Exponential family

"Natural parameter" links here. For the usage of this term in differential geometry, see differential geometry of curves.

In probability and statistics, an **exponential family** is a parametric set of probability distributions of a certain form, specified below. This special form is chosen for mathematical convenience, based on some useful algebraic properties, as well as for generality, as exponential families are in a sense very natural sets of distributions to consider. The term **exponential class** is sometimes used in place of "exponential family",[1] or the older term **Koopman-Darmois family**. The terms "distribution" and "family" are often used loosely: properly, *an* exponential family is a *set* of distributions, where the specific distribution varies with the parameter;[note 1] however, a parametric *family* of distributions is often referred to as "*a* distribution" (like "the normal distribution", meaning "the family of normal distributions"), and the set of all exponential families is sometimes loosely referred to as "the" exponential family.

The concept of exponential families is credited to[2] E. J. G. Pitman,[3] G. Darmois,[4] and B. O. Koopman[5] in 1935–36. Exponential families of distributions provides a general framework for selecting a possible alternative parameterisation of a parametric family of distributions, in terms of **natural parameters**, and for defining useful sample statistics, called the **natural sufficient statistics** of the family.

Contents

Definition

Most of the commonly used distributions form an exponential family or subset of an exponential family, listed in the subsection below. The subsections following it are a sequence of increasingly more general mathematical definitions of an exponential family. A casual reader may wish to restrict attention to the first and simplest definition, which corresponds to a single-parameter family of discrete or continuous probability distributions.

Examples of exponential family distributions

Exponential families include many of the most common distributions. Among many others, exponential families includes the following:

- normal
- exponential
- gamma
- chi-squared
- beta
- Dirichlet
- Bernoulli
- categorical
- Poisson
- Wishart
- inverse Wishart
- geometric

A number of common distributions are exponential families, but only when certain parameters are fixed and known. For example:

- binomial (with fixed number of trials)
- multinomial (with fixed number of trials)
- negative binomial (with fixed number of failures)

Notice that in each case, the parameters which must be fixed determine a limit on the size of observation values.

Examples of common distributions that are *not* exponential families are Student's *t*, most mixture distributions, and even the family of uniform distributions when the bounds are not fixed. See the section below on examples for more discussion.

Scalar parameter

A single-parameter exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form

$$f_X(x \mid \theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\theta))$$

where $T(x)$, $h(x)$, $\eta(\theta)$, and $A(\theta)$ are known functions.

An alternative, equivalent form often given is

$$f_X(x \mid \theta) = h(x)g(\theta)\exp(\eta(\theta) \cdot T(x))$$

or equivalently

$$f_X(x \mid \theta) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$$

The value θ is called the parameter of the family.

In addition, the support of $f_X(x \mid \theta)$ (i.e. the set of all x for which $f_X(x \mid \theta)$ is greater than 0) does not depend on θ.[6] This can be used to exclude a parametric family distribution from being an exponential family. For example, the Pareto distribution has a pdf which is defined for $x \geq x_m$ (x_m being the scale parameter) and its support, therefore, has a lower limit of x_m. Since the support of $f_{\alpha,x_m}(x)$ is dependent on the value of the parameter, the family of Pareto distributions does not form an exponential family of distributions.

Note that x is often a vector of measurements, in which case $T(x)$ may be a function from the space of possible values of x to the real numbers. More generally, $\eta(\theta)$ and $T(x)$ can each be vector-valued such that $\eta(\theta)' \cdot T(x)$ is real-valued.

If $\eta(\theta) = \theta$, then the exponential family is said to be in *canonical form*. By defining a transformed parameter $\eta = \eta(\theta)$, it is always possible to convert an exponential family to canonical form. The canonical form is non-unique, since $\eta(\theta)$ can be multiplied by any nonzero constant, provided that $T(x)$ is multiplied by that constant's reciprocal, or a constant c can be added to $\eta(\theta)$ and $h(x)$ multiplied by $\exp(-c \cdot T(x))$ to offset it.

Even when x is a scalar, and there is only a single parameter, the functions $\eta(\theta)$ and $T(x)$ can still be vectors, as described below.

Note also that the function $A(\theta)$, or equivalently $g(\theta)$, is automatically determined once the other functions have been chosen, since it must assume a form that causes the distribution to be normalized (sum or integrate to one over the entire domain). Furthermore, both of these functions can always be written as functions of η, even when $\eta(\theta)$ is not a one-to-one function, i.e. two or more different values of θ map to the same value of $\eta(\theta)$, and hence $\eta(\theta)$ cannot be inverted. In such a case, all values of θ mapping to the same $\eta(\theta)$ will also have the same value for $A(\theta)$ and $g(\theta)$.

Factorization of the variables involved

What is important to note, and what characterizes all exponential family variants, is that the parameter(s) and the observation variable(s) must factorize (can be separated into products each of which involves only one type of variable), either directly or within either part (the base or exponent) of an exponentiation operation. Generally, this means that all of the factors constituting the density or mass function must be of one of the following forms:

$$f(x), g(\theta), c^{f(x)}, c^{g(\theta)}, [f(x)]^c, [g(\theta)]^c, [f(x)]^{g(\theta)}, [g(\theta)]^{f(x)}, [f(x)]^{h(x)g(\theta)}, \text{ or } [g(\theta)]^{h(x)j(\theta)},$$

where f and h are arbitrary functions of x; g and j are arbitrary functions of θ; and c is an arbitrary "constant" expression (i.e. an expression not involving x or θ).

There are further restrictions on how many such factors can occur. For example, the two expressions:

$$[f(x)g(\theta)]^{h(x)j(\theta)}, \qquad [f(x)]^{h(x)j(\theta)}[g(\theta)]^{h(x)j(\theta)},$$

are the same, i.e. a product of two "allowed" factors. However, when rewritten into the factorized form,

$$[f(x)g(\theta)]^{h(x)j(\theta)} = [f(x)]^{h(x)j(\theta)}[g(\theta)]^{h(x)j(\theta)} = e^{[h(x)\,\log f(x)]j(\theta)+h(x)[j(\theta)\,\log g(\theta)]},$$

it can be seen that it cannot be expressed in the required form. (However, a form of this sort is a member of a *curved exponential family*, which allows multiple factorized terms in the exponent.)

To see why an expression of the form

$$[f(x)]^{g(\theta)}$$

qualifies, note that

$$[f(x)]^{g(\theta)} = e^{g(\theta)\,\log f(x)}$$

and hence factorizes inside of the exponent. Similarly,

$$[f(x)]^{h(x)g(\theta)} = e^{h(x)g(\theta)\,\log f(x)} = e^{[h(x)\,\log f(x)]g(\theta)}$$

and again factorizes inside of the exponent.

Note also that a factor consisting of a sum where both types of variables are involved (e.g. a factor of the form $1 + f(x)g(\theta)$) cannot be factorized in this fashion (except in some cases where occurring directly in an exponent); this is why, for example, the Cauchy distribution and Student's *t* distribution are not exponential families.

Vector parameter

The definition in terms of one *real-number* parameter can be extended to one *real-vector* parameter

$$\boldsymbol\theta = (\theta_1, \theta_2, \ldots, \theta_s)^T.$$

A family of distributions is said to belong to a vector exponential family if the probability density function (or probability mass function, for discrete distributions) can be written as

$$f_X(x \mid \boldsymbol\theta) = h(x)\exp\left(\sum_{i=1}^{s}\eta_i(\boldsymbol\theta)T_i(x) - A(\boldsymbol\theta)\right)$$

Or in a more compact form,

$$f_X(x \mid \boldsymbol\theta) = h(x)\exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(x) - A(\boldsymbol\theta)\Big)$$

This form writes the sum as a dot product of vector-valued functions $\boldsymbol\eta(\boldsymbol\theta)$ and $\mathbf{T}(x)$.

