Exponential Families

Exponential families are perhaps the most important statistical models, and nearly every model we discuss in this course is an exponential family. Today we will examine exponential families carefully, examining carefully their algebraic and combinatorial structure and how we have already seen it appearing

Regular Exponential Families

- $\mathcal{X}$ a sample space with σ-algebra $\mathcal{A}$ with (σ-finite) measure ν.
- $T : \mathcal{X} \longrightarrow \mathbb{R}^k$ a statistic (i.e. a measurable map)
- $h : \mathcal{X} \longrightarrow \mathbb{R}_{\geq 0}$ a measurable function
- The natural parameter space is

$$N = \left\{ \eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x) e^{\eta^t T(x)} d\nu(x) < \infty \right\}.$$

- For $\eta \in N$ get a probability density function p_η on $\mathcal{X}$:

$$p_\eta(x) := h(x) e^{\eta^t T(x) - \phi(\eta)}, \text{ where}$$

$$\phi(\eta) := \log \int_{\mathcal{X}} h(x) e^{\eta^t T(x)} d\nu(x).$$

- P_η := probability measure on $(\mathcal{X}, \mathcal{A})$ with ν-density p_η.
- set $\nu^T := \nu \circ T^{-1}$, a measure on the Borel σ-algebra $\mathcal{B}_k$ of $\mathbb{R}^k$:

$$\nu^T(B) = \nu(T^{-1}(B)) \quad \text{for} \quad B \in \mathcal{B}_k$$

(smallest σ-algebra containing all $\pi_k(-\infty, b_k)$)

- The support of ν^T is the set

$$\operatorname{supp}(\nu^T) := \bigcap_{A \subseteq \mathbb{R}^k \text{ closed}: \nu^T(\mathbb{R}^k \setminus A) = 0} A.$$

(Part of $\mathbb{R}^k$ carrying info about $\mathcal{X}$...)

Definition (16) Let k be a positive integer. The probability distributions

$$\mathcal{P} := \{ P_\eta : \eta \in N \}$$

form a regular exponential family of order k if N is an open subset in $\mathbb{R}^k$ and the affine dimension of $\operatorname{supp}(\nu^T)$ is equal to k. The statistic T inducing $\mathcal{P}$ is called the canonical sufficient statistic.

- It is clear from the definition that T is a sufficient statistic, but in fact T has a stronger property when $\mathcal{P}$ is regular that corresponds to the condition on the dimension of $\text{supp}(\nu^T)$:
- The canonical sufficient statistic T of a regular exponential family is minimal: For any sufficient statistic T' and for every $x, y \in \mathcal{X}$

$$T(x) = T(y) \quad \text{whenever} \quad T'(x) = T'(y).$$

That is, T is a function of T': $\exists f$ s.t. $T(x) = f(T'(x))$.

- It turns out that any exponential family that is not regular can be reduced by reparametrization to a regular exponential family of order $< k$. Here, the statistic for the model is also transformed into a minimal one...
- Today, we want to understand the algebra and geometry of exponential families that are submodels of regular exponential families. We start in the discrete setting.

Discrete Exponential Families

Example 1 (Discrete Random Variables)

- $\mathcal{X} := [r] = \{1, \dots, r\}$.
- ν = the counting measure on $[r]$: $\nu(A) = |A| \quad \forall A \subseteq [r]$.
- $T: [r] \to \mathbb{R}^{r-1}$ where $T: i \mapsto (\delta_{i=1}, \delta_{i=2}, \dots, \delta_{i=r-1})^t \in \mathbb{R}^{r-1}$,

where $\delta_{i=j} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases}$

- $h(x) \equiv 1$.
- Note: $\text{Supp}(\nu^T) = \{0, e_1, \dots, e_{r-1}\} \subseteq \mathbb{R}^{r-1}$ since

$$\nu^T(B) = \nu(T^{-1}(B)) = 0 \iff T^{-1}(B) \cap \{0, e_1, \dots, e_{r-1}\} = \emptyset.$$

$\Rightarrow \dim(\text{supp}(\nu^T)) = r-1.$

Given $\eta \in \mathbb{R}^{r-1}$

$$\phi(\eta) = \log \int_{\mathcal{X}} h(x) e^{\eta^t T(x)} d\nu(x) = \log\left(\sum_{i=1}^{r} e^{\eta^t T(i)}\right) = \log\left(1 + \sum_{i=1}^{r-1} e^{\eta_i}\right) < \infty.$$

$\Rightarrow$ The natural parameter space is $N = \mathbb{R}^{r-1}$ (and hence N open)

$\Rightarrow$ $\mathcal{P} = \{ P_\eta : P_\eta(i) = e^{\eta^t T(i) - \phi(\eta)} \ \forall i \in [r] \}$

is a regular exponential family of order $r-1$.

