Exponential families

1. definition
2. examples – Poisson, Bernoulli
3. properties of exp. family
4. conjugate priors

Exponential family:

sufficient statistic $\mathcal{X} \to \mathbb{R}^d$

log normalizer

$$p(x \mid \eta) = \pi(x) \exp\{ \eta \cdot t(x) - a(\eta) \}$$

natural parameter $\eta \in \mathbb{R}^d$

base measure
carrier distribution

log normalizer

$$a(\eta) = \log \int \pi(x) \exp\{\eta \cdot t(x)\}\, dx$$

$$p(x \mid \eta) = \frac{\pi(x) \exp\{\eta \cdot t(x)\}}{\int \pi(x') \exp\{\eta \cdot t(x')\}\, dx'}$$

integrates to 1

Bernoulli

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x} \qquad x \in \{0,1\}, \quad \theta \in [0,1] \qquad \theta: \text{prob of "heads"}$$

$$= \exp\{x \log \theta + (1-x) \log(1-\theta)\}$$

$$= \exp\{x \log(\theta/1-\theta) + \log(1-\theta)\}$$

Brown (1986)
Efron (2023)
Wainwright + Jordan (2008)

$$\left.\begin{aligned} t(x) &= x \\ \pi(x) &= 1 \\ \eta &= \log(\theta/1-\theta), \quad \theta = 1/(1+e^{-\eta}) \\ a(\eta) &= -\log(1-\theta) = \log(1+e^{\eta}) \end{aligned}\right]$$

exponential family form of the Bernoulli

Poisson

$$p(x \mid \lambda) = \frac{1}{x!} \lambda^x \exp\{-\lambda\} \qquad x \in \{0, 1, 2, \ldots\}$$
$$\lambda > 0$$

$$= \frac{1}{x!} \exp\{ x \log \lambda - \lambda \}$$

$\eta = \log \lambda \,, \quad \lambda = \exp\{\eta\}$

$t(x) = x$

$\pi(x) = 1/x!$

$a(\eta) = \lambda = \exp\{\eta\}$

Moments of an exp. family

$$\mathbb{E}[t(X)] = \nabla_\eta a(\eta)$$

$$\nabla_\eta a(\eta) = \nabla_\eta \left\{ \log \int \exp\{\eta \cdot t(x)\} \pi(x)\, dx \right\}$$

$t(x) : \mathcal{X} \to$ d-vector.

$\eta : \mathbb{R}^d$

$$= \frac{\nabla_\eta \int \exp\{\eta \cdot t(x)\} \pi(x)\, dx}{\int \exp\{\eta \cdot t(x)\} \pi(x)\, dx}$$

$$= \int t(x) \underbrace{\boxed{\frac{\exp\{\eta \cdot t(x)\} \pi(x)}{\int \exp\{\eta \cdot t(x')\} \pi(x')\, dx'}}}_{p(x \mid \eta)} dx$$

$$= \int t(x)\, p(x \mid \eta)\, dx = \mathbb{E}[t(X)]$$

$$\frac{\partial^2 a(\eta)}{\partial \eta_i \partial \eta_j} = \mathbb{E}[t_i(X) t_j(X)] - \mathbb{E}[t_i(X)]\, \mathbb{E}[t_j(X)] = \mathrm{Cov}(t_i(X), t_j(X))$$

Example: Poisson

$\mathbb{E}[X] = \lambda$

$$\frac{da}{d\eta} = \exp\{\eta\} = \lambda$$

$$\frac{d^2 a}{d\eta^2} = \exp\{\eta\} = \lambda$$

<u>mean parameterization</u>

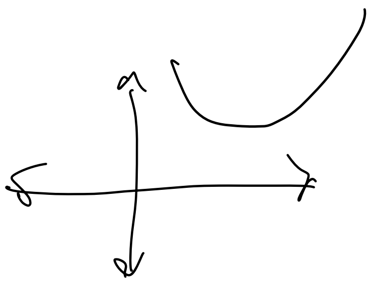
1-1 relationship b/w η and $\mathbb{E}_\eta[t(X)]$ ← mean parameterization.

why?

1d expfam.