An alternative, equivalent form often seen is

$$f_X(x \mid \boldsymbol\theta) = h(x)g(\boldsymbol\theta)\exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(x)\Big)$$

As in the scalar valued case, the exponential family is said to be in canonical form if

$$\forall i: \quad \eta_i(\boldsymbol\theta) = \theta_i.$$

A vector exponential family is said to be *curved* if the dimension of

$$\boldsymbol\theta = (\theta_1, \theta_2, \ldots, \theta_d)^T$$

is less than the dimension of the vector

$$\boldsymbol\eta(\boldsymbol\theta) = (\eta_1(\boldsymbol\theta), \eta_2(\boldsymbol\theta), \ldots, \eta_s(\boldsymbol\theta))^T.$$

That is, if the *dimension* of the parameter vector is less than the *number of functions* of the parameter vector in the above representation of the probability density function. Note that most common distributions in the exponential family are *not* curved, and many algorithms designed to work with any exponential family implicitly or explicitly assume that the distribution is not curved.

Note that, as in the above case of a scalar-valued parameter, the function $A(\boldsymbol\theta)$ or equivalently $g(\boldsymbol\theta)$ is automatically determined once the other functions have been chosen, so that the entire distribution is normalized. In addition, as above, both of these functions can always be written as functions of $\boldsymbol\eta$, regardless of the form of the transformation that generates $\boldsymbol\eta$ from $\boldsymbol\theta$. Hence an exponential family in its "natural form" (parametrized by its natural parameter) looks like

$$f_X(x \mid \boldsymbol\eta) = h(x) \exp\Big(\boldsymbol\eta \cdot \mathbf{T}(x) - A(\boldsymbol\eta)\Big)$$

or equivalently

$$f_X(x \mid \boldsymbol\eta) = h(x) g(\boldsymbol\eta) \exp\Big(\boldsymbol\eta \cdot \mathbf{T}(x)\Big)$$

Note that the above forms may sometimes be seen with $\boldsymbol\eta^T \mathbf{T}(x)$ in place of $\boldsymbol\eta \cdot \mathbf{T}(x)$. These are exactly equivalent formulations, merely using different notation for the dot product.

Vector parameter, vector variable

The vector-parameter form over a single scalar-valued random variable can be trivially expanded to cover a joint distribution over a vector of random variables. The resulting distribution is simply the same as the above distribution for a scalar-valued random variable with each occurrence of the scalar *x* replaced by the vector

$$\mathbf{x} = (x_1, x_2, \cdots, x_k)\,.$$

Note that the dimension *k* of the random variable need not match the dimension *d* of the parameter vector, nor (in the case of a curved exponential function) the dimension *s* of the natural parameter $\boldsymbol\eta$ and sufficient statistic *T*(**x**).

The distribution in this case is written as

$$f_X(\mathbf{x} \mid \boldsymbol\theta) = h(\mathbf{x}) \exp\left(\sum_{i=1}^{s} \eta_i(\boldsymbol\theta) T_i(\mathbf{x}) - A(\boldsymbol\theta)\right)$$

Or more compactly as

$$f_X(\mathbf{x} \mid \boldsymbol\theta) = h(\mathbf{x}) \exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(\mathbf{x}) - A(\boldsymbol\theta)\Big)$$

Or alternatively as

$$f_X(\mathbf{x} \mid \boldsymbol\theta) = h(\mathbf{x})\ g(\boldsymbol\theta)\ \exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(\mathbf{x})\Big)$$

Measure-theoretic formulation

We use cumulative distribution functions (cdf) in order to encompass both discrete and continuous distributions.

Suppose *H* is a non-decreasing function of a real variable. Then Lebesgue–Stieltjes integrals with respect to *dH*(*x*) are integrals with respect to the "reference measure" of the exponential family generated by *H*.

Any member of that exponential family has cumulative distribution function

$$dF(\mathbf{x} \mid \boldsymbol\eta) = e^{\boldsymbol\eta^{\mathrm{T}} \mathbf{T}(\mathbf{x}) - A(\boldsymbol\eta)}\, dH(\mathbf{x}).$$

If *F* is a continuous distribution with a density, one can write *dF*(*x*) = *f*(*x*) *dx*.

H(*x*) is a Lebesgue–Stieltjes integrator for the *reference measure*. When the reference measure is finite, it can be normalized and *H* is actually the cumulative distribution function of a probability distribution. If *F* is absolutely continuous with a density, then so is *H*, which can then be written *dH*(*x*) = *h*(*x*) *dx*. If *F* is discrete, then *H* is a step function (with steps on the support of *F*).

Interpretation

In the definitions above, the functions *T*(*x*), *η*(*θ*) and *A*(*η*) were apparently arbitrarily defined. However, these functions play a significant role in the resulting probability distribution.

- *T*(*x*) is a *sufficient statistic* of the distribution. For exponential families, the sufficient statistic is a function of the data that holds all information the data $\boldsymbol{x}$ provides with regard to the unknown parameter values. This means that, for any data sets $\boldsymbol{x}$ and $\boldsymbol{y}$, the likelihood ratio is the same (that is, $\frac{f(x;\theta_1)}{f(x;\theta_2)} = \frac{f(y;\theta_1)}{f(y;\theta_2)}$) if $T(x) = T(y)$. This is true even if *x* and *y* are quite different—that is, $d(x, y) > 0$. The dimension of *T*(*x*) equals the number of parameters of *θ* and encompasses all of the information regarding the data related to the parameter *θ*. The sufficient statistic of a set of independent identically distributed data observations is simply the sum of individual sufficient statistics, and encapsulates all the information needed to describe the posterior distribution of the parameters, given the data (and hence to derive any desired estimate of the parameters). This important property is further discussed below.
- *η* is called the *natural parameter*. The set of values of *η* for which the function $f_X(x;\theta)$ is finite is called the *natural parameter space*. It can be shown that the natural parameter space is always convex.
- *A*(*η*) is called the **log-partition function** because it is the logarithm of a normalization factor, without which $f_X(x;\theta)$ would not be a probability distribution ("partition function" is often used in statistics as a synonym of "normalization factor"):

$$A(\eta) = \log\left(\int_x h(x) \exp(\eta(\theta) \cdot T(x))\, \mathrm{d}x\right)$$

The function *A* is important in its own right, because the mean, variance and other moments of the sufficient statistic *T*(*x*) can be derived simply by differentiating *A*(*η*). For example, because log(*x*) is one of the components of the sufficient statistic of the gamma distribution, $\mathbf{E}[\log x]$ can be easily determined for this distribution using *A*(*η*). Technically, this is true because

$$K(u \mid \eta) = A(\eta + u) - A(\eta),$$

is the cumulant generating function of the sufficient statistic.

Properties

Exponential families have a large number of properties that make them extremely useful for statistical analysis. In many cases, it can be shown that, except in a few exceptional cases, *only* exponential families have these properties. Examples:

- Exponential families have sufficient statistics that can summarize arbitrary amounts of independent identically distributed data using a fixed number of values.
- Exponential families have conjugate priors, an important property in Bayesian statistics.
- The posterior predictive distribution of an exponential-family random variable with a conjugate prior can always be written in closed form (provided that the normalizing factor of the exponential-family distribution can itself be written in closed form). Note that these distributions are often not themselves exponential families. Common examples of non-exponential families arising from exponential ones are the Student's *t*-distribution, beta-binomial distribution and Dirichlet-multinomial distribution.
- In the mean-field approximation in variational Bayes (used for approximating the posterior distribution in large Bayesian networks), the best approximating posterior distribution of an exponential-family node (a node is a random variable in the context of Bayesian networks) with a conjugate prior is in the same family as the node.[7]

Examples

It is critical, when considering the examples in this section, to remember the discussion above about what it means to say that a "distribution" is an exponential family, and in particular to keep in mind that the set of parameters that are allowed to vary is critical in determining whether a "distribution" is or is not an exponential family.

The normal, exponential, log-normal, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson, geometric, inverse Gaussian, von Mises and von Mises-Fisher distributions are all exponential families.

Some distributions are exponential families only if some of their parameters are held fixed. The family of Pareto distributions with a fixed minimum bound x_m form an exponential family. The families of binomial and multinomial distributions with fixed number of trials n but unknown probability parameter(s) are exponential families. The family of negative binomial distributions with fixed number of failures (a.k.a. stopping-time parameter) r is an exponential family. However, when any of the above-mentioned fixed parameters are allowed to vary, the resulting family is not an exponential family.