- We claim $\mathcal{P} \simeq \Delta^\circ_{r-1}$: P_η is a vector in $\mathbb{R}^r$. For $1 \le i \le r-1$

$$P_\eta(i) = e^{\eta^t T(i) - \phi(\eta)} = \frac{e^{\eta_i}}{1 + \sum_{i=1}^{r-1} e^{\eta_i}}, \text{ and}$$

$$P_\eta(r) = 1 - \sum_{i=1}^{r-1} P_\eta(i) = \frac{1}{1 + \sum_{i=1}^{r-1} e^{\eta_i}}.$$

- Given $(p_1, \dots, p_r) \in \Delta^\circ_{r-1}$, taking $\eta_i = \log\left(\frac{p_i}{p_r}\right)$ establishes $\mathbb{R}^{r-1} = N \simeq \Delta^\circ_{r-1}$.
- Hence all discrete exponential families are submodels of a regular exponential family.
- To get such a submodel, we consider the following:
  - $T : [r] \longrightarrow \mathbb{R}^k$ a measurable map. $\longleftrightarrow$ an assignment of vectors $T(i) \in \mathbb{R}^k \ \forall i \in [r] \dots$
  - $h : [r] \longrightarrow \mathbb{R}_{>0}$ a vector: $h = (h_1, \dots, h_r)^t \in \mathbb{R}^r_{>0}$ $(h_i := h(i))$
- Since measurability is trivial in the discrete case, any collection of vectors works for T and h is any positive vector.
- Each choice $\{ T(i) \in \mathbb{R}^k : i \in [r] \}$ and $h \in \mathbb{R}^r_{>0}$ defines an exponential subfamily of Δ°_{r-1}.
- Let $Z(\eta) := \sum_{i \in [r]} h_i e^{\eta^t T(i)}$
- Write $T(i) := (a_{1i}, \dots, a_{ki})^t \in \mathbb{R}^k$ and $\eta = (\eta_1, \dots, \eta_k)^t \in \mathbb{R}^k$.

Then $\forall i \in [r]$,

$$P_\eta(i) = h(i) e^{\eta^t T(i) - \phi(\eta)} = h_i e^{\eta^t T(i) - \log(Z(\eta))}$$

$$= \frac{h_i e^{\eta^t T(i)}}{Z(\eta)} = \frac{1}{Z(\eta)} h_i \prod_{i \in [k]} e^{\eta_j a_{ji}}.$$

- Setting $\theta_i := \exp(\eta_i)$ $\forall i \in [r]$, we get

$$P_\theta(i) = \frac{1}{Z(\theta)} h_i \prod_{j \in [k]} \theta_j^{a_{ji}}.$$

- Hence, given any T and h the resulting exponential family is

$$\mathcal{P}_{T,h} := \left\{ P_\eta : \eta \in \mathbb{R}^{r-1} \right\} \simeq \left\{ P_\theta \in \Delta^\circ_{r-1} : P_\theta(i) = \frac{1}{Z(\theta)} h_i \prod_{j \in [k]} \theta_j^{a_{ji}} \right\}.$$

- Look familiar?

Example (2) Γ a simplicial complex with set of facets $\text{facets}(\Gamma)$.

