

Exponential families

Outline

- 1) Exponential families
- 2) Differential identities
- 3) MGF

Exponential Families

An s-parameter exponential family is a family

$\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ with densities of the form

$$P_\eta(x) = e^{\eta^T T(x) - A(\eta)} h(x)$$

wrt base measure μ on sample space \mathcal{X}

Components of P_η :

- $\eta \in \Xi \subseteq \mathbb{R}^s$ called natural parameter
- $T(x)$ is s-dimensional sufficient statistic
factorization theorem: $g_\eta(T(x)) = e^{\eta^T T(x) - A(\eta)}$
- $h(x) \geq 0$ called base density or carrier density
can be absorbed into base measure μ
- $A(\eta)$ called log-partition function or normalizing const.
 $A(\cdot)$ determined by T, h, μ :

$$A(\eta) = \log \left[\int_{\mathcal{X}} e^{\eta^T T(x)} h(x) d\mu(x) \right] \leq \infty$$

The natural parameter space is the set of all η that give us normalizable p_η

$$\Xi_1 = \{ \eta : A(\eta) < \infty \}$$

Note: \mathcal{P} can use strict subset ($\Xi \subsetneq \Xi_1$) if scientific considerations constrain η

$A(\eta)$ is always a convex function $\Rightarrow \Xi_1$ convex set

If we absorb h into μ , log-densities almost linear

$$\log p_\eta(x) = \eta^T T(x) - A(\eta) \quad (\text{wrt } h, d\mu)$$

Can think of $T(x)$ as basis

Very nice structure when we multiply densities

- Combining evidence from independent obs.
- Prior \times likelihood in Bayesian calculations

or divide them

- Calculating conditional probabilities
- Likelihood ratios
- Relative densities

Examples

Poisson (single obs)

$$X \sim \text{Pois}(\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in 0, 1, \dots$$

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

$$\eta(\lambda) = \log \lambda \quad T(x) = x$$

$$A(\eta) = \lambda = e^\eta \quad h(x) = \frac{1}{x!}$$

Poisson (n obs) $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\lambda)$

$$p_\lambda(x) = \prod_{i=1}^n \exp \{ (\log \lambda) x_i - \lambda \} \frac{1}{x_i!}$$

$$= \exp \{ (\log \lambda) (\sum_i x_i) - n\lambda \} \prod_i \frac{1}{x_i!}$$

$$\eta(\lambda) = \log \lambda \quad T(x) = \sum_i x_i$$

$$A(\eta) = ne^\eta \quad h(x) = \prod_i \frac{1}{x_i!}$$

Generic (n obs) $X_1, \dots, X_n \stackrel{iid}{\sim} p_\eta^{(n)}(x) = e^{\eta' T^{(n)}(x) - A^{(n)}(\eta)} h^{(n)}(x)$

$$p_\eta(x) = \prod_{i=1}^n \exp \{ \eta' T^{(n)}(x_i) - A^{(n)}(\eta) \} h^{(n)}(x_i)$$

$$= \exp \left\{ \underbrace{\eta' \left(\sum_i T^{(n)}(x_i) \right)}_{T(x)} - \underbrace{n A^{(n)}(\eta)}_{A(\eta)} \right\} \underbrace{\prod_i h^{(n)}(x_i)}_{h(x)}$$

\Rightarrow Dimension of $T(x)$ doesn't grow with n

Differential Identities

Write $e^{A(\eta)} = \int e^{\eta^T T(x)} h(x) d\mu(x)$ (*)

We can derive lots of useful identities by differentiating (*) on both sides, pulling derivative inside } [not always allowed]

Keener Thm 2.4 for $f: \mathcal{X} \rightarrow \mathbb{R}$ let

$$\Xi_f = \left\{ \eta \in \mathbb{R}^s : \int |f| e^{\eta^T T} h d\mu < \infty \right\}$$

Then $g(\eta) = \int f e^{\eta^T T} h d\mu$ has cts partial derivatives of all orders for $\eta \in \Xi_f^o$. & we can get them by differentiating under the \int sign.