As mentioned above, as a general rule, the support of an exponential family must remain the same across all parameter settings in the family. This is why the above cases (e.g. binomial with varying number of trials, Pareto with varying minimum bound) are not exponential families — in all of the cases, the parameter in question affects the support (particularly, changing the minimum or maximum possible value). For similar reasons, neither the discrete uniform distribution nor continuous uniform distribution are exponential families as one or both bounds vary. If both bounds are held fixed, the result is a single distribution; this can be considered a zero-dimensional exponential family, and is the only zero-dimensional exponential family with a given support, but this is generally considered too trivial to consider as a family.

The Weibull distribution with fixed shape parameter k is an exponential family. Unlike in the previous examples, the shape parameter does not affect the support; the fact that allowing it to vary makes the Weibull non-exponential is due rather to the particular form of the Weibull's probability density function (k appears in the exponent of an exponent).

In general, distributions that result from a finite or infinite mixture of other distributions, e.g. mixture model densities and compound probability distributions, are *not* exponential families. Examples are typical Gaussian mixture models as well as many heavy-tailed distributions that result from compounding (i.e. infinitely mixing) a distribution with a prior distribution over one of its parameters, e.g. the Student's t-distribution (compounding a normal distribution over a gamma-distributed precision prior), and the beta-binomial and Dirichlet-multinomial distributions. Other examples of distributions that are not exponential families are the F-distribution, Cauchy distribution, hypergeometric distribution and logistic distribution.

Following are some detailed examples of the representation of some useful distribution as exponential families.

Normal distribution: unknown mean, known variance

As a first example, consider a random variable distributed normally with unknown mean μ and *known* variance σ^2. The probability density function is then

$$f_\sigma(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

This is a single-parameter exponential family, as can be seen by setting

$$\begin{aligned}
h_\sigma(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)} \\
T_\sigma(x) &= \frac{x}{\sigma} \\
A_\sigma(\mu) &= \frac{\mu^2}{2\sigma^2} \\
\eta_\sigma(\mu) &= \frac{\mu}{\sigma}.
\end{aligned}$$

If $\sigma = 1$ this is in canonical form, as then $\eta(\mu) = \mu$.

Normal distribution: unknown mean and unknown variance

Next, consider the case of a normal distribution with unknown mean and unknown variance. The probability density function is then

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This is an exponential family which can be written in canonical form by defining

$$\begin{aligned}
\boldsymbol\eta &= \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^{\mathrm{T}} \\
h(x) &= \frac{1}{\sqrt{2\pi}} \\
T(x) &= \left(x, x^2\right)^{\mathrm{T}} \\
A(\boldsymbol\eta) &= \frac{\mu^2}{2\sigma^2} + \log|\sigma| = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left|\frac{1}{2\eta_2}\right|
\end{aligned}$$

Binomial distribution

As an example of a discrete exponential family, consider the binomial distribution with *known* number of trials n. The probability mass function for this distribution is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, 2, \ldots, n\}.$$

This can equivalently be written as

$$f(x) = \binom{n}{x} \exp\left(x \log\left(\frac{p}{1-p}\right) + n \log(1-p)\right),$$

which shows that the binomial distribution is an exponential family, whose natural parameter is

$$\eta = \log\frac{p}{1-p}.$$

This function of p is known as logit.

Table of distributions

The following table shows how to rewrite a number of common distributions as exponential-family distributions with natural parameters. Refer to the flashcards[8] for main exponential families.

For a scalar variable and scalar parameter, the form is as follows:

$$f_X(x \mid \theta) = h(x) \exp\Big(\eta(\theta) T(x) - A(\eta)\Big)$$

For a scalar variable and vector parameter:

$$f_X(x \mid \boldsymbol\theta) = h(x) \exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(x) - A(\boldsymbol\eta)\Big)$$
$$f_X(x \mid \boldsymbol\theta) = h(x) g(\boldsymbol\theta) \exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(x)\Big)$$

For a vector variable and vector parameter:

$$f_X(\mathbf{x} \mid \boldsymbol\theta) = h(\mathbf{x}) \exp\Big(\boldsymbol\eta(\boldsymbol\theta) \cdot \mathbf{T}(\mathbf{x}) - A(\boldsymbol\eta)\Big)$$

The above formulas choose the functional form of the exponential-family with a log-partition function $A(\boldsymbol\eta)$. The reason for this is so that the moments of the sufficient statistics can be calculated easily, simply by differentiating this function. Alternative forms involve either parameterizing this function in terms of the normal parameter $\boldsymbol\theta$ instead of the natural parameter, and/or using a factor $g(\boldsymbol\eta)$ outside of the exponential. The relation between the latter and the former is:

$$A(\boldsymbol\eta) = -\log g(\boldsymbol\eta)$$
$$g(\boldsymbol\eta) = e^{-A(\boldsymbol\eta)}$$

To convert between the representations involving the two types of parameter, use the formulas below for writing one type of parameter in terms of the other.