- $h = \underline{1} = (1, \dots, 1)^t \in \mathbb{R}^{|\mathcal{R}|}$, where $\mathcal{R}$ is state space for $(X_1, \dots, X_m)$ discrete and $[m] = G(\Gamma)$.
- Define the statistic $T : \mathcal{R} \longrightarrow \mathbb{R}^{k \times |\mathcal{R}_F|}$ where $T : \underline{i} \mapsto \sum_{\underline{j} \in \mathcal{R}_F,\ F \in \text{facets}(\Gamma)} \delta_{\underline{i}_F = \underline{j}}$.

$$\Rightarrow a_{\underline{j}\,\underline{i}} = \begin{cases} 1 & \text{if } \underline{i} \text{ restricted to } F \text{ is } \underline{j} \\ 0 & \text{otherwise.} \end{cases}$$

- Given $\eta \in \mathbb{R}^{k \times |\mathcal{R}_F|}$, set $\theta_{\underline{j}}^{(F)} := \exp(\eta_{\underline{j},F})$.
- Then $P_\theta(\underline{i}) = \frac{1}{Z(\theta)} \prod_{F \in \text{facets}(\Gamma)} \theta^{(F)}_{\underline{i}_F}$.

$\Rightarrow \mathcal{P}_{T,h}$ = the log-linear hierarchical model for Γ.

- In fact, all log-linear models are discrete regular exponential families...

- Given $A = [a_{ji}]_{j=1,\, i=1}^{k,\, r} \in \mathbb{Z}^{k \times r}$, define the statistic $T(i) := (a_{1i}, \dots, a_{ki})^t$ to be the columns of A.
- Take some $h \in \mathbb{R}^r_{>0}$.
- For any $p_\eta \in \mathcal{P}_{T,h}$ we have that

$$p_\eta(i) = h_i e^{\eta^t T(i) - \phi(\eta)}$$

$$\log(p_\eta(i)) = \log(h_i) + \eta^t T(i) - \phi(\eta), \text{ or equivalently,}$$

- $\log(p_\theta(i)) = \log(h_i) + \log(\theta)^t A - \log(Z(\theta))$.
- If we assume $\mathbb{1} \in \text{rowspan}(A)$ then $b^t A = \mathbb{1}$ for some b and

- $$\mathcal{P}_{T,h} = \left\{ p \in \Delta^\circ_{r-1} : \log(p) \in \log(h) + \text{rowspan}(A) \right\}.$$

Definition (17) Let $A \in \mathbb{Z}^{k \times r}$ be an integer matrix such that $\mathbb{1} \in \text{rowspan}(A)$ and let $h \in \mathbb{R}^r_{>0}$. The log-affine model associated to A and h is

$$M_{A,h} := \left\{ p \in \Delta^\circ_{r-1} : \log(p) \in \log(h) + \text{rowspan}(A) \right\}.$$

- So we just proved the following:

Proposition (1) Discrete exponential families are the log-affine models.

- This gives a geometric interpretation of discrete exponential families as arising from taking linear subspaces of the natural parameter space of the full regular exponential family for discrete random variables:

$$p_\eta : \log(p_\eta) \in \log(h) + (\eta + \phi(\eta) b)^t A. \qquad (b^t A = \mathbb{1})$$

- Letting $h = \mathbb{1}$, we recover the log-linear models: $M_A := M_{A,\mathbb{1}}$ from chapter 1.
- We called log-linear models toric models. Hence, discrete exponential families are the toric models up to rescaling...

- For $A \in \mathbb{Z}^{k \times r}$ and $h \in \mathbb{R}^r_{>0}$ with $\mathbb{1} \in \text{rowspan}(A)$, define the monomial map

$$\phi^{A,h}: \mathbb{R}^k \to \mathbb{R}^r \quad \text{s.t.} \quad \phi_i^{A,h}(\theta) = h_i \prod_{j \in [k]} \theta_j^{a_{ji}}.$$

- The vanishing ideal

$$I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k)) \subseteq \mathbb{R}[P] := \mathbb{R}[p_1, \dots, p_r]$$

is the toric ideal of $M_{A,h}$.

- By definition, $V(I_{A,h})$ is the Zariski closure of $M_{A,h}$.

- mapping $p_j \mapsto p_j/h_j$ we can recover the generators of $I_{A,h}$ from $I_{A,\mathbb{1}}$. $\Rightarrow$ discrete exponential families are toric models up to rescaling.

