The most interesting aspect of this approach is that it only draw samples from univariate distributions through the course of iterations.

Inputs : State-space $S \subset \mathbb{R}^p$, μ a probability measure on S.
Initial configuration X_0.

for $n = 1 \cdots$ **do**
  Draw sample $X_{n,2}$ from the univariate distribution $\mu(\cdot | x_{n-1,2}, \cdots, x_{n-1,p})$;
  for $i = 2$ **to** p **do**
    Draw sample $X_{n,i}$ from the univariate distribution $\mu(\cdot | X_{n,1}, \cdots, X_{n,i-1}, x_{n-1,i+1}, \cdots, x_{n-1,p})$.

Proposition *(i) The Gibbs sampler is a special case of MH.*

(ii) $\{X_n\}_{n\geq 0}$ is an irreducible Markov chain on S.

(iii) μ is a stationary/invariant probability distribution for $\{X_n\}_{n\geq 0}$.

(iv) The LLN on Markov chains applies.

- Is popular for hierarchical Bayesian modeling.
- e.g. in Markov random fields with Ising model.
- Improved estimators can be obtained via variance-reduction (Rao-Blackwell theorem).

Langevin diffusion

- A Langevin diffusion X in $\mathbb{R}^p$, is a homogeneous Markov process defined by the SDE

Drift — Diffusion

$$dX(t) = \frac{1}{2}\rho(X(t))dt + dW(t), \; t > 0, \; X(0) = \boldsymbol{x}_0,$$

  - $\rho = -\nabla \log \mu$, μ is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^p$;
  - W is a p-dimensional Brownian process.

Langevin diffusion

- A Langevin diffusion X in $\mathbb{R}^p$, is a homogeneous Markov process defined by the SDE

$$dX(t) = \frac{1}{2}\underbrace{\rho(X(t))}_{\text{Drift}}dt + \underbrace{dW(t)}_{\text{Diffusion}},\ t > 0,\ X(0) = \boldsymbol{x}_0,$$

  - $\rho = -\nabla \log \mu$, μ is everywhere non-zero and suitably smooth target density function on $\mathbb{R}^p$;
  - W is a p-dimensional Brownian process.

- Under mild assumptions, the SDE has a unique strong solution and $X(t)$ has a stationary distribution with density precisely μ.
- Opens the door to approximating integrals $\int_{\mathbb{R}^p} g(\theta)\mu(\theta)d\theta$ by the average value of the Langevin diffusion path

$$\frac{1}{T}\int_0^T g(X(t))dt, \quad \text{for large enough } T.$$

Langevin diffusion

- Euler (forward) discretization

$$X_{n+1} = X_n + \frac{\delta}{2}\rho(X_n) + \sqrt{\delta} Z_n, \quad X_0 = \boldsymbol{x}_0,$$
$$= X_n - \frac{\delta}{2}\nabla \log \mu(X_n) + \sqrt{\delta} Z_n$$

  - $\delta > 0$: the discretization step-size ;
  - $\{Z_n\}_{n \geq 0}$ i.i.d. $\sim \mathcal{N}(0, \mathrm{I}_p)$.

Langevin diffusion

- Euler (forward) discretization

$$X_{n+1} = X_n + \frac{\delta}{2}\rho(X_n) + \sqrt{\delta} Z_n, \quad X_0 = \boldsymbol{x}_0,$$
$$= X_n - \frac{\delta}{2}\nabla \log \mu(X_n) + \sqrt{\delta} Z_n$$

  - $\delta > 0$: the discretization step-size ;
  - $\{Z_n\}_{n\geq 0}$ i.i.d. $\sim \mathcal{N}(0, \mathrm{I}_p)$.
- $\{X_n\}_{n\geq 0}$ is a Markov chain.
- The Girsanov formula implies :

$$X^\delta(t) \overset{\text{def}}{=} X_0 + \frac{1}{2}\int_0^t \rho(\overline{X}(s))ds + \int_0^t dW(s)ds,$$

$\overline{X}(t) = X_n$ for $t \in [n\delta, (n+1)\delta[$.

$$\mathrm{KL}\left(\mu(\{X(t) : t \in [0,T]\}), \mu(\left\{X^\delta(t) : t \in [0,T]\right\})\right) \underset{h\to 0}{\longrightarrow} 0.$$

- The average value can then be naturally approximated via

$$\frac{\delta}{T}\sum_{n=0}^{\lfloor T/\delta \rfloor} X_n.$$

Langevin diffusion

- Euler (forward) discretization

$$
\begin{aligned}
X_{n+1} &= X_n + \frac{\delta}{2}\rho(X_n) + \sqrt{\delta} Z_n, \quad X_0 = \boldsymbol{x}_0, \\
&= X_n - \frac{\delta}{2}\nabla \log \mu(X_n) + \sqrt{\delta} Z_n
\end{aligned}
$$

$$
X^\delta(t) \stackrel{\text{def}}{=} X_0 + \frac{1}{2}\int_0^t \rho(\overline{X}(s))ds + \int_0^t dW(s)ds, \qquad \overline{X}(t) = X_n \text{ for } t \in [n\delta, (n+1)\delta[.
$$

Theorem *Assume that ρ is locally Lipchitz continuous and verifies an appropriate growth condition. Then,*

$$
\left\| \mathbb{E}\left[X^\delta(T)\right] - \mathbb{E}\left[X(T)\right] \right\|_2 \leq \mathbb{E}\Big[\sup_{0 \leq t \leq T} \left\|X^\delta(t) - X(t)\right\|_2\Big] \xrightarrow[\delta \to 0]{} 0.
$$

If ρ is uniformly Lipschitz continuous, the optimal consistency rate $\delta^{1/2}$ is achieved.

Take-away messages

- Bayesian modeling is a flexible paradigm.
- Bayesian inference involves optimization and integration.
- Bayesian interpretation is not universal: all PMLE are NOT MAP.
- Bayesian computation is essentially easier for MAP.
- For MMSE, MCMC methods are general and versatile, though scaling with dimension can be an issue.
- A variety of applications: signal and image processing, communication, biostatistics, classification, machine learning.

https://fadili.users.greyc.fr/

Thanks
Any questions ?