Distribution	Parameter(s) $\boldsymbol{\theta}$	Natural parameter(s) $\boldsymbol{\eta}$	Inverse parameter mapping	Base measure $h(\boldsymbol{x})$	Sufficient statistic $T(\boldsymbol{x})$	Log-partition $A(\boldsymbol{\eta})$	Log-partition $A(\boldsymbol{\theta})$
Bernoulli distribution	p	$\log \frac{p}{1-p}$ ▪ This is the logit function.	$\frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}}$ ▪ This is the logistic function.	1	x	$\log(1+e^{\eta})$	$-\log(1-p)$
binomial distribution with known number of trials *n*	p	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}}$	$\binom{n}{x}$	x	$n\log(1+e^{\eta})$	$-n\log(1-p)$
Poisson distribution	λ	$\log \lambda$	e^{η}	$\frac{1}{x!}$	x	e^{η}	λ
negative binomial distribution with known number of failures *r*	p	$\log p$	e^{η}	$\binom{x+r-1}{x}$	x	$-r\log(1-e^{\eta})$	$-r\log(1-p)$
exponential distribution	λ	$-\lambda$	$-\eta$	1	x	$-\log(-\eta)$	$-\log\lambda$
Pareto distribution with known minimum value x_{m}	α	$-\alpha - 1$	$-1-\eta$	1	$\log x$	$-\log(-1-\eta) + (1+\eta)\log x_{\mathrm{m}}$	$-\log\alpha - \alpha\log x_{\mathrm{m}}$
Weibull distribution with known shape *k*	λ	$-\frac{1}{\lambda^k}$	$(-\eta)^{-\frac{1}{k}}$	x^{k-1}	x^k	$-\log(-\eta) - \log k$	$k\log\lambda - \log k$
Laplace distribution with known mean μ	b	$-\frac{1}{b}$	$-\frac{1}{\eta}$	1	$\lvert x-\mu\rvert$	$\log\left(-\frac{2}{\eta}\right)$	$\log 2b$
chi-squared distribution	ν	$\frac{\nu}{2} - 1$	$2(\eta+1)$	$e^{-\frac{x}{2}}$	$\log x$	$\log\Gamma(\eta+1) + (\eta+1)\log 2$	$\log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log 2$
normal distribution known variance	μ	$\frac{\mu}{\sigma}$	$\sigma\eta$	$\frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$	$\frac{x}{\sigma}$	$\frac{\eta^2}{2}$	$\frac{\mu^2}{2\sigma^2}$
normal distribution	μ,σ^2	$\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$	$\frac{\mu^2}{2\sigma^2} + \log\sigma$
lognormal distribution	μ,σ^2	$\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}x}$	$\begin{bmatrix} \log x \\ (\log x)^2 \end{bmatrix}$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$	$\frac{\mu^2}{2\sigma^2} + \log\sigma$
inverse Gaussian distribution	μ,λ	$\begin{bmatrix} -\frac{\lambda}{2\mu^2} \\ -\frac{\lambda}{2} \end{bmatrix}$	$\begin{bmatrix} \sqrt{\frac{\eta_2}{\eta_1}} \\ -2\eta_2 \end{bmatrix}$	$\frac{1}{\sqrt{2\pi}x^{\frac{3}{2}}}$	$\begin{bmatrix} x \\ \frac{1}{x} \end{bmatrix}$	$2\sqrt{\eta_1\eta_2} - \frac{1}{2}\log(-2\eta_2)$	$-\frac{\lambda}{\mu} - \frac{1}{2}\log\lambda$
gamma distribution	α,β	$\begin{bmatrix} \alpha - 1 \\ -\beta \end{bmatrix}$	$\begin{bmatrix} \eta_1 + 1 \\ -\eta_2 \end{bmatrix}$	1	$\begin{bmatrix} \log x \\ x \end{bmatrix}$	$\log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2)$	$\log\Gamma(\alpha) - \alpha\log\beta$
	k, *θ*	$\begin{bmatrix} k - 1 \\ -\frac{1}{\theta} \end{bmatrix}$	$\begin{bmatrix} \eta_1 + 1 \\ -\frac{1}{\eta_2} \end{bmatrix}$				$\log\Gamma(k) + k\log\theta$
inverse gamma distribution	α,β	$\begin{bmatrix} -\alpha - 1 \\ -\beta \end{bmatrix}$	$\begin{bmatrix} -\eta_1 - 1 \\ -\eta_2 \end{bmatrix}$	1	$\begin{bmatrix} \log x \\ \frac{1}{x} \end{bmatrix}$	$\log\Gamma(-\eta_1-1) - (-\eta_1-1)\log(-\eta_2)$	$\log\Gamma(\alpha) - \alpha\log\beta$
scaled inverse chi-squared distribution	ν,σ^2	$\begin{bmatrix} -\frac{\nu}{2} - 1 \\ -\frac{\nu\sigma^2}{2} \end{bmatrix}$	$\begin{bmatrix} -2(\eta_1+1) \\ \frac{\eta_2}{\eta_1+1} \end{bmatrix}$	1	$\begin{bmatrix} \log x \\ \frac{1}{x} \end{bmatrix}$	$\log\Gamma(-\eta_1-1) - (-\eta_1-1)\log(-\eta_2)$	$\log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\log\frac{\nu\sigma^2}{2}$
beta distribution	α,β	$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$	$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$	$\frac{1}{x(1-x)}$	$\begin{bmatrix} \log x \\ \log(1-x) \end{bmatrix}$	$\log\Gamma(\eta_1) + \log\Gamma(\eta_2) - \log\Gamma(\eta_1+\eta_2)$	$\log\Gamma(\alpha) + \log\Gamma(\beta) - \log\Gamma(\alpha+\beta)$
multivariate normal distribution	$\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$	$\begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 \\ -\frac{1}{2}\boldsymbol{\eta}_2^{-1} \end{bmatrix}$	$(2\pi)^{-\frac{k}{2}}$	$\begin{bmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^{\mathrm{T}} \end{bmatrix}$	$-\frac{1}{4}\boldsymbol{\eta}_1^{\mathrm{T}}\boldsymbol{\eta}_2^{-1}\boldsymbol{\eta}_1 - \frac{1}{2}\log\lvert -2\boldsymbol{\eta}_2\rvert$	$\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log\lvert\boldsymbol{\Sigma}\rvert$
categorical distribution (variant 1)	p$_1$,…,p$_k$ where $\sum_{i=1}^{k} p_i = 1$	$\begin{bmatrix} \log p_1 \\ \vdots \\ \log p_k \end{bmatrix}$	$\begin{bmatrix} e^{\eta_1} \\ \vdots \\ e^{\eta_k} \end{bmatrix}$ where $\sum_{i=1}^{k} e^{\eta_i} = 1$	1	$\begin{bmatrix} [x=1] \\ \vdots \\ [x=k] \end{bmatrix}$ ▪ $[x=i]$ is the Iverson bracket (1 if $x=i$, 0 otherwise).	0	0
categorical distribution	p$_1$,…,p$_k$			1		0	0

(variant 2)	where $\sum_{i=1}^{k} p_i = 1$	$\begin{bmatrix} \log p_1 + C \\ \vdots \\ \log p_k + C \end{bmatrix}$	$\begin{bmatrix} \frac{1}{C} e^{\eta_1} \\ \vdots \\ \frac{1}{C} e^{\eta_k} \end{bmatrix} = \begin{bmatrix} \frac{e^{\eta_1}}{\sum_{i=1}^{k} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_k}}{\sum_{i=1}^{k} e^{\eta_i}} \end{bmatrix}$ where $\sum_{i=1}^{k} e^{\eta_i} = C$		$\begin{bmatrix} [x = 1] \\ \vdots \\ [x = k] \end{bmatrix}$ ▪ $[x = i]$ is the Iverson bracket (1 if $x = i$, 0 otherwise).		
categorical distribution (variant 3)	p1,...,pk where $p_k = 1 - \sum_{i=1}^{k-1} p_i$	$\begin{bmatrix} \log \frac{p_1}{p_k} \\ \vdots \\ \log \frac{p_{k-1}}{p_k} \\ 0 \end{bmatrix} = \begin{bmatrix} \log \frac{p_1}{1 - \sum_{i=1}^{k-1} p_i} \\ \vdots \\ \log \frac{p_{k-1}}{1 - \sum_{i=1}^{k-1} p_i} \\ 0 \end{bmatrix}$ ▪ This is the inverse softmax function, a generalization of the logit function.	$\begin{bmatrix} \frac{e^{\eta_1}}{\sum_{i=1}^{k} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_k}}{\sum_{i=1}^{k} e^{\eta_i}} \end{bmatrix} = \begin{bmatrix} \frac{e^{\eta_1}}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_{k-1}}}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \\ \frac{1}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \end{bmatrix}$ ▪ This is the softmax function, a generalization of the logistic function.	1	$\begin{bmatrix} [x = 1] \\ \vdots \\ [x = k] \end{bmatrix}$ ▪ $[x = i]$ is the Iverson bracket (1 if $x = i$, 0 otherwise).	$\log\left(\sum_{i=1}^{k} e^{\eta_i}\right) = \log\left(1 + \sum_{i=1}^{k-1} e^{\eta_i}\right)$	$-\log p_k = -\log\left(1 - \sum_{i=1}^{k-1} p_i\right)$
multinomial distribution (variant 1) with known number of trials *n*	p1,...,pk where $\sum_{i=1}^{k} p_i = 1$	$\begin{bmatrix} \log p_1 \\ \vdots \\ \log p_k \end{bmatrix}$	$\begin{bmatrix} e^{\eta_1} \\ \vdots \\ e^{\eta_k} \end{bmatrix}$ where $\sum_{i=1}^{k} e^{\eta_i} = 1$	$\frac{n!}{\prod_{i=1}^{k} x_i!}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$	0	0
multinomial distribution (variant 2) with known number of trials *n*	p1,...,pk where $\sum_{i=1}^{k} p_i = 1$	$\begin{bmatrix} \log p_1 + C \\ \vdots \\ \log p_k + C \end{bmatrix}$	$\begin{bmatrix} \frac{1}{C} e^{\eta_1} \\ \vdots \\ \frac{1}{C} e^{\eta_k} \end{bmatrix} = \begin{bmatrix} \frac{e^{\eta_1}}{\sum_{i=1}^{k} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_k}}{\sum_{i=1}^{k} e^{\eta_i}} \end{bmatrix}$ where $\sum_{i=1}^{k} e^{\eta_i} = C$	$\frac{n!}{\prod_{i=1}^{k} x_i!}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$	0	0
multinomial distribution (variant 3) with known number of trials *n*	p1,...,pk where $p_k = 1 - \sum_{i=1}^{k-1} p_i$	$\begin{bmatrix} \log \frac{p_1}{p_k} \\ \vdots \\ \log \frac{p_{k-1}}{p_k} \\ 0 \end{bmatrix} = \begin{bmatrix} \log \frac{p_1}{1 - \sum_{i=1}^{k-1} p_i} \\ \vdots \\ \log \frac{p_{k-1}}{1 - \sum_{i=1}^{k-1} p_i} \\ 0 \end{bmatrix}$	$\begin{bmatrix} \frac{e^{\eta_1}}{\sum_{i=1}^{k} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_k}}{\sum_{i=1}^{k} e^{\eta_i}} \end{bmatrix} = \begin{bmatrix} \frac{e^{\eta_1}}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \\ \vdots \\ \frac{e^{\eta_{k-1}}}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \\ \frac{1}{1 + \sum_{i=1}^{k-1} e^{\eta_i}} \end{bmatrix}$	$\frac{n!}{\prod_{i=1}^{k} x_i!}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}$	$n \log\left(\sum_{i=1}^{k} e^{\eta_i}\right) = n \log\left(1 + \sum_{i=1}^{k-1} e^{\eta_i}\right)$	$-n \log p_k = -n \log\left(1 - \sum_{i=1}^{k-1} p_i\right)$
Dirichlet distribution	α1,...,αk	$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}$	$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_k \end{bmatrix}$	$\frac{1}{\prod_{i=1}^{k} x_i}$	$\begin{bmatrix} \log x_1 \\ \vdots \\ \log x_k \end{bmatrix}$	$\sum_{i=1}^{k} \log \Gamma(\eta_i) - \log \Gamma\left(\sum_{i=1}^{k} \eta_i\right)$	$\sum_{i=1}^{k} \log \Gamma(\alpha_i) - \log \Gamma\left(\sum_{i=1}^{k} \alpha_i\right)$
Wishart distribution	**V**,n	$\begin{bmatrix} -\frac{1}{2}\mathbf{V}^{-1} \\ \frac{n-p-1}{2} \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}{\eta_1}^{-1} \\ 2\eta_2 + p + 1 \end{bmatrix}$	1	$\begin{bmatrix} \mathbf{X} \\ \log\lvert\mathbf{X}\rvert \end{bmatrix}$	$-\left(\eta_2 + \frac{p+1}{2}\right) \log\lvert -\eta_1 \rvert$	$\frac{n}{2}(p \log 2 + \log\lvert\mathbf{V}\rvert) + \log \Gamma_p\left(\frac{n}{2}\right)$