- We now prove a claim made in chapter 1 that the toric variety $V(I_L)$ for $L = \ker_{\mathbb{Z}} A$ is the Zariski closure of $M_{A,\mathbb{1}}$:

Proposition (2) Let $A \in \mathbb{Z}^{k \times r}$. Then the toric ideal $I_{A,\mathbb{1}}$ is a binomial ideal and

$$I_{A,\mathbb{1}} = \langle p^u - p^v : u, v \in \mathbb{N}^r \text{ and } Au = Av \rangle.$$

Proof: Any $p^u - p^v$ such that $Au = Av$ is in $I_{A,\mathbb{1}}$ since for any point in $\phi^{A,\mathbb{1}}(\mathbb{R}^k)$ we get $p_i = \prod_{j \in [k]} \theta_j^{a_{ji}}$ and

$$p^u = \prod_{j \in [k]} \theta_j^{\langle u, (a_{j1}, \dots, a_{jr}) \rangle} = \prod_{j \in [k]} \theta_j^{\langle v, (a_{j1}, \dots, a_{jr}) \rangle} = p^v.$$

($Au = Av$.)

- To show $p^u - p^v$ with $Au = Av$ generate $I_{A,\mathbb{1}}$, it suffices to show that they generate it as a vector space:

- IF $f(p) \in I_{A,\mathbb{1}}$ and $c_u p^u$ a monomial in $f(p)$ with $c_u \neq 0$ then for any $\theta \in \mathbb{R}^k$ we have $f(\phi^{A,\mathbb{1}}(\theta)) = 0$.

$$\Rightarrow f(\phi^{A,\mathbb{1}}(\theta)) = 0 \text{ as a polynomial in } \theta.$$

- plugging $\phi^{A,\mathbb{1}}(\theta)$ into $c_u p^u$ we get $c_u \theta^{Au}$ which must cancel...

$\Rightarrow$ There is another monomial $c_v p^v$ such that plugging in $\phi^{A,Z}(\theta)$ yields the same monomial $c_v \theta^{Au} \Rightarrow Au = Av$.

$\rightsquigarrow$ subtract off: $f - c_u(p^u - p^v)$, then induct. ▨

- Aside from being a beautiful correspondence between discrete exponential families and toric varieties, the geometric connection can help us solve important statistical problems...

Theorem (1) The maximum likelihood estimator of a discrete exponential family exists and is unique if and only if the observed value $T(x)$ of the sufficient statistic lies in the interior of the conic hull of $\mathrm{supp}(v^T)$: $\mathrm{cone}(v^T) = \left\{ \sum_{v^{(i)} \in \mathrm{supp}(v^T)} \lambda_i v^{(i)} : \lambda_i \geq 0 \right\}$.

- The geometry of a toric variety V_A defined by $A \in \mathbb{Z}^{k \times r}$ is encoded in the polytope $P_A = \mathrm{conv}(A)$.

$$\mathrm{cone}(P_A) = \mathrm{cone}(A) = \mathrm{cone}(\mathrm{supp}(v^T)).$$

Gaussian Regular Exponential Families

- We can similarly generalize the observation made last week to show that all multivariate Gaussian models are regular exponential families.

- Let $N_m := \{ N_m(\mu, \Sigma) : \mu \in \mathbb{R}^m, \Sigma \in PD_m \}$.

- $\mathcal{X} = \mathbb{R}^m$ with σ-algebra $\mathcal{B}_m$ and Lesbesgue measure ν.

- $T(x) := \left(x_1, \dots, x_m, -x_1^2/2, \dots, -x_m^2/2, -x_1 x_2, \dots, -x_{m-1} x_m\right)^t \in \mathbb{R}^m \times \mathbb{R}^{\binom{m+1}{2}}$.

$\hookrightarrow$ polynomial components linearly independent $\Rightarrow \dim(\mathrm{supp}(v^T)) = m + \binom{m+1}{2}$.

- $h(x) \equiv 1$

- $\eta \in \mathbb{R}^m \times \mathbb{R}^{\binom{m+1}{2}}$, let η_{cmq} be its first m components and $\eta_{\mathrm{cmxm}} \in S_{m \times m}$ the symmetric $m \times m$ matrix formed by the last $\binom{m+1}{2}$ components.

Fact $x \mapsto e^{\eta^t T(x)}$ integrable $\Leftrightarrow$ $\eta_{[m \times m]}$ is PD.

$\Rightarrow N = \mathbb{R}^m \times PD_m$ is the natural parameter space.