$t(x)$ scalar, $\eta \in \mathbb{R}$

$$\frac{d^2 a}{d\eta^2} = \mathrm{Var}(X)$$



Variance > 0

$a(\eta)$ convex

1-1 relationship b/w $\frac{da}{d\eta}$ and η

$\frac{da}{d\eta} = \mathbb{E}[t(X)]$ 1-1

<u>notation</u> $\mu \triangleq \mathbb{E}[t(X)]$

$\eta_\mu(\mu)$ maps μ to η

$\mu_\eta(\eta)$ maps η to μ

<u>Maximum likelihood estimation</u>

$X_i \sim \mathrm{expfam}(\eta)$

$$\hat{\eta}_{MLE} = \arg\max_\eta \sum_{i=1}^n \log p(x_i \mid \eta)$$

1 – MLE for expfam

2 – conjugate priors for EF

3 – GLMs

4 – hierarchical GLMs

anonclass.com

Code: 5596
1782

$$\mathcal{L}(\eta) = \sum_{i=1}^{n} \log p(x_i \mid \eta)$$

$$= \sum_{i=1}^{n} \left(\log \pi(x_i) + \eta \cdot t(x) - a(\eta) \right)$$

$$= \sum_{i=1}^{n} \log \pi(x_i) + \eta \left(\sum_{i=1}^{n} t(x) \right) - n\, a(\eta)$$

$$\nabla_\eta \mathcal{L}(\eta) = \left(\sum_i t(x_i) \right) - n \nabla_\eta a(\eta)$$

$$\mu_\eta(\eta) \triangleq \nabla_\eta a(\eta) = \mathbb{E}[t(x); \eta]$$

$$\left(\sum_i t(x_i) \right) - n\mu = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_i t(x_i)$$

$$\hat{\eta}_{MLE} = \eta_\mu(\hat{\mu}_{MLE})$$

Conjugacy

expfam EDM:

$$\eta \sim f(\eta; \lambda)$$

$$x_i \mid \eta \sim \text{expfam}(\eta)$$

observe $x_{1:n}$, goal: $p(\eta \mid x_{1:n}; \lambda)$

$$p(\eta \mid \mathbf{x}; \lambda) \propto p(\eta; \lambda) \prod_{i=1}^{n} p(x_i \mid \eta)$$

Suppose $p(\eta \mid \mathbf{x}; \lambda) = f(\eta; \hat{\lambda})$ where $\hat{\lambda}$ is a function of λ, $\mathbf{x}$

$$p(x \mid \eta) = \pi_\ell(x) \exp\{\eta \cdot t_\ell(x) - a_\ell(\eta)\}$$ likelihood model

$$p(\eta; \lambda) = \pi_c(\eta) \exp\{\lambda_1 \cdot \eta + \lambda_2(-a_\ell(\eta)) - a_c(\lambda_1, \lambda_2)\}$$

$$t_c(\eta) = (\underbrace{\eta}_{d\text{-vector}}, \underbrace{-a_\ell(\eta)}_{\text{scalar}}) \qquad \mathbb{R}^d \to \mathbb{R}^{d+1}$$

λ_1 : d-vector

λ_2 : scalar

$\lambda = (\lambda_1, \lambda_2)$

$$p(\eta \mid \mathbf{x}; \lambda) \propto \overbrace{\pi_c(\eta) \exp\{\lambda_1 \cdot \eta + \lambda_2(-a(\eta))\}}^{p(\eta) \text{ unnormalized}} \overbrace{\exp\{\eta \cdot (\sum_{i=1}^{n} t(x_i)) - n\, a_\ell(\eta)\}}^{p(\mathbf{x} \mid \eta)}$$

$$= \pi_c(\eta) \exp\{(\lambda_1 + \Sigma t(x_i)) \cdot \eta + (\lambda_2 + n)(-a(\eta))\}$$ ← unnormalized

$$t(\eta) = (\eta, -a(\eta)) \qquad a_c(\hat{\lambda}_1, \hat{\lambda}_2)$$

$$\hat{\lambda}_1 = \lambda_1 + \sum_i t(x_i)$$

$$\hat{\lambda}_2 = \lambda_2 + n$$

$\mu(\eta)$

$$\mathbb{E}_\lambda[\mu(\eta)] = \lambda_1 / \lambda_2$$ prior expectation of the mean parameter

$$\mathbb{E}[\mu(\eta) \mid \mathbf{x}; \lambda] = \frac{\hat{\lambda}_1}{\hat{\lambda}_2} = \frac{\lambda_1 + \sum_{i=1}^{n} t(x_i)}{\lambda_2 + n}$$