						$+\log\Gamma_p\left(\eta_2+\frac{p+1}{2}\right)=$ $-\frac{n}{2}\log\lvert-\boldsymbol{\eta}_1\rvert+\log\Gamma_p\left(\frac{n}{2}\right)=$ $\left(\eta_2+\frac{p+1}{2}\right)(p\log 2+\log\lvert\mathbf{V}\rvert)$ $+\log\Gamma_p\left(\eta_2+\frac{p+1}{2}\right)$ ▪ Three variants with different parameterizations are given, to facilitate computing moments of the sufficient statistics.	
	NOTE: Uses the fact that $\mathbf{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})=\mathbf{vec}(\mathbf{A})\cdot\mathbf{vec}(\mathbf{B})$, i.e. the trace of a matrix product is much like a dot product. The matrix parameters are assumed to be vectorized (laid out in a vector) when inserted into the exponential form. Also, **V** and **X** are symmetric, so e.g. $\mathbf{V}^{\mathrm{T}}=\mathbf{V}$.						
inverse Wishart distribution	**Ψ**,m	$\begin{bmatrix}-\frac{1}{2}\boldsymbol{\Psi}\\-\frac{m+p+1}{2}\end{bmatrix}$	$\begin{bmatrix}-2\boldsymbol{\eta}_1\\-(2\eta_2+p+1)\end{bmatrix}$	1	$\begin{bmatrix}\mathbf{X}^{-1}\\\log\lvert\mathbf{X}\rvert\end{bmatrix}$	$\left(\eta_2+\frac{p+1}{2}\right)\log\lvert-\boldsymbol{\eta}_1\rvert$ $+\log\Gamma_p\left(-\left(\eta_2+\frac{p+1}{2}\right)\right)=$ $-\frac{m}{2}\log\lvert-\boldsymbol{\eta}_1\rvert+\log\Gamma_p\left(\frac{m}{2}\right)=$ $-\left(\eta_2+\frac{p+1}{2}\right)(p\log 2-\log\lvert\boldsymbol{\Psi}\rvert)$ $+\log\Gamma_p\left(-\left(\eta_2+\frac{p+1}{2}\right)\right)$	$\frac{m}{2}(p\log 2-\log\lvert\boldsymbol{\Psi}\rvert)+\log\Gamma_p\left(\frac{m}{2}\right)$
normal-gamma distribution	α,β,μ,λ	$\begin{bmatrix}\alpha-\frac{1}{2}\\-\beta-\frac{\lambda\mu^2}{2}\\\lambda\mu\\-\frac{\lambda}{2}\end{bmatrix}$	$\begin{bmatrix}\eta_1+\frac{1}{2}\\-\eta_2+\frac{\eta_3^2}{4\eta_4}\\-\frac{\eta_3}{2\eta_4}\\-2\eta_4\end{bmatrix}$	$\frac{1}{\sqrt{2\pi}}$	$\begin{bmatrix}\log\tau\\\tau\\\tau x\\\tau x^2\end{bmatrix}$	$\log\Gamma\left(\eta_1+\frac{1}{2}\right)-\frac{1}{2}\log(-2\eta_4)-$ $-\left(\eta_1+\frac{1}{2}\right)\log\left(-\eta_2+\frac{\eta_3^2}{4\eta_4}\right)$	$\log\Gamma(\alpha)-\alpha\log\beta-\frac{1}{2}\log\lambda$

The three variants of the categorical distribution and multinomial distribution are due to the fact that the parameters $\boldsymbol{p_i}$ are constrained, such that

$$\sum_{i=1}^{k} p_i = 1.$$

Thus, there are only k−1 independent parameters.

- Variant 1 uses k natural parameters with a simple relation between the standard and natural parameters; however, only k−1 of the natural parameters are independent, and the set of k natural parameters is nonidentifiable. The constraint on the usual parameters translates to a similar constraint on the natural parameters.
- Variant 2 demonstrates the fact that the entire set of natural parameters is nonidentifiable: Adding any constant value to the natural parameters has no effect on the resulting distribution. However, by using the constraint on the natural parameters, the formula for the normal parameters in terms of the natural parameters can be written in a way that is independent on the constant that is added.
- Variant 3 shows how to make the parameters identifiable in a convenient way by setting $C=-\log p_k$. This effectively "pivots" around p_k and causes the last natural parameter to have the constant value of 0. All the remaining formulas are written in a way that does not access p_k, so that effectively the model has only k−1 parameters, both of the usual and natural kind.

Note also that variants 1 and 2 are not actually standard exponential families at all. Rather they are *curved exponential families*, i.e. there are k−1 independent parameters embedded in a k-dimensional parameter space.[9] Many of the standard results for exponential families do not apply to curved exponential families. An example is the log-partition function $A(x)$, which has the value of 0 in the curved cases. In standard exponential families, the derivatives of this function correspond to the moments (more technically, the cumulants) of the sufficient statistics, e.g. the mean and variance. However, a value of 0 suggests that the mean and variance of all the sufficient statistics are uniformly 0, whereas in fact the mean of the ith sufficient statistic should be p_i. (This does emerge correctly when using the form of $A(x)$ in variant 3.)

Moments and cumulants of the sufficient statistic

Normalization of the distribution

We start with the normalization of the probability distribution. In general, any non-negative function $f(x)$ that serves as the kernel of a probability distribution (the part encoding all dependence on x) can be made into a proper distribution by normalizing: i.e.

$$p(x)=\frac{1}{Z}f(x)$$

where

$$Z=\int_x f(x)\,dx.$$

The factor Z is sometimes termed the *normalizer* or *partition function*, based on an analogy to statistical physics.

In the case of an exponential family where

$$p(x;\boldsymbol{\eta})=g(\boldsymbol{\eta})h(x)e^{\boldsymbol{\eta}\cdot\mathbf{T}(x)},$$

the kernel is

$$K(x)=h(x)e^{\boldsymbol{\eta}\cdot\mathbf{T}(x)}$$

and the partition function is

$$Z=\int_x h(x)e^{\boldsymbol{\eta}\cdot\mathbf{T}(x)}\,dx.$$

Since the distribution must be normalized, we have

$$1=\int_x g(\boldsymbol{\eta})h(x)e^{\boldsymbol{\eta}\cdot\mathbf{T}(x)}\,dx=g(\boldsymbol{\eta})\int_x h(x)e^{\boldsymbol{\eta}\cdot\mathbf{T}(x)}\,dx=g(\boldsymbol{\eta})Z.$$

In other words,

$$g(\boldsymbol{\eta})=\frac{1}{Z}$$

or equivalently

$$A(\boldsymbol{\eta})=-\log g(\boldsymbol{\eta})=\log Z.$$

This justifies calling A the *log-normalizer* or *log-partition function*.

