Special Families of Models

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

Focus on Parametric Families

Focus on Parametric Families

Recall our setup:

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, \ldots, X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

The Problem of Point Estimation

1. Assume that F_θ is known up to the parameter θ which is unknown
2. Let $(x_1, \ldots, x_n)$ be a realization of $\mathbf{X} \sim F_\theta$ which is available to us
3. Estimate the value of θ that generated the sample given $(x_1, \ldots, x_n)$

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data:

$\hookrightarrow$ Anything we "do" will be a function of the data $g(x_1, \ldots, x_n)$

So far have concentrated on aspects of data: approximate distributions $+$ data reduction...... But what about $\mathcal{F}$?

Focus on Parametric Families

We describe $\mathcal{F}$ by a *parametrization* $\Theta \ni \theta \mapsto F_\theta$:

Definition (Parametrization)

Let Θ be a set, $\mathcal{F}$ be a family of distributions and $g : \Theta \to \mathcal{F}$ an onto mapping. The pair (Θ, g) is called a *parametrization* of $\mathcal{F}$.

$\hookrightarrow$ assigns a label $\theta \in \Theta$ to each member of $\mathcal{F}$

Definition (Parametric Model)

A *parametric model* with parameter space $\Theta \subseteq \mathbb{R}^d$ is a family of probability models $\mathcal{F}$ parametrized by Θ, $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

So far have seen a number of examples of distributions...
...have worked out certain properties individually

Question

Are there more general families that contain the standard ones as special cases and for which a general and abstract study can be pursued?

Exponential Families of Distributions

Exponential Families of Distributions

Definition (Exponential Family)

Let $\mathbf{X} = (X_1, ..., X_n)$ have joint distribution F_θ with parameter $\theta \in \mathbb{R}^p$. We say that the family of distributions F_θ is a k-parameter exponential family if the joint density or joint frequency function of $(X_1, ..., X_n)$ admits the form

$$f(\mathbf{x}; \theta) = \exp\left\{\sum_{i=1}^{k} c_i(\theta) T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right\}, \quad \mathbf{x} \in \mathcal{X}, \theta \in \Theta,$$

with $\text{supp}\{f(\cdot; \theta)\} = \mathcal{X}$ is independent of θ.

- k need not be equal to p, although they sometimes coincide.
- The value of k may be reduced if c or T satisfy linear constraints.
- We will assume that the representation above is minimal.
  $\hookrightarrow$ Can re-parametrize via $\phi_i = c_i(\theta)$, the **natural parameter**.

Motivation: Maximum Entropy Under Constraints

Consider the following variational problem:

Determine the probability distribution f supported on $\mathcal{X}$ with maximum entropy

$$H(f) = -\int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}$$

subject to the linear constraints

$$\int_{\mathcal{X}} T_i(\mathbf{x}) f(\mathbf{x}) dx = \alpha_i, \qquad i = 1, ..., k$$

Philosophy: How to choose a probability model for a given situation?

Maximum entropy approach:

- In any given situation, choose the distribution that gives *highest uncertainty* while satisfying situation–specific required constraints.

Proposition.

When a solution to the constrained optimisation problem exists, it is unique and has the form

$$f(\mathbf{x}) = Q(\lambda_1, ..., \lambda_k) \exp\left\{\sum_{i=1}^{k} \lambda_i T_i(\mathbf{x})\right\}$$

Proof.

Let $g(\mathbf{x})$ be a density also satisfying the constraints. Then,

$$\begin{aligned}
H(g) &= -\int_{\mathcal{X}} g(\mathbf{x}) \log g(\mathbf{x}) d\mathbf{x} = -\int_{\mathcal{X}} g(\mathbf{x}) \log\left[\frac{g(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x})\right] d\mathbf{x} \\
&= -\underbrace{KL(g\|f)}_{\geq 0} \quad -\int_{\mathcal{X}} g(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\
&\leq -\log Q \underbrace{\int_{\mathcal{X}} g(\mathbf{x}) dx}_{=1} - \int_{\mathcal{X}} g(\mathbf{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\mathbf{x})\right) d\mathbf{x}
\end{aligned}$$

But g also satisfies the moment constraints, so the last term is

$$\begin{aligned} &= -\log Q - \int_{\mathcal{X}} f(\mathbf{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\mathbf{x}) \right) d\mathbf{x} = \int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= H(f) \end{aligned}$$

Uniqueness of the solution follows from the fact that strict equality can only follow when $KL(g\| f) = 0$, which happens if and only if $g = f$. $\square$

- The λ's are the Lagrange multipliers derived by the Lagrange form of the optimisation problem.
- These are derived so that the constraints are satisfied.
- They give us the $c_i(\theta)$ in our definition of exponential families.
- Note that the presence of $S(\mathbf{x})$ in our definition is compatible: $S(\mathbf{x}) = c_{k+1} T_{k+1}(\mathbf{x})$, where c_{k+1} *does not* depend on θ. (provision for a multiplier that may not depend on parameter)

Example (Binomial Distribution)

Let $X \sim$Binomial(n, θ) with n known. Then

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \exp\left[\log\left(\frac{\theta}{1-\theta}\right)x + n\ln(1-\theta) + \log\binom{n}{x}\right]$$

Example (Gamma Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim}$Gamma with unknown shape parameter α and unknown scale parameter λ. Then,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\alpha,\lambda) &= \prod_{i=1}^{n} \frac{\lambda^\alpha x_i^{\alpha-1}\exp(-\lambda x_i)}{\Gamma(\alpha)} \\ &= \exp\left[(\alpha-1)\sum_{i=1}^{n}\log x_i - \lambda\sum_{i=1}^{n}x_i + n\alpha\log\lambda - n\log\Gamma(\alpha)\right] \end{aligned}$$

Example (Heteroskedastic Gaussian Distribution)

Let $X_1, ..., X_n \stackrel{iid}{\sim} \mathcal{N}(\theta, \theta^2)$. Then,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}; \theta) &= \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left[-\frac{1}{2\theta^2}(x_i - \theta)^2\right] \\ &= \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{2}\left\{(1 + 2\log\theta) + \log(2\pi)\right\}\right] \end{aligned}$$

Notice that even though $k = 2$ here, the dimension of the parameter space is 1. This is an example of a *curved exponential family*.

Example (Uniform Distribution)

Let $X \sim \mathcal{U}[0, \theta]$. Then, $f_X(x; \theta) = \frac{\mathbf{1}\{x \in [0,\theta]\}}{\theta}$. Since the support of f, $\mathcal{X}$, depends on θ, we do *not* have an exponential family.

Proposition

Suppose that $\mathbf{X} = (X_1, ..., X_n)$ has a one-parameter exponential family distribution with density or frequency function

$$f(\mathbf{x}; \theta) = \exp\left[c(\theta) T(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right]$$

for $x \in \mathcal{X}$ where

(a) the parameter space Θ is open,

(b) $c(\cdot)$ is twice continuously differentiable with non vanishing derivative

Then, d is twice differentiable and

$$\mathbb{E} T(\mathbf{X}) = \frac{d'(\theta)}{c'(\theta)} \quad \& \quad \mathrm{Var}[T(\mathbf{X})] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{[c'(\theta)]^3}$$

Proof.

Define $\phi = c(\theta)$ the *natural parameter* of the exponential family. Let θ_0 be the true parameter. Since $c \in C^2$ and $c' \neq 0$, there exists an open neighbourghood U of $\phi_0 = \eta(\theta_0)$ such that $c^{-1}(\phi)$ exists and is continuously differentiable on U, with derivative

$$\frac{d}{d\phi} c^{-1}(\phi) = \frac{1}{c'(c^{-1}(\phi))}.$$

Since U is open, there exists s sufficiently small so that $\phi_0 + s \in U$. Letting Let $\gamma(\phi) = d(c^{-1}(\phi))$ on U, observe that the m.g.f. of T is

$$\begin{aligned}
\mathbb{E} \exp[sT(\mathbf{X})] &= \int e^{sT(\mathbf{x})} e^{\phi_0 T(\mathbf{x}) - \gamma(\phi_0) + S(\mathbf{x})} d\mathbf{x} \\
&= e^{\gamma(\phi_0+s)-\gamma(\phi_0)} \underbrace{\int e^{(\phi_0+s)T(\mathbf{x}) - \gamma(\phi_0+s) + S(\mathbf{x})} d\mathbf{x}}_{=1} \\
&= \exp[\gamma(\phi_0 + s) - \gamma(\phi_0)]
\end{aligned}$$

Proof.

It follows that:

- $M_T(s) < \infty$ for s sufficiently small, and thus:
  - all moments of T exist,
  - and $M_T(s)$ is infinitely differentiable on an open neighbourhood of 0.
- It consequently follows that $\gamma(s + \phi_0)$ is infinitely differentiable for s small enough, or equivalently, γ is infinitely differentiable in an open neighbourhood of ϕ_0.

We we may differentiate w.r.t. s, and, setting $s = 0$, we get $\mathbb{E}[T(\mathbf{X})] = \gamma'(\phi)$ and $\mathrm{Var}[T(\mathbf{X})] = \gamma''(\phi)$. To complete the proof, we recall that $\gamma(\phi) = d(c^{-1}(\phi))$. Using the fact that $c \in C^2$ and $\gamma \in C^\infty$, a short exercise with the inverse function theorem yields

$$\gamma'(\phi) = d'(\theta)/c'(\theta) \text{ and } \gamma''(\phi) = [d''(\theta)c'(\theta) - d'(\theta)c''(\theta)]/[c'(\theta)]^3$$

$\square$

Exponential Families and Sufficiency

Exercise

Extend the result to the the means, variances and covariances of the random variables $T_1(\mathbf{X}), ..., T_k(\mathbf{X})$ in a k-parameter exponential family

Lemma

Suppose that $\mathbf{X} = (X_1, ..., X_n)$ has a k-parameter exponential family distribution with density or frequency function

$$f(\mathbf{x};\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right]$$

for $x \in \mathcal{X}$. Then, the statistic $(T_1(\mathbf{x}), ..., T_k(\mathbf{x}))$ is sufficient for θ

Proof.

Set $g(\mathbf{T}(\mathbf{x}); \boldsymbol{\theta}) = \exp\{\sum_i T_i(\mathbf{x}) c_i(\boldsymbol{\theta}) + d(\theta)\}$ and $h(\mathbf{x}) = e^{S(\mathbf{x})}\mathbf{1}\{\mathbf{x} \in \mathcal{X}\}$, and apply the factorization theorem. $\square$

Sampling Exponential Families

- The families of distributions obtained by sampling from exponential families are themselves exponential families.
- Let $X_1, ..., X_n$ be iid distributed according to a k-parameter exponential family. Consider the density (or frequency function) of $\mathbf{X} = (X_1, ..., X_n)$,

$$
\begin{aligned}
f(\mathbf{x}; \theta) &= \prod_{j=1}^{n} \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(x_j) - d(\theta) + S(x_j)\right] \\
&= \exp\left[\sum_{i=1}^{k} c_i(\theta)\tau_i(\mathbf{x}) - nd(\theta) + \sum_{j=1}^{n} S(x_j)\right]
\end{aligned}
$$

  for $\tau_i(\mathbf{X}) = \sum_{j=1}^{n} T_i(X_j)$ the *natural statistics*, $i = 1, ..., k$.
- Note that the natural sufficient statistic is k-dimensional $\forall\ n$.
- What about the distribution of $\boldsymbol{\tau} = (\tau_1(\mathbf{X}), ..., \tau_k(\mathbf{X}))$?

The Natural Statistics

Lemma

The joint distribution of $\boldsymbol{\tau} = (\tau_1(\mathbf{X}), \dots, \tau_k(\mathbf{X}))$ is of exponential family form with natural parameters $c_1(\theta), \dots, c_k(\theta)$.

Proof. (discrete case).

Let $\mathcal{T}_{\mathbf{y}} = (\mathbf{x} : \tau_1(\mathbf{x}) = y_1, \dots, \tau_k(\mathbf{x}) = y_k)$ be the level set of $\mathbf{y} \in \mathbb{R}^k$.

$$\begin{aligned}
\mathbb{P}[\boldsymbol{\tau}(\mathbf{X}) = \mathbf{y}] &= \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{y}}} \mathbb{P}[\mathbf{X} = \mathbf{x}] = \delta(\theta) \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{y}}} \exp\left[\sum_{i=1}^{k} c_i(\theta)\tau_i(\mathbf{x}) + \sum_{j=1}^{n} S(x_j)\right] \\
&= \delta(\theta) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right] \sum_{\mathbf{x} \in \mathcal{T}_{\mathbf{y}}} \exp\left[\sum_{j=1}^{n} S(x_j)\right] \\
&= \delta(\theta) \mathcal{S}(\mathbf{y}) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right].
\end{aligned}$$

$\square$

The Natural Statistics

Lemma

For any $A \subseteq \{1, \dots, k\}$, the joint distribution of $\{\tau_i(\mathbf{X}); i \in A\}$ conditional on $\{\tau_i(\mathbf{X}); i \in A^c\}$ is of exponential family form, and depends only on $\{c_i(\theta); i \in A\}$.

Proof. (discrete case).

Let $\mathcal{T}_i = \tau_i(\mathbf{X})$. Have $\mathbb{P}[\mathcal{T} = \mathbf{y}] = \delta(\theta)\mathcal{S}(\mathbf{y}) \exp\left[\sum_{i=1}^{k} c_i(\theta) y_i\right]$, so

$$\mathbb{P}[\mathcal{T}_A = \mathbf{y}_A | \mathcal{T}_{A^c} = \mathbf{y}_{A^c}] = \frac{\mathbb{P}[\mathcal{T}_A = \mathbf{y}_A, \mathcal{T}_{A^c} = \mathbf{y}_{A^c}]}{\sum_{\mathbf{w} \in \mathbb{R}^l} \mathbb{P}[\mathcal{T}_A = \mathbf{w}, \mathcal{T}_{A^c} = \mathbf{y}_{A^c}]}$$

$$= \frac{\delta(\theta)\mathcal{S}((\mathbf{y}_A, \mathbf{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right] \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right]}{\delta(\theta) \exp\left[\sum_{i \in A^c} c_i(\theta) y_i\right] \sum_{\mathbf{w} \in \mathbb{R}^l} \mathcal{S}((\mathbf{w}, \mathbf{y}_{A^c})) \exp\left[\sum_{i \in A} c_i(\theta) w_i\right]}$$

$$= \Delta(\{c_i(\theta) : i \in A\}) h(\mathbf{y}_A) \exp\left[\sum_{i \in A} c_i(\theta) y_i\right] \qquad \square$$

The Natural Statistics and Sufficiency

Look at the previous results through the prism of the canonical parametrisation:

- Already know that $\boldsymbol{\tau}$ is sufficient for $\phi = c(\theta)$.
- But result tells us something even stronger:

that each τ_i is sufficient for $\phi_i = c_i(\theta)$

- In fact any $\boldsymbol{\tau}_A$ is sufficient for ϕ_A, $\forall\ A \subseteq \{1, ..., k\}$
- Therefore, each natural statistic contains the relevant information for each natural parameter
- A useful result that is by no means true for any distribution.

Exponential Families and Completeness

Theorem

Suppose that $\mathbf{X} = (X_1, ..., X_n)$ *has a* k*-parameter exponential family distribution with density or frequency function*

$$f(\mathbf{x};\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta) T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x})\right]$$

for $x \in \mathcal{X}$. *Define* $C = \{(c_1(\theta), ..., c_k(\theta)) : \theta \in \Theta\}$. *If the set* C *contains an open set, then the statistic* $(T_1(\mathbf{X}), ..., T_k(\mathbf{X}))$ *is complete for* θ, *and so minimally sufficient.*

- Intuitively, result says that a k-dimensional sufficient statistic in a k-parameter exponential family will also be complete provided that the effective dimension of the natural parameter space is k.

Proof. (Case $k = 1$)

Recall that T also has a 1-parameter exponential family law, also with parameter $c(\theta)$, with density

$$f_T(t) = \delta(\theta)\mathcal{S}(t)\exp\{c(\theta)t\}$$

where we recall that $\mathcal{S}(t) \geq 0$. Let $g(\cdot)$ be such that $\mathbb{E}_\theta[g(T)] = 0$ for all $\theta \in \Theta$. This translates to

$$\delta(\theta)\int_{\mathbb{R}} g(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt = 0, \qquad \forall\theta \in \Theta.$$

Write $g = g^+ - g^- = g(t)\mathbf{1}\{g(t) \geq 0\} - |g(t)|\mathbf{1}\{g(t) < 0\}$ for the decomposition of g into its positive and negative parts. This yields

$$\int_{\mathbb{R}} g^+(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt = \int_{\mathbb{R}} g^-(t)\mathcal{S}(t)\exp\{c(\theta)t\}dt, \qquad \forall\theta \in \Theta.$$

Since $\mathbb{E}_\theta[g(T)]$ exists for all θ, the two terms above are finite $\forall\theta$.

Our trick now will be to view the two terms integrands as probability densities. Let θ_0 be such that $c(\theta_0)$ is in the interior of C (we can choose such a θ_0 by our assumption on C containing an open set). Let r be equal to the value of either side when $\theta = \theta_0$. Then,

$$F(u) = \int_{-\infty}^{u} \frac{1}{r} g^{+}(t)\mathcal{S}(t)\exp\{c(\theta_0)t\}dt$$

$$G(u) = \int_{-\infty}^{u} \frac{1}{r} g^{-}(t)\mathcal{S}(t)\exp\{c(\theta_0)t\}dt$$

define two probability distributions, with densities given by the integrands. With this definition, our previous equality can be written as

$$\mathbb{E}[\exp\{[c(\theta) - c(\theta_0)]Z\}] = \mathbb{E}[\exp\{[c(\theta) - c(\theta_0)]W\}]$$

for $Z \sim F$ and $W \sim G$. These equalities are valid for all θ, and so for an open neighbourhood of $\phi = c(\theta) - c(\theta_0)$ containing zero. By the characterization property of MGFs, it must be that $F = G$, and so $g^{+} = g^{-}$ almost everywhere, i.e. $g = 0$ a.e., so that T is complete. QED.

Summary on exponential families

An exponential family gives a max-entropy model of the data.

Every natural statistic $T_i(\mathbf{X})$ is a sufficient statistic for the natural parameter: $\phi_i = c_i(\theta)$.

If the mapping $\theta \to \phi$ is nice, then every natural statistic $T_i(\mathbf{X})$ is also complete.

The conjunction of "sufficient" + "complete" almost never occurs outside of exponential families.

KEY LESSON: it's better to have a better model of the data, with more inconvenient properties than to have a worse model of the data.

Transformation Families

Groups Acting on the Sample Space

Basic Idea

Often can generate a family of distributions of the same form (but with different parameters) by letting a group act on our data space $\mathcal{X}$.

Recall: a group is a set G along with a binary operator $\circ$ such that:

1. $g, g' \in G \implies g \circ g' \in G$
2. $(g \circ g') \circ g'' = g \circ (g' \circ g''), \ \forall g, g', g'' \in G$
3. $\exists\ e \in G : e \circ g = g \circ e = g, \ \forall g \in G$
4. $\forall g \in G\ \exists\ g^{-1} \in G : g \circ g^{-1} = g^{-1} \circ g = e$

Often groups are sets of transformations and the binary operator is the composition operator (e.g. $SO(2)$ the group of rotations of $\mathbb{R}^2$):

$$\begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} = \begin{bmatrix} \cos(\phi+\psi) & -\sin(\phi+\psi) \\ \sin(\phi+\psi) & \cos(\phi+\psi) \end{bmatrix}$$

Groups Acting on the Sample Space

- Have a group of transformations G, with $G \ni g : \mathcal{X} \to \mathcal{X}$
- $gX := g(X)$ and $(g_2 \circ g_1)X := g_2(g_1(X))$
- Obviously $\text{dist}(gX)$ changes as g ranges in G.
- Is this change completely arbitrary or are there situations where it has a simple structure?

Definition (Transformation Family)

Let G be a group of transformations acting on $\mathcal{X}$ and let $\{f_\theta(x); \theta \in \Theta\}$ be a parametric family of densities on $\mathcal{X}$. If there exists a bijection $h : G \to \Theta$ then the family $\{f_\theta\}_{\theta \in \Theta}$ will be called a *(group) transformation family* if:

$$X \sim f_\theta \Rightarrow g(X) \sim f_{h(g)*\theta}$$

Hence Θ admits a group structure $\bar{G} := (\Theta, *)$ via:

$$\theta_1 * \theta_2 := h(h^{-1}(\theta_1) \circ h^{-1}(\theta_2))$$

Usually write $g_\theta = h^{-1}(\theta)$, so $g_\theta \circ g_{\theta'} = g_{\theta * \theta'}$

Invariance and Equivariance

Define an equivalence relation on $\mathcal{X}$ via G:

$$x \overset{G}{\equiv} x' \iff \exists\, g \in G : x' = g(x)$$

Partitions $\mathcal{X}$ into equivalence classes called the *orbits* of $\mathcal{X}$ under G

Definition (Invariant Statistic)

A statistic T that is constant on the orbits of $\mathcal{X}$ under G is called an *invariant statistic*. That is, T is invariant with respect to G if, for any arbitrary $x \in \mathcal{X}$, we have $T(x) = T(gx)\ \forall g \in G$.

Notice that it may be that $T(x) = T(y)$ but x, y are not in the same orbit, i.e. in general the orbits under G are subsets of the level sets of an invariant statistic T. When orbits and level sets coincide, we have:

Definition (Maximal Invariant)

A statistic T will be called a *maximal invariant* for G when

$$T(x) = T(y) \iff x \overset{G}{\equiv} y$$

Invariance and Equivariance

- Intuitively, a maximal invariant is a reduced version of the data that represent it as closely as possible, under the requirement of remaining invariant with respect to G.
- If T is an invariant statistic with respect to the group defining a transformation family, then it is ancillary.

Definition (Equivariance)

A statistic $S : \mathcal{X} \to \Theta$ will be called equivariant for a transformation family if $S(g_\theta x) = \theta * s(x), \quad \forall\ g_\theta \in G\ \&\ x \in \mathcal{X}$.

- Equivariance may be a natural property to require if S is used as an *estimator* of the true parameter $\theta \in \Theta$, as it suggests that a transformation of a sample by g_ψ would yield an estimator that is the original one transformed by ψ.

Invariance and Equivalance

Lemma (Constructing Maximal Invariants)

Let $S : \mathcal{X} \to \Theta$ be an equivariant statistic for a transformation family with parameter space Θ and transformation group G. Then, $T(X) = g^{-1}_{S(X)}X$ defines a maximally invariant statistic.

Proof.

$$T(g_\theta x) \stackrel{def}{=} (g^{-1}_{S(g_\theta x)} \circ g_\theta)x \stackrel{eqv}{=} (g^{-1}_{\theta * S(x)} \circ g_\theta)x = [(g^{-1}_{S(x)} \circ g^{-1}_\theta) \circ g_\theta]x = T(x)$$

so that T is invariant. To show maximality, notice that

$$T(x) = T(y) \implies g^{-1}_{S(x)}x = g^{-1}_{S(y)}y \implies y = \underbrace{g_{S(y)} \circ g^{-1}_{S(x)}}_{=g \in G}x$$

so that $\exists g \in G$ with $y = gx$ which completes the proof. □

An important transformation family is the *location-scale* model:

- Let $X = \eta + \tau\varepsilon$ with $\varepsilon \sim f$ completely known.
- Parameter is $\theta = (\eta, \tau) \in \Theta = \mathbb{R} \times \mathbb{R}_+$.
- Define set of transformations on $\mathcal{X}$ by $g_\theta x = g_{(\eta,\tau)}x = \eta + \tau x$ so

$$g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}x = \eta + \tau\mu + \tau\sigma x = g_{(\eta+\tau\mu,\tau\sigma)}x$$

- set of transformations is closed under composition
- $g_{(0,1)} \circ g_{(\eta,\tau)} = g_{\eta,\tau} \circ g_{(0,1)} = g_{(\eta,\tau)}$ (so $\exists$ identity)
- $g(-\eta/\tau, \tau^{-1}) \circ g_{(\eta,\tau)} = g_{(\eta,\tau)} \circ g(-\eta/\tau, \tau^{-1}) = g_{(0,1)}$ (so $\exists$ inverse)
- Hence $G = \{g_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+\}$ is a group under $\circ$.
- Action of G on random sample $\mathbf{X} = \{X_i\}_{i=1}^n$ is $g_{(\eta,\tau)}\mathbf{X} = \eta\mathbf{1}_n + \tau\mathbf{X}$.
- Induced group action on Θ is $(\eta, \tau) * (\mu, \sigma) = (\eta + \tau\mu, \tau\sigma)$.

Location-Scale Families

- The sample mean and sample variance are equivariant, because with $S(\mathbf{X}) = (\bar{X}, V^{1/2})$ where $V = \frac{1}{n-1}\sum(X_j - \bar{X})^2$:

$$
\begin{aligned}
S(g_{(\eta,\tau)}\mathbf{X}) &= \left(\overline{\eta + \tau\mathbf{X}}, \left\{\frac{1}{n-1}\sum(\eta + \tau X_j - \overline{(\eta + \tau X)})^2\right\}^{1/2}\right) \\
&= \left(\eta + \tau\bar{X}, \left\{\frac{1}{n-1}\sum(\eta + \tau X_j - \eta - \tau\bar{X})^2\right\}^{1/2}\right) \\
&= (\eta + \tau\bar{X}, \tau V^{1/2}) = (\eta, \tau) * S(\mathbf{X})
\end{aligned}
$$

- A maximal invariant is given by $A = g^{-1}_{S(\mathbf{X})}\mathbf{X}$ the corresponding parameter being $(-\bar{X}/V^{1/2}, V^{-1/2})$. Hence the vector of residuals is a maximal invariant:

$$
A = \frac{(\mathbf{X} - \bar{X}\mathbf{1}_n)}{V^{1/2}} = \left(\frac{X_1 - \bar{X}}{V^{1/2}}, \ldots, \frac{X_n - \bar{X}}{V^{1/2}}\right)
$$

Example (The Multivariate Gaussian Distribution)

- Let $\mathbf{Z} \sim \mathcal{N}_d(0, I)$ and consider $\mathbf{X} = \boldsymbol{\mu} + \Omega\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Omega\Omega^{\mathsf{T}})$
- Parameter is $(\boldsymbol{\mu}, \Omega) \in \mathbb{R}^d \times \mathrm{GL}(d)$
- Set of transformations is closed under $\circ$
- $g_{(0,I)} \circ g_{(\boldsymbol{\mu},\Omega)} = g_{\boldsymbol{\mu},\Omega} \circ g_{(0,I)} = g_{(\boldsymbol{\mu},\Omega)}$
- $g_{(}-\Omega^{-1}\boldsymbol{\mu}, \Omega^{-1}) \circ g_{(\boldsymbol{\mu},\Omega)} = g_{(\boldsymbol{\mu},\Omega)} \circ g(-\Omega^{-1}\boldsymbol{\mu}, \Omega^{-1}) = g_{(0,I)}$
- Hence $G = \{g_\theta : \theta \in \mathbb{R} \times \mathbb{R}_+\}$ is a group under $\circ$ (affine group).
- Action of G on $\mathbf{X}$ is $g_{(\boldsymbol{\mu},\Omega)}\mathbf{X} = \boldsymbol{\mu} + \Omega\mathbf{X}$.
- Induced group action on Θ is $(\boldsymbol{\mu}, \Omega) * (\boldsymbol{\nu}, \Psi) = (\boldsymbol{\nu} + \Psi\mu, \Psi\Omega)$.

Summary

We have presented two good types of models for data:

- Exponential families: defined from a max-entropy principle.
  Most often, $\mathbf{T}(\mathbf{X})$ is a complete minimally sufficient statistic.
- Transformation families, most often of the form $\mathbf{X} = \mu + \sigma\boldsymbol{\eta}$

We will further study these two types of models in the remainder of the cours. We will focus on exponential families.