- posterior variance goes down as $1/n$
- posterior predictive distribution.

$$p(x_{new} \mid \mathbf{x}; \lambda) = \pi_\ell(x) \exp\{a_c(\hat{\lambda}_1 + t(x_{new}), \hat{\lambda}_2 + 1) - a_c(\hat{\lambda}_1, \hat{\lambda}_2)\}$$

- Beta-Bernoulli

Generalized linear models (GLMs)

$$\mathbb{E}[Y \mid X = x_i, \beta] = f(\beta \cdot x_i)$$

Y ↑ response $\quad x_i$ ↑ features $\quad \beta$ ↑ coefficients ("weights")

in linear regression $f(\eta) = \eta$

logistic regression $f(\eta) = \sigma(\eta)$

anon class – ~~4955~~ 4501

GLM.

given features x_i

$$w_i = \beta \cdot x_i$$

$$\mu_i = f(w_i)$$

$$\eta_i = \eta_\mu(\mu_i)$$

$$y_i \sim \text{expfam}(\eta_i)$$

f: link function or response function

$\mu_i \triangleq \mathbb{E}[Y \mid X = x_i, \beta]$

$$\begin{tikzcd} \beta \arrow[dr] & & & & \\ & w_i \arrow[r, "f = \eta_\mu^{-1}"] & \mu_i \arrow[r, "\eta_\mu"] & \eta_i \arrow[r] & y_i \\ x_i \arrow[ur] & & & & \end{tikzcd}$$

Choices : (1) link function

(2) response distribution

$$\begin{tikzcd} \beta \arrow[dr] & & \\ & \eta_i \arrow[r] & y_i \\ x_i \arrow[ur] & & \end{tikzcd}$$

put a prior on coefficients $\beta \rightarrow$ Bayesian GLM.

$$\left.\begin{aligned} &\beta \sim \mathcal{N}_p(0, \lambda^2) \\ &y_i \mid x_i, \beta \sim \text{GLM}(x_i, \beta) \qquad i = 1..n \end{aligned}\right]$$

special link : canonical link $\quad f = \eta_\mu^{-1}$

$$\eta_i = \beta \cdot x_i$$

$$y_i \mid \eta_i \sim \text{expfam}(\eta_i)$$

MAP inference

$$\beta \sim \mathcal{N}_p(0, \lambda^2)$$

$$y_i \mid x_i, \beta \sim \text{GLM}(x_i, \beta) \qquad i = 1..n$$

$$\mathcal{L}(\beta) = -\frac{1}{2\lambda^2}\sum_{j=1}^{p} \beta_j^2 + \sum_{i=1}^{n} \overbrace{(\eta_i \cdot y_i - a(\eta_i))}^{\ell_i}$$

$\log p(y_i \mid x_i, \beta)$

$$\ell_i \triangleq \eta_i \cdot y_i - a(\eta_i)$$

$$\eta_i = \beta \cdot x_i$$

$$\nabla_\beta \ell_i = \nabla_\eta \ell_i \, \nabla_\beta \eta_i$$

$$= (y_i - \nabla_\eta a(\eta_i)) \nabla_\beta \eta_i$$

$$= (y_i - \mathbb{E}[Y \mid X = x_i, \beta]) \, x_i$$

Exp Fams: Efron.
GLMs: Nelder & McCullough

$$\nabla_\beta \mathcal{L} = -\frac{1}{\lambda^2}\beta + \sum_{i=1}^{n} \underbrace{(y_i - \mathbb{E}[Y \mid X = x_i, \beta])}_{\text{signed residual}} x_i$$