- For $\eta \in N$,

$$\phi(\eta) = -\frac{1}{2}\left(\log\det(\eta_{[m\times m]}) - \eta_{[m]}^t \eta_{[m\times m]} \eta_{[m]} - m\log(2\pi)\right).$$

- Hence, the densities P_η are of the form

$$P_\eta(x) = \frac{1}{\sqrt{(2\pi)^m \det(\eta^{-1}_{[m\times m]})}} \exp\left(\eta^t T(x) - \eta_{[m]}^t \eta_{[m\times m]} \eta_{[m]}/2\right)$$

$$\eta^t T(x) = \eta^t\left(x_1, \dots, x_m, -x_1^2/2, \dots, -x_m^2/2, -x_1x_2, \dots, -x_{m-1}x_m\right)^t$$

$$= \eta_{[m]}^t x + \left(\frac{-1}{2}\right) x^t \eta_{[m\times m]} x \quad \text{, where } x = (x_1, \dots, x_m)^t.$$

- Setting $\Sigma = \eta^{-1}_{[m\times m]}$ and $\mu = \Sigma^{-1}\eta_{[m]}$ we get

$$P_{\mu,\Sigma} = \frac{1}{\sqrt{(2\pi)^m \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right).$$

- Just as in the discrete setting, the submodels of N_m that are also exponential families arise from taking linear subspaces of the natural parameter space $\mathbb{R}^m \times PD_m$.

exponential subfamilies of N_m $\longleftrightarrow$ linear subspaces of $\mathbb{R}^m \times PD_m$ $\longleftrightarrow$ $T(x)$ that maps $x \in \mathbb{R}^m$ to a vector of degree two polynomials in x with no constant term.

- In general take $L \cap (\mathbb{R}^m \times PD_m)$ for any linear subspace L of $\mathbb{R}^m \times \mathbb{S}_{m\times m}$.
- In practice, $L = L_1 \times L_2$ for linear subspaces $L_1 \subseteq \mathbb{R}^m$ and $L_2 \subseteq \mathbb{S}_{m\times m}$.
- Last time: $L_1 = \mathbb{R}^m$ and $L_2 = I_{m\times m}$.

Example (3) Let $L_1 := \mathbb{R}^2$ and L_2 be the linear subspace of $S_{2\times 2}$:

$$\left\{ \eta_{11}\begin{pmatrix} 1 & 1/2 \\ 1/2 & 0 \end{pmatrix} + \eta_{12}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \eta_{22}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} : \eta_{11}, \eta_{22}, \eta_{12} \in \mathbb{R} \right\}.$$

Check that $(L_1 \times L_2) \cap (\mathbb{R}^2 \times PD_2)$ is the collection of parameters for an exponential family given by the statistic

$$T(x) = (x_1, x_2, (-x_1^2/2) + x_1x_2, -x_2^2/2, -x_1x_2)^t.$$

- Since the models are linear in the inverse covariance matrix, the vanishing ideals of $N_m(0, \Sigma)$ also have interesting properties...
- If we set $L_1 = 0$ then the vanishing ideal of the model is

$$I(L_2^{-1}) \subseteq \mathbb{R}[\sigma_{ij} : 1 \le i \le j \le m] \text{ where}$$

$$L_2^{-1} := \{ \Sigma = [\sigma_{ij}] \in PD_m : \Sigma^{-1} \in L_2 \cap PD_m \}$$ is the inverse linear space of L_2.

- Since the map $\Sigma \mapsto \Sigma^{-1}$ is rational, $I(L_2^{-1})$ will have nice properties... (examples coming!)
- Today, we didn't say "diffeomorphism." This idea only becomes important when we want to discuss algebraic exponential families: where we use a nonalgebraic map to transform the natural parameter space and the talk about a semialgebraic subset of the transformed space to define our model:

Definition (18) Let $\{P_\eta : \eta \in N\}$ be a regular exponential family of order k. The subfamily induced by the set $M \subseteq N$ is an algebraic exponential family if there exists an open subset $\bar{N} \subseteq \mathbb{R}^k$, a diffeomorphism $g : N \to \bar{N}$, and a semialgebraic set $A \subseteq \mathbb{R}^k$ such that

$$M = g^{-1}(A \cap \bar{N}).$$