Moment-generating function of the sufficient statistic

Now, the moment-generating function of $T(x)$ is

$$M_T(u) \equiv E[e^{u^{\mathrm{t}}T(x)} \mid \eta] = \int_x h(x) e^{(\eta+u)^{\mathrm{t}}T(x)-A(\eta)}\,dx = e^{A(\eta+u)-A(\eta)}$$

where t means transpose, proving the earlier statement that

$$K(u \mid \eta) = A(\eta+u) - A(\eta)$$

is the cumulant generating function for T.

An important subclass of exponential families are the natural exponential families, which have a similar form for the moment-generating function for the distribution of x.

Differential identities for cumulants

In particular, using the properties of the cumulant generating function,

$$\mathrm{E}(T_j) = \frac{\partial A(\eta)}{\partial \eta_j}$$

and

$$\operatorname{cov}(T_i, T_j) = \frac{\partial^2 A(\eta)}{\partial \eta_i\, \partial \eta_j}.$$

The first two raw moments and all mixed second moments can be recovered from these two identities. Higher-order moments and cumulants are obtained by higher derivatives. This technique is often useful when T is a complicated function of the data, whose moments are difficult to calculate by integration.

Another way to see this that does not rely on the theory of cumulants is to begin from the fact that the distribution of an exponential family must be normalized, and differentiate. We illustrate using the simple case of a one-dimensional parameter, but an analogous derivation holds more generally.

In the one-dimensional case, we have

$$p(x) = g(\eta)h(x)e^{\eta T(x)}.$$

This must be normalized, so

$$1 = \int_x p(x)\,dx = \int_x g(\eta)h(x)e^{\eta T(x)}\,dx = g(\eta)\int_x h(x)e^{\eta T(x)}\,dx.$$

Take the derivative of both sides with respect to η:

$$\begin{aligned}
0 &= g(\eta)\frac{d}{d\eta}\int_x h(x)e^{\eta T(x)}\,dx + g'(\eta)\int_x h(x)e^{\eta T(x)}\,dx \\
&= g(\eta)\int_x h(x)\left(\frac{d}{d\eta}e^{\eta T(x)}\right)dx + g'(\eta)\int_x h(x)e^{\eta T(x)}\,dx \\
&= g(\eta)\int_x h(x)e^{\eta T(x)}T(x)\,dx + g'(\eta)\int_x h(x)e^{\eta T(x)}\,dx \\
&= \int_x T(x)g(\eta)h(x)e^{\eta T(x)}\,dx + \frac{g'(\eta)}{g(\eta)}\int_x g(\eta)h(x)e^{\eta T(x)}\,dx \\
&= \int_x T(x)p(x)\,dx + \frac{g'(\eta)}{g(\eta)}\int_x p(x)\,dx \\
&= \mathrm{E}[T(x)] + \frac{g'(\eta)}{g(\eta)} \\
&= \mathrm{E}[T(x)] + \frac{d}{d\eta}\log g(\eta)
\end{aligned}$$

Therefore,

$$\mathrm{E}[T(x)] = -\frac{d}{d\eta}\log g(\eta) = \frac{d}{d\eta}A(\eta).$$

Example 1

As an introductory example, consider the gamma distribution, whose distribution is defined by

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}.$$

Referring to the above table, we can see that the natural parameter is given by

$$\begin{aligned}
\eta_1 &= \alpha - 1, \\
\eta_2 &= -\beta,
\end{aligned}$$

the reverse substitutions are

$$\begin{aligned}
\alpha &= \eta_1 + 1, \\
\beta &= -\eta_2,
\end{aligned}$$

the sufficient statistics are $(\log x, x)$, and the log-partition function is

$$A(\eta_1, \eta_2) = \log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2).$$

We can find the mean of the sufficient statistics as follows. First, for η_1:

$$\begin{aligned}\mathbf{E}[\log x] &= \frac{\partial A(\eta_1, \eta_2)}{\partial \eta_1} = \frac{\partial}{\partial \eta_1}\left(\log \Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2)\right) \\ &= \psi(\eta_1 + 1) - \log(-\eta_2) \\ &= \psi(\alpha) - \log \beta,\end{aligned}$$

Where $\psi(x)$ is the digamma function (derivative of log gamma), and we used the reverse substitutions in the last step.

Now, for η_2:

$$\begin{aligned}\mathbf{E}[x] &= \frac{\partial A(\eta_1, \eta_2)}{\partial \eta_2} = \frac{\partial}{\partial \eta_2}\left(\log \Gamma(\eta_1 + 1) - (\eta_1 + 1)\log(-\eta_2)\right) \\ &= -(\eta_1 + 1)\frac{1}{-\eta_2}(-1) = \frac{\eta_1 + 1}{-\eta_2} \\ &= \frac{\alpha}{\beta},\end{aligned}$$

again making the reverse substitution in the last step.

To compute the variance of *x*, we just differentiate again:

$$\begin{aligned}\mathbf{Var}(x) &= \frac{\partial^2 A\left(\eta_1, \eta_2\right)}{\partial \eta_2^2} = \frac{\partial}{\partial \eta_2}\frac{\eta_1 + 1}{-\eta_2} \\ &= \frac{\eta_1 + 1}{\eta_2^2} \\ &= \frac{\alpha}{\beta^2}.\end{aligned}$$

All of these calculations can be done using integration, making use of various properties of the gamma function, but this requires significantly more work.

Example 2

As another example consider a real valued random variable *X* with density

$$p_\theta(x) = \frac{\theta e^{-x}}{\left(1 + e^{-x}\right)^{\theta+1}}$$

indexed by shape parameter $\theta \in (0, \infty)$ (this is called the skew-logistic distribution). The density can be rewritten as

$$\frac{e^{-x}}{1 + e^{-x}} \exp\left(-\theta \log\left(1 + e^{-x}\right) + \log(\theta)\right)$$

Notice this is an exponential family with natural parameter

$$\eta = -\theta,$$

sufficient statistic

$$T = \log\left(1 + e^{-x}\right),$$

and log-partition function

$$A(\eta) = -\log(\theta) = -\log(-\eta)$$

So using the first identity,

$$\mathbf{E}(\log(1 + e^{-X})) = \mathbf{E}(T) = \frac{\partial A(\eta)}{\partial \eta} = \frac{\partial}{\partial \eta}[-\log(-\eta)] = \frac{1}{-\eta} = \frac{1}{\theta},$$

and using the second identity

$$\mathbf{var}\left(\log\left(1 + e^{-X}\right)\right) = \frac{\partial^2 A(\eta)}{\partial \eta^2} = \frac{\partial}{\partial \eta}\left[\frac{1}{-\eta}\right] = \frac{1}{(-\eta)^2} = \frac{1}{\theta^2}.$$

This example illustrates a case where using this method is very simple, but the direct calculation would be nearly impossible.

Example 3

The final example is one where integration would be extremely difficult. This is the case of the Wishart distribution, which is defined over matrices. Even taking derivatives is a bit tricky, as it involves matrix calculus, but the respective identities are listed in that article.

From the above table, we can see that the natural parameter is given by

$$\begin{aligned}\boldsymbol{\eta}_1 &= -\frac{1}{2}\mathbf{V}^{-1}, \\ \eta_2 &= \frac{n - p - 1}{2},\end{aligned}$$

the reverse substitutions are

$$\begin{aligned}\mathbf{V} &= -\frac{1}{2}{\boldsymbol{\eta}_1}^{-1}, \\ n &= 2\eta_2 + p + 1,\end{aligned}$$

and the sufficient statistics are $(\mathbf{X}, \log|\mathbf{X}|)$.