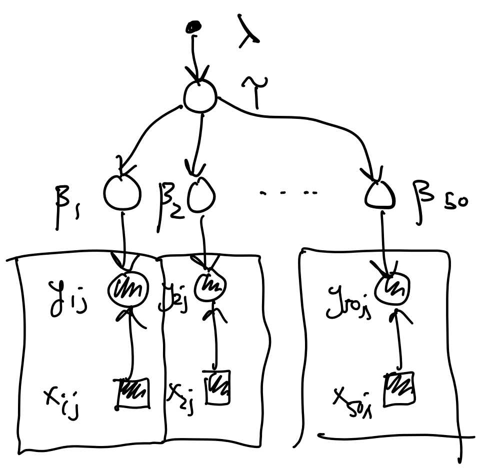
general link function: $\eta_i = \eta_\mu(f(\beta \cdot x_i))$

Hierarchical GLMs

data from each state $\mathcal{D}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$

How to analyze?

- separate GLMs for each state
- pool all the data, one GLM
- hierarchical model



hierarchical GLM.

$\tau \sim p(\tau; \lambda)$

For each $i \in \{1, \ldots, 50\}$

$\quad \beta_i \sim p(\beta \mid \tau)$

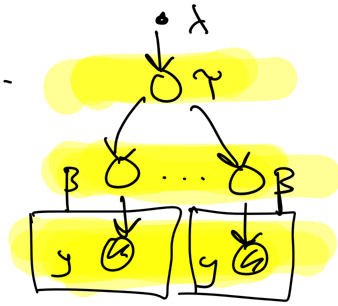
$\quad$ For each $j \in \{1 \ldots n_i\}$

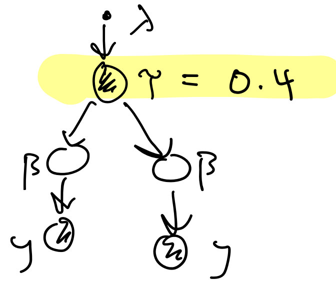
$\quad\quad y_{ij} \mid x_{ij}, \beta_i \sim \mathrm{GLM}(\beta_i)$

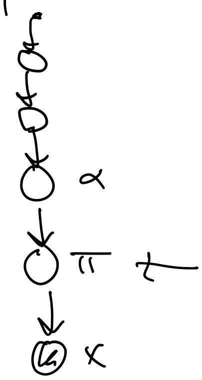
$$\log p(-) = \log p(\tau; \lambda) + \sum_{i=1} \left(\log p(\beta_i | \tau) + \sum_{j=1} \log p(y_{ij} | \beta_i, x_{ij}) \right)$$

$$p(\beta_i | \mathbf{x}, \mathbf{y}) \propto \mathbb{E}_{\lambda}\left[p(\beta_i | \tau) \,|\, \mathbf{x}_{-i}, \mathbf{y}_{-i} \right] p(\mathbf{y}_i | \mathbf{x}_i, \beta_i)$$

Empirical Bayes.
Hierarchical Bayes.
└ Gelman & Hill.







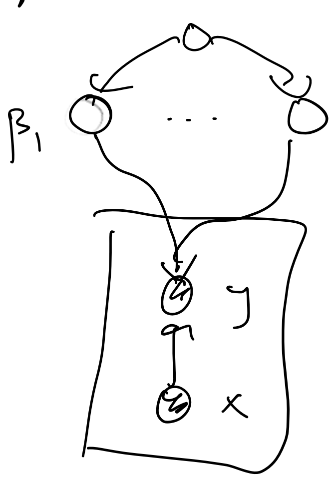
Gaussian process : $f(\cdot) \sim GP(\mu(\cdot), K(\cdot,\cdot))$ — Bayesian nonparametrics.

$$\mathbb{E}_{\mathcal{D}}\left[d(\beta^*_{1:50}, \hat{\beta}_{1:50}) \right] = R(\beta^*)$$ Risk.

$$\mathbb{E}_{\tau^*}\left[\mathbb{E}_{\mathcal{D}}\left[d(\beta^*_{1:50}, \hat{\beta}_{1:50}) \right] \right] = R_{\text{Bayes}}$$

$$R^{\text{MLE}}_{\text{Bayes}} > R^{\text{Shrinkage}}_{\text{Bayes}}$$

Robbins, Efron, James-Stein . . .



— ARD
— Horseshoe priors.