\Rightarrow on Ξ_f^o , $A(\eta)$ has all partial derivatives

Differentiate once:

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \frac{\partial}{\partial \eta_j} \int e^{\eta^T T(x)} h(x) d\mu(x)$$

$$\cancel{e^{A(\eta)}} \frac{\partial A}{\partial \eta_j}(\eta) = \int T_j(x) e^{\eta^T T(x) - A(\eta)} h(x) d\mu(x)$$

$$\Rightarrow \frac{\partial A}{\partial \eta_j}(\eta) = \mathbb{E}_\eta [T_j(x)]$$

$$\nabla A(\eta) = \mathbb{E}_\eta [T(x)]$$

Diff twice:

$$\frac{\partial^2}{\partial \eta_i \partial \eta_k} e^{A(\eta)} = \frac{\partial^2}{\partial \eta_i \partial \eta_k} \int e^{\eta' T} h d\mu$$

$$e^{A(\eta)} \left(\frac{\partial^2 A}{\partial \eta_i \partial \eta_k} + \underbrace{\frac{\partial A}{\partial \eta_i}}_{\mathbb{E}[T_i]} \underbrace{\frac{\partial A}{\partial \eta_k}}_{\mathbb{E}[T_k]} \right) = \underbrace{\int T_i T_k e^{\eta' T - A(\eta)} h d\mu}_{\mathbb{E}[T_i T_k]}$$

$$\frac{\partial^2 A}{\partial \eta_i \partial \eta_k}(\eta) = \text{Cov}_\eta(T_i, T_k)$$

$$\nabla^2 A(\eta) = \text{Var}_\eta(T(x)) \in \mathbb{R}^{s \times s}$$

Example: Poisson: $T(x) = X$, $A(\eta) = e^\eta (= \lambda)$

$$\mathbb{E}_\eta[X] = \frac{d}{d\eta} e^\eta = e^\eta = \lambda$$

$$\text{Var}_\eta(X) = \frac{d^2}{d\eta^2} e^\eta = e^\eta = \lambda$$

NB: We would get wrong answer by differentiating wrt λ

Moment-generating function

We can get k^{th} order moments of $T(X)$ by

1) Differentiating (*) k times, then

2) Dividing by $e^{A(\eta)}$

That is because $M_{\eta}^T(u) = e^{A(\eta+u) - A(\eta)}$
is the moment-generating function (mgf)
of $T(X)$ when $X \sim P_{\eta}$

$$\begin{aligned} M_{\eta}^{T(X)}(u) &= \mathbb{E}_{\eta} \left[e^{u'T(X)} \right] \\ &= \int e^{u'T} e^{\eta'T - A(\eta)} h d\mu \\ &= e^{A(\eta+u) - A(\eta)} \underbrace{\int e^{(\eta+u)'T - A(\eta+u)} h d\mu}_{= 1} \end{aligned}$$

Useful for

- finding moments
- finding dist. of sums of indep. RVs

Cumulant-generating function

$$K_{\eta}^T(u) = \log M_{\eta}^T(u) = A(\eta+u) - A(\eta) \quad (A \text{ is sometimes called cgf})$$

Other Parameterizations

Sometimes it is more convenient to use a different parameterization:

$$p_{\theta}(x) = e^{\eta(\theta)'T(x) - B(\theta)} h(x)$$
$$B(\theta) = A(\eta(\theta))$$

Many, many examples, sometimes requires massaging to see that they are exp. fam.s:

Ex: Normal $X \sim N(\mu, \sigma^2)$ $\mu \in \mathbb{R}$ $\sigma^2 > 0$

Let $\theta = (\mu, \sigma^2)$

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu-x)^2}{2\sigma^2}}$$

$$= \exp \left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right\}$$

$$\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \quad T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad h(x) = 1$$

$$B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2)$$

Natural parameterization

$$p_{\eta}(x) = e^{\eta_1' \begin{pmatrix} x \\ x^2 \end{pmatrix} - A(\eta)}$$

$$A(\eta) = \frac{-\eta_1^2}{4\eta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\eta_2}\right)$$

More examples

Binomial

$$X \sim \text{Binom}(n, \theta)$$

$$p_{\theta}(x) = \theta^x (1-\theta)^{n-x} \binom{n}{x} \quad x = 0, \dots, n$$

$$= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \binom{n}{x}$$

$$= \exp \left\{ \log\left(\frac{\theta}{1-\theta}\right) \cdot x + n \log(1-\theta) \right\} \binom{n}{x}$$

$$\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \quad \text{"log odds ratio"}$$

Beta

$$X \sim \text{Beta}(\alpha, \beta)$$

$$p_{\alpha, \beta}(x) = x^{\alpha-1} (1-x)^{\beta-1} / B(\alpha, \beta)$$

← Beta function

$$= \exp \left\{ \alpha \log x + \beta \log(1-x) - \log B(\alpha, \beta) \right\} \frac{1}{x(1-x)}$$

$$\eta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad T(x) = \begin{pmatrix} \log x \\ \log(1-x) \end{pmatrix} \quad h(x) = \frac{1}{x(1-x)}$$

Practically everything else on wikipedia too:

Beta, Gamma, Multinom., Dirichlet, Pareto, Wishart...

Interpretation: Exponential tilting

Can think of $p_\eta(x) = e^{\eta'T(x) - A(\eta)} h(x)$ as
an exponential tilt of the carrier $h(x)$

1) Start with carrier $h(x)$

2) Multiply by $e^{\eta'T(x)}$

3) Re-normalize by $e^{-A(\eta)}$

$T(x) = (T_1(x), \dots, T_s(x))$ gives linear space of directions
in which we can tilt $h(x)$

$\Xi_\eta =$ all tilts after which normalization is possible

\Rightarrow Decomposition into η, T, h, A very non-unique

1) Only $\text{span}(T_1, \dots, T_s)$ matters

2) Could absorb h into μ ($d\nu(x) = h d\mu(x)$)
(wlog $h(x) \equiv 1$ if we want)

3) Can add constant to $T(x)$

\vdots many others

Distribution of $T(x)$

Suppose $X \sim p_\eta(x) = e^{\eta' T(x) - A(\eta)}$ wrt μ
(wlog $h \equiv 1$)

Then $T(X) \sim q_\eta(t) = e^{\eta' t - A(\eta)}$ wrt ν ,

where ν is the measure μ "pushed forward"
through $T: \mathcal{X} \rightarrow \mathbb{R}^S$

$$\nu(B) \triangleq \mu(\{x: T(x) \in B\})$$

$$\begin{aligned} \mathbb{P}_\eta(T(X) \in B) &= \int \mathbb{1}_B(T(x)) e^{\eta' T(x) - A(\eta)} d\mu(x) \\ &= \int \mathbb{1}_B(t) e^{\eta' t - A(\eta)} d\nu(t) \end{aligned}$$

Simplest in discrete case: (drop $h \equiv 1$ assumption)

$$\begin{aligned} \mathbb{P}_\eta(T(X) = t) &= \sum_{x: T(x) = t} e^{\eta' T(x) - A(\eta)} h(x) \mu(\{x\}) \\ &= e^{\eta' t - A(\eta)} \underbrace{\sum_{x: T(x) = t} h(x) \mu(\{x\})}_{\nu(\{t\})} \end{aligned}$$

Canonical Form

The structure is most evident when:

- $T(x) = x$ (wlog: sufficiency reduction)
- $h(x) \equiv 1$ (wlog: absorb h into μ)
- $\theta = \eta$ (wlog: parameterize by η)

Then, we say the family is in canonical form:

$$p_{\eta}(x) = e^{\eta'x - A(\eta)}$$

Minimal form

Form of $p_\eta(x) = e^{z^T T(x) - A(z)} h(x)$ minimal if

$z \in \Xi$ and $T(x)$ satisfy no linear constraints:

no $a \neq 0$, $b \in \mathbb{R}$ s.t. $z^T a = b$ for all $z \in \Xi$
or $T(x)^T a = b$ \mathcal{P} -a.s.

Otherwise we can represent \mathcal{P} as an r -dim. ex. fam.

for some $r < s$

If p_η minimal, then $T(x)$ is minimal suff.

Need to show $l(\cdot; x) = l(\cdot; y) + c_{xy} \Rightarrow T(x) = T(y)$
(\Leftarrow holds by suff.)

$$l(z; x) - l(z; y) = z^T \underbrace{(T(x) - T(y))}_a$$

Can find $\eta, \zeta \in \Xi$ s.t. $\eta^T a \neq \zeta^T a$ unless $a = 0$

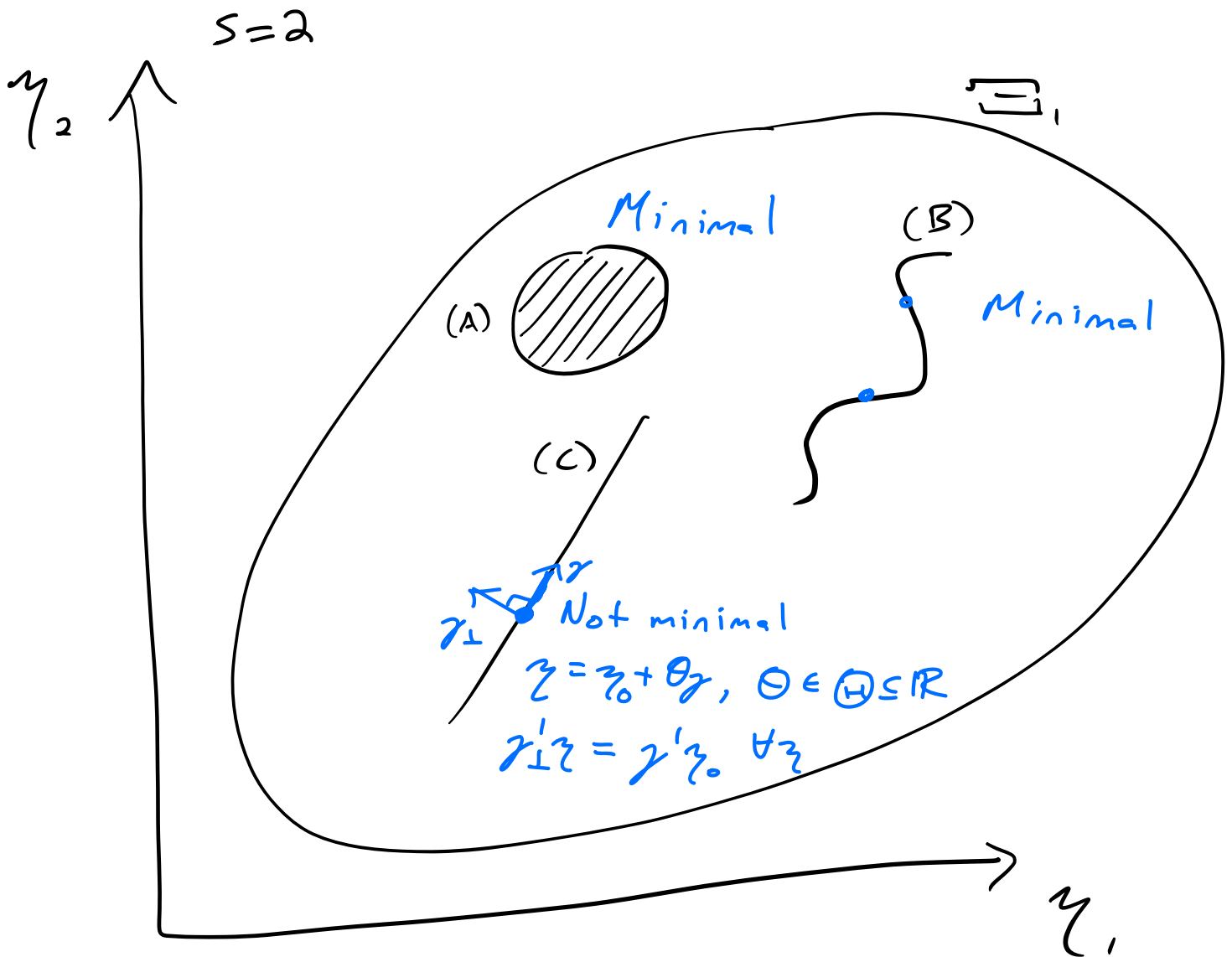
$$\Rightarrow T(x) = T(y)$$

Important: Converse is not true: reducing dimension of $T(x)$ may or may not be a data reduction (See multinomial problem in hw)

Diagram

$$p_{\eta}(x) = e^{\eta' T(x) - A(\eta)} h(x)$$

Three subfamilies $\{P_{\eta} : \eta \in \Xi \subseteq \Xi_1\}$



Can make (C) minimal for $s=1$

$$e^{\eta' T(x) - A(\eta)} h(x) = e^{\underbrace{\theta (\eta' T(x))}_{\text{new } T(x)} - A(\eta_0 + \theta \eta)} h(x)$$