The log-partition function is written in various forms in the table, to facilitate differentiation and back-substitution. We use the following forms:

$$A(\boldsymbol{\eta}_1, n) = -\frac{n}{2}\log|-\boldsymbol{\eta}_1| + \log\Gamma_p\left(\frac{n}{2}\right),$$

$$A(\mathbf{V}, \eta_2) = \left(\eta_2 + \frac{p+1}{2}\right)(p\log 2 + \log|\mathbf{V}|) + \log\Gamma_p\left(\eta_2 + \frac{p+1}{2}\right).$$

Expectation of X (associated with $\boldsymbol{\eta}_1$)

To differentiate with respect to $\boldsymbol{\eta}_1$, we need the following matrix calculus identity:

$$\frac{\partial \log|a\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^{\mathrm{T}}$$

Then:

$$\begin{aligned}
\mathbf{E}[\mathbf{X}] &= \frac{\partial A(\boldsymbol{\eta}_1, \cdots)}{\partial \boldsymbol{\eta}_1} \\
&= \frac{\partial}{\partial \boldsymbol{\eta}_1}\left[-\frac{n}{2}\log|-\boldsymbol{\eta}_1| + \log\Gamma_p\left(\frac{n}{2}\right)\right] \\
&= -\frac{n}{2}(\boldsymbol{\eta}_1^{-1})^{\mathrm{T}} \\
&= \frac{n}{2}(-\boldsymbol{\eta}_1^{-1})^{\mathrm{T}} \\
&= n(\mathbf{V})^{\mathrm{T}} \\
&= n\mathbf{V}
\end{aligned}$$

The last line uses the fact that $\mathbf{V}$ is symmetric, and therefore it is the same when transposed.

Expectation of log |X| (associated with η_2)

Now, for η_2, we first need to expand the part of the log-partition function that involves the multivariate gamma function:

$$\log\Gamma_p(a) = \log\left(\pi^{\frac{p(p-1)}{4}}\prod_{j=1}^{p}\Gamma\left(a + \frac{1-j}{2}\right)\right) = \frac{p(p-1)}{4}\log\pi + \sum_{j=1}^{p}\log\Gamma\left[a + \frac{1-j}{2}\right]$$

We also need the digamma function:

$$\psi(x) = \frac{d}{dx}\log\Gamma(x).$$

Then:

$$\begin{aligned}
\mathbf{E}[\log|\mathbf{X}|] &= \frac{\partial A(\ldots, \eta_2)}{\partial \eta_2} \\
&= \frac{\partial}{\partial \eta_2}\left[-\left(\eta_2 + \frac{p+1}{2}\right)(p\log 2 + \log|\mathbf{V}|) + \log\Gamma_p\left(\eta_2 + \frac{p+1}{2}\right)\right] \\
&= \frac{\partial}{\partial \eta_2}\left[\left(\eta_2 + \frac{p+1}{2}\right)(p\log 2 + \log|\mathbf{V}|) + \frac{p(p-1)}{4}\log\pi + \sum_{j=1}^{p}\log\Gamma\left(\eta_2 + \frac{p+1}{2} + \frac{1-j}{2}\right)\right] \\
&= p\log 2 + \log|\mathbf{V}| + \sum_{j=1}^{p}\psi\left(\eta_2 + \frac{p+1}{2} + \frac{1-j}{2}\right) \\
&= p\log 2 + \log|\mathbf{V}| + \sum_{j=1}^{p}\psi\left(\frac{n-p-1}{2} + \frac{p+1}{2} + \frac{1-j}{2}\right) \\
&= p\log 2 + \log|\mathbf{V}| + \sum_{j=1}^{p}\psi\left(\frac{n+1-j}{2}\right)
\end{aligned}$$

This latter formula is listed in the Wishart distribution article. Both of these expectations are needed when deriving the variational Bayes update equations in a Bayes network involving a Wishart distribution (which is the conjugate prior of the multivariate normal distribution).

Computing these formulas using integration would be much more difficult. The first one, for example, would require matrix integration.

Entropy

Relative entropy

The relative entropy (Kullback–Leibler divergence, KL divergence) of two distributions in an exponential family has a simple expression as the Bregman divergence between the natural parameters with respect to the log-normalizer.[10] The relative entropy is defined in terms of an integral, while the Bregman divergence is defined in terms of a derivative and inner product, and thus is easier to calculate and has a closed-form expression (assuming the derivative has a closed-form expression). Further, the Bregman divergence in terms of the natural parameters and the log-normalizer equals the Bregman divergence of the dual parameters (expectation parameters), in the opposite order, for the convex conjugate function.

Fixing an exponential family with log-normalizer A (with convex conjugate A^*), writing $P_{A,\theta}$ for the distribution in this family corresponding a fixed value of the natural parameter θ (writing θ' for another value, and with η, η' for the corresponding dual expectation/moment parameters), writing KL for the KL divergence, and B_A for the Bregman divergence, the divergences are related as:

$$\mathrm{KL}(P_{A,\theta} \parallel P_{A,\theta'}) = B_A(\theta' \parallel \theta) = B_{A^*}(\eta \parallel \eta').$$

The KL divergence is conventionally written with respect to the *first* parameter, while the Bregman divergence is conventionally written with respect to the *second* parameter, and thus this can be read as "the relative entropy is equal to the Bregman divergence defined by the log-normalizer on the swapped natural parameters", or equivalently as "equal to the Bregman divergence defined by the dual to the log-normalizer on the expectation parameters".

Maximum entropy derivation

Exponential families arise naturally as the answer to the following question: what is the maximum-entropy distribution consistent with given constraints on expected values?

The information entropy of a probability distribution *dF*(*x*) can only be computed with respect to some other probability distribution (or, more generally, a positive measure), and both measures must be mutually absolutely continuous. Accordingly, we need to pick a *reference measure dH*(*x*) with the same support as *dF*(*x*).

The entropy of *dF*(*x*) relative to *dH*(*x*) is

$$S[dF \mid dH] = -\int \frac{dF}{dH}\log\frac{dF}{dH}\,dH$$

or

$$S[dF \mid dH] = \int \log \frac{dH}{dF}\, dF$$

where *dF*/*dH* and *dH*/*dF* are Radon–Nikodym derivatives. Note that the ordinary definition of entropy for a discrete distribution supported on a set *I*, namely

$$S = -\sum_{i \in I} p_i \log p_i$$

assumes, though this is seldom pointed out, that *dH* is chosen to be the counting measure on *I*.

Consider now a collection of observable quantities (random variables) T_i. The probability distribution *dF* whose entropy with respect to *dH* is greatest, subject to the conditions that the expected value of T_i be equal to t_i, is an exponential family with *dH* as reference measure and $(T_1, ..., T_n)$ as sufficient statistic.

The derivation is a simple variational calculation using Lagrange multipliers. Normalization is imposed by letting $T_0 = 1$ be one of the constraints. The natural parameters of the distribution are the Lagrange multipliers, and the normalization factor is the Lagrange multiplier associated to T_0.

For examples of such derivations, see Maximum entropy probability distribution.

Role in statistics

Classical estimation: sufficiency

According to the **Pitman–Koopman–Darmois theorem**, among families of probability distributions whose domain does not vary with the parameter being estimated, only in exponential families is there a sufficient statistic whose dimension remains bounded as sample size increases.

Less tersely, suppose X_k, (where $k = 1, 2, 3, \ldots n$) are independent, identically distributed random variables. Only if their distribution is one of the *exponential family* of distributions is there a sufficient statistic $\boldsymbol{T}(X_1, ..., X_n)$ whose number of scalar components does not increase as the sample size *n* increases; the statistic $\boldsymbol{T}$ may be a vector or a single scalar number, but whatever it is, its size will neither grow nor shrink when more data are obtained.

As a counterexample if these conditions are relaxed, note that the family of uniform distributions (either discrete or continuous, with either or both bounds unknown) has a sufficient statistic, namely the sample maximum, sample minimum, and sample size, but does not form an exponential family, as the domain varies with the parameters.

Bayesian estimation: conjugate distributions

Exponential families are also important in Bayesian statistics. In Bayesian statistics a prior distribution is multiplied by a likelihood function and then normalised to produce a posterior distribution. In the case of a likelihood which belongs to an exponential family there exists a conjugate prior, which is often also in an exponential family. A conjugate prior π for the parameter $\boldsymbol{\eta}$ of an exponential family

$$f(x \mid \boldsymbol{\eta}) = h(x) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{T}(x) - A(\boldsymbol{\eta})\right)$$

is given by

$$p_\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi} - \nu A(\boldsymbol{\eta})\right),$$

or equivalently

$$p_\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^\nu \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}\right), \qquad \boldsymbol{\chi} \in \mathbb{R}^s$$

where *s* is the dimension of $\boldsymbol{\eta}$ and $\nu > 0$ and $\boldsymbol{\chi}$ are hyperparameters (parameters controlling parameters). ν corresponds to the effective number of observations that the prior distribution contributes, and $\boldsymbol{\chi}$ corresponds to the total amount that these pseudo-observations contribute to the sufficient statistic over all observations and pseudo-observations. $f(\boldsymbol{\chi}, \nu)$ is a normalization constant that is automatically determined by the remaining functions and serves to ensure that the given function is a probability density function (i.e. it is normalized). $A(\boldsymbol{\eta})$ and equivalently $g(\boldsymbol{\eta})$ are the same functions as in the definition of the distribution over which π is the conjugate prior.

A conjugate prior is one which, when combined with the likelihood and normalised, produces a posterior distribution which is of the same type as the prior. For example, if one is estimating the success probability of a binomial distribution, then if one chooses to use a beta distribution as one's prior, the posterior is another beta distribution. This makes the computation of the posterior particularly simple. Similarly, if one is estimating the parameter of a Poisson distribution the use of a gamma prior will lead to another gamma posterior. Conjugate priors are often very flexible and can be very convenient. However, if one's belief about the likely value of the theta parameter of a binomial is represented by (say) a bimodal (two-humped) prior distribution, then this cannot be represented by a beta distribution. It can however be represented by using a mixture density as the prior, here a combination of two beta distributions; this is a form of hyperprior.

An arbitrary likelihood will not belong to an exponential family, and thus in general no conjugate prior exists. The posterior will then have to be computed by numerical methods.

To show that the above prior distribution is a conjugate prior, we can derive the posterior.

First, assume that the probability of a single observation follows an exponential family, parameterized using its natural parameter:

$$p_F(x \mid \boldsymbol{\eta}) = h(x) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{T}(x)\right)$$

Then, for data $\mathbf{X} = (x_1, \ldots, x_n)$, the likelihood is computed as follows:

$$p(\mathbf{X} \mid \boldsymbol{\eta}) = \left(\prod_{i=1}^{n} h(x_i)\right) g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i)\right)$$

Then, for the above conjugate prior:

$$p_\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \propto g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi})$$

We can then compute the posterior as follows:

$$\begin{aligned}
p(\boldsymbol{\eta} \mid \mathbf{X}, \boldsymbol{\chi}, \nu) &\propto p(\mathbf{X} \mid \boldsymbol{\eta}) p_\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) \\
&= \left(\prod_{i=1}^{n} h(x_i)\right) g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i)\right) f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \\
&\propto g(\boldsymbol{\eta})^n \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{T}(x_i)\right) g(\boldsymbol{\eta})^\nu \exp(\boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}) \\
&\propto g(\boldsymbol{\eta})^{\nu+n} \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \left(\boldsymbol{\chi} + \sum_{i=1}^{n} \mathbf{T}(x_i)\right)\right)
\end{aligned}$$

The last line is the kernel of the posterior distribution, i.e.

$$p(\boldsymbol\eta \mid \mathbf{X}, \boldsymbol\chi, \nu) = p_\pi \left(\boldsymbol\eta \mid \boldsymbol\chi + \sum_{i=1}^{n} \mathbf{T}(x_i), \nu + n \right)$$

This shows that the posterior has the same form as the prior.

Note in particular that the data **X** enters into this equation *only* in the expression

$$\mathbf{T}(\mathbf{X}) = \sum_{i=1}^{n} \mathbf{T}(x_i),$$

which is termed the sufficient statistic of the data. That is, the value of the sufficient statistic is sufficient to completely determine the posterior distribution. The actual data points themselves are not needed, and all sets of data points with the same sufficient statistic will have the same distribution. This is important because the dimension of the sufficient statistic does not grow with the data size — it has only as many components as the components of $\boldsymbol\eta$ (equivalently, the number of parameters of the distribution of a single data point).

The update equations are as follows:

$$\begin{aligned}
\boldsymbol\chi' &= \boldsymbol\chi + \mathbf{T}(\mathbf{X}) \\
&= \boldsymbol\chi + \sum_{i=1}^{n} \mathbf{T}(x_i) \\
\nu' &= \nu + n
\end{aligned}$$

This shows that the update equations can be written simply in terms of the number of data points and the sufficient statistic of the data. This can be seen clearly in the various examples of update equations shown in the conjugate prior page. Note also that because of the way that the sufficient statistic is computed, it necessarily involves sums of components of the data (in some cases disguised as products or other forms — a product can be written in terms of a sum of logarithms). The cases where the update equations for particular distributions don't exactly match the above forms are cases where the conjugate prior has been expressed using a different parameterization than the one that produces a conjugate prior of the above form — often specifically because the above form is defined over the natural parameter $\boldsymbol\eta$ while conjugate priors are usually defined over the actual parameter $\boldsymbol\theta$.

Hypothesis testing: uniformly most powerful tests

A one-parameter exponential family has a monotone non-decreasing likelihood ratio in the sufficient statistic $T(x)$, provided that $\eta(\theta)$ is non-decreasing. As a consequence, there exists a uniformly most powerful test for testing the hypothesis H_0: $\theta \geq \theta_0$ *vs.* H_1: $\theta < \theta_0$.

Generalized linear models

Exponential families form the basis for the distribution functions used in generalized linear models, a class of model that encompass many of the commonly used regression models in statistics.

See also

- Natural exponential family
- Exponential dispersion model
- Gibbs measure

Notes

1. For example, the family of normal distributions includes the standard normal distribution $N(0, 1)$ with mean 0 and variance 1, as well as other normal distributions with different mean and variance.

References

Citations

1. Kupperman, M. (1958) "Probabilities of Hypotheses and Information-Statistics in Sampling from Exponential-Class Populations", Annals of Mathematical Statistics, 9 (2), 571–575 JSTOR 2237349 (https://www.jstor.org/stable/2237349)
2. Andersen, Erling (September 1970). "Sufficiency and Exponential Families for Discrete Sample Spaces". *Journal of the American Statistical Association*. Journal of the American Statistical Association, Vol. 65, No. 331. **65** (331): 1248–1255. doi:10.2307/2284291 (https://doi.org/10.2307%2F2284291). JSTOR 2284291 (https://www.jstor.org/stable/2284291). MR 0268992 (https://www.ams.org/mathscinet-getitem?mr=0268992).
3. Pitman, E.; Wishart, J. (1936). "Sufficient statistics and intrinsic accuracy". *Mathematical Proceedings of the Cambridge Philosophical Society*. **32** (4): 567–579. Bibcode:1936PCPS...32..567P (http://adsabs.harvard.edu/abs/1936PCPS...32..567P). doi:10.1017/S0305004100019307 (https://doi.org/10.1017%2FS0305004100019307).
4. Darmois, G. (1935). "Sur les lois de probabilites a estimation exhaustive". *C. R. Acad. Sci. Paris* (in French). **200**: 1265–1266.
5. Koopman, B (1936). "On distribution admitting a sufficient statistic". *Transactions of the American Mathematical Society*. American Mathematical Society. **39** (3): 399–409. doi:10.2307/1989758 (https://doi.org/10.2307%2F1989758). JSTOR 1989758 (https://www.jstor.org/stable/1989758). MR 1501854 (https://www.ams.org/mathscinet-getitem?mr=1501854).
6. Abramovich & Ritov (2013). *Statistical Theory: A Concise Introduction*. Chapman & Hall. ISBN 978-1439851845.
7. Blei, David. "Variational Inference" (https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf) (PDF). *Princeton*.
8. Nielsen, Frank; Garcia, Vincent (2009). "Statistical exponential families: A digest with flash cards". arXiv:0911.4863 (https://arxiv.org/abs/0911.4863).
9. van Garderen, Kees Jan (1997). "Curved Exponential Models in Econometrics". *Econometric Theory*. **13** (6): 771–790. doi:10.1017/S0266466600006253 (https://doi.org/10.1017%2FS0266466600006253).
10. Nielsen & Nock 2010, 4. Bregman Divergences and Relative Entropy of Exponