Fundamental Theory of Statistical Inference

G. Alastair Young

Department of Mathematics
Imperial College London

LTCC, 2017

Two general classes of models particularly relevant in theory and practice are:

- exponential families
- transformation families

Exponential Families

Suppose that Y depends on parameter $\phi = (\phi^1, \dots, \phi^m)^T$, to be called natural parameters, through a density of the form

$$f_Y(y;\phi) = h(y)\exp\{s^T\phi - K(\phi)\}, \quad y \in \mathcal{Y},$$

where $\mathcal{Y}$ is a set not depending on ϕ. Here $s \equiv s(y) = (s_1(y), \dots, s_m(y))^T$, are called natural statistics.

The value of m may be reduced if either $s = (s_1, \dots, s_m)^T$ or $\phi = (\phi^1, \dots, \phi^m)^T$ satisfies a linear constraint (with probability one). Assume that representation is minimal, in that m is as small as possible.

Full Exponential Family

Provided the natural parameter space Ω_ϕ consists of all ϕ such that

$$\int h(y) \exp\{s^T \phi\} dy < \infty,$$

we refer to the family $\mathcal{F}$ as a full exponential model, or an (m, m) exponential family.

Moments of natural statistics

The moment generating function of the random variable S corresponding to s is

$$\begin{aligned} M(S;t,\phi) &= E\{\exp(S^T t)\} \\ &= \int h(y)\exp\{s^T(\phi+t)\}dy\,\exp\{-K(\phi)\} \\ &= \exp\{K(\phi+t)-K(\phi)\}. \end{aligned}$$

Then

$$E(S_i;\phi) = \frac{\partial K(\phi)}{\partial \phi^i},$$

Also,

$$\mathrm{cov}(S_i, S_j;\phi) = \frac{\partial^2 K(\phi)}{\partial \phi^i \partial \phi^j}.$$

To compute $E(S_i)$ etc. it is only necessary to know the function $K(\phi)$.

Properties of exponential families

Let $s(y) = (t(y), u(y))$ be a partition of the vector of natural statistics, where t has k components and u is $m - k$ dimensional. Consider the corresponding partition of the natural parameter $\phi = (\tau, \xi)$.

The density of a generic element of the family can be written as

$$f_Y(y; \tau, \xi) = \exp\{\tau^T t(y) + \xi^T u(y) - K(\tau, \xi)\} h(y).$$

Two key results hold which allow inference about components of the natural parameter, in the absence of knowledge about the other components.

Result 1

The family of marginal distributions of $U = u(Y)$ is an $m - k$ dimensional exponential family,

$$f_U(u; \tau, \xi) = \exp\{\xi^T u - K_\tau(\xi)\} h_\tau(u),$$

say.

Result 2

The family of conditional distributions of $T = t(Y)$ given $u(Y) = u$ is a k dimensional exponential family, and the conditional densities are free of ξ, so that

$$f_{T|U=u}(t \mid u; \tau) = \exp\{\tau^T t - K_u(\tau)\} h_u(t),$$

say.

Curved exponential families

In the above, both the natural statistic and the natural parameter lie in m-dimensional regions.

Sometimes, ϕ may be restricted to lie in a d-dimensional subspace, $d < m$.

This is most conveniently expressed by writing $\phi = \phi(\theta)$ where θ is a d-dimensional parameter.

We then have

$$f_Y(y;\theta) = h(y)\exp[s^T\phi(\theta) - K\{\phi(\theta)\}]$$

where $\theta \in \Omega_\theta \subset \mathbb{R}^d$.

We call this system an (m, d) exponential family, or curved exponential family, noting that we required that $(\phi^1, \dots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$.

Think of the case $m = 2, d = 1$: $\{\phi^1(\theta), \phi^2(\theta)\}$ describes a curve as θ varies.

Transformation families

A transformation family is defined by a group of transformations acting on the sample space which generates a family of distributions all of the same form, but with different values of the parameters.

A reminder

A group G is a mathematical structure having a binary operation $\circ$ such that

- if $g, g' \in G$, then $g \circ g' \in G$;
- if $g, g', g'' \in G$, then $(g \circ g') \circ g'' = g \circ (g' \circ g'')$;
- G contains an identity element e such that $e \circ g = g \circ e = g$, for each $g \in G$; and
- each $g \in G$ possesses an inverse $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

Present context

Concerned with group G of transformations acting on sample space $\mathcal{Y}$ of random variable Y, binary operation $\circ$ is composition of functions. Have $e(x) = x, (g_1 \circ g_2)(x) = g_1(g_2(x))$.

The group elements typically correspond to elements of a parameter space Ω_θ, transformation may be written as g_θ. The family of densities of $g_\theta(Y)$, for $g_\theta \in G$ is called a (group) transformation family.

Discussion

Setting $y \approx y'$ iff there is a $g \in G$ such that $y = g(y')$ gives an equivalence relation, which partitions $\mathcal{Y}$ into equivalence classes called orbits. These may be labelled by an index a, say.

Each y belongs to precisely one orbit, and can be represented by a (which identifies the orbit) and its position on the orbit.

Maximal invariant

We say that the statistic t is invariant to the action of the group G if its value does not depend on whether y or $g(y)$ was observed, for any $g \in G : t(y) = t(g(y))$.

The statistic t is maximal invariant if every other invariant statistic is a function of it, or equivalently, $t(y) = t(y')$ implies that $y' = g(y)$ for some $g \in G$.

Group action on Ω_θ

Typically, there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ.

Assume this.

Then the action of G on $\mathcal{Y}$ requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$.

Group action on $\mathcal{Y}$ induces group action on Ω_θ. If $\bar{G}$ denotes induced group, associated with each $g_\theta \in G$ is a $\bar{g}_\theta \in \bar{G}$, satisfying $\bar{g}_\theta(\phi) = \theta * \phi$.

Distribution constant statistic

If t is an invariant statistic, the distribution of $t(Y)$ is the same as that of $t(g(Y))$ for all g. If, as we assume, elements of G are identified with parameter values, this means distribution of $T = t(Y)$ does not depend on the parameter and is known in principle.

T is said to be distribution constant.

Equivariant statistic

A statistic $S = s(Y)$ defined on $\mathcal{Y}$ and taking values in the parameter space Ω_θ is said to be equivariant if $s(g_\theta(y)) = \bar{g}_\theta(s(y))$ for all $g_\theta \in G$ and $y \in \mathcal{Y}$.

Equivariant estimator

Often S is chosen to be an estimator of θ, and it is then called an equivariant estimator. An equivariant estimator can be used to construct a maximal invariant.

A maximal invariant

Consider $t(Y) = g^{-1}_{s(Y)}(Y)$.

This is invariant, since

$$\begin{aligned} t(g_\theta(y)) &= g^{-1}_{s(g_\theta(y))}(g_\theta(y)) = g^{-1}_{\bar{g}_\theta(s(y))}(g_\theta(y)) = g^{-1}_{\theta * s(y)}(g_\theta(y)) \\ &= g^{-1}_{s(y)}\{g_\theta^{-1}(g_\theta(y))\} = g^{-1}_{s(y)}(y) = t(y). \end{aligned}$$

If $t(y) = t(y')$, then $g^{-1}_{s(y)}(y) = g^{-1}_{s(y')}(y')$, and it follows that $y' = g_{s(y')} \circ g^{-1}_{s(y)}(y)$, which shows that $t(Y)$ is maximal invariant.

Location-scale model

Let $Y = \eta + \tau\epsilon$, where ϵ has a known density f, and the parameter $\theta = (\eta, \tau) \in \Omega_\theta = \mathbb{R} \times \mathbb{R}_+$. Define a group action by $g_\theta(y) = g_{(\eta,\tau)}(y) = \eta + \tau y$, so

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

The set of such transformations is closed with identity $g_{(0,1)}$. It is easy to check that $g_{(\eta,\tau)}$ has inverse $g_{(-\eta/\tau,\tau^{-1})}$. Hence, $G = \{g_{(\eta,\tau)} : (\eta, \tau) \in \mathbb{R} \times \mathbb{R}_+\}$ constitutes a group under the composition of functions operation $\circ$.

The action of $g_{(\eta,\tau)}$ on a random sample $Y = (Y_1, \ldots, Y_n)$ is $g_{(\eta,\tau)}(Y) = \eta + \tau Y$, with $\eta \equiv \eta 1_n$, where 1_n denotes the $n \times 1$ vector of 1's, and Y is written as an $n \times 1$ vector.

The induced group action on Ω_θ is given by $\bar{g}_{(\eta,\tau)}((\mu, \sigma)) \equiv (\eta, \tau) * (\mu, \sigma) = (\eta + \tau\mu, \tau\sigma)$.

The sample mean and standard deviation are equivariant, because with $s(Y) = (\bar{Y}, V^{1/2})$, where $V = (n-1)^{-1}\sum(Y_j - \bar{Y})^2$, we have

$$
\begin{aligned}
s(g_{(\eta,\tau)}(Y)) &= \left(\overline{\eta + \tau Y}, \left\{(n-1)^{-1}\sum(\eta + \tau Y_j - \overline{(\eta + \tau Y)})^2\right\}^{1/2}\right) \\
&= \left(\eta + \tau\bar{Y}, \left\{(n-1)^{-1}\sum(\eta + \tau Y_j - \eta - \tau\bar{Y})^2\right\}^{1/2}\right) \\
&= \left(\eta + \tau\bar{Y}, \tau V^{1/2}\right) \\
&= \bar{g}_{(\eta,\tau)}(s(Y)).
\end{aligned}
$$

Maximal invariant

A maximal invariant is $A = g^{-1}_{s(Y)}(Y)$, and the parameter corresponding to $g^{-1}_{s(Y)}$ is $(-\bar{Y}/V^{1/2}, V^{-1/2})$.

Hence a maximal invariant is the vector of residuals

$$A = (Y - \bar{Y})/V^{1/2} = \left(\frac{Y_1 - \bar{Y}}{V^{1/2}}, \ldots, \frac{Y_n - \bar{Y}}{V^{1/2}} \right)^T,$$

called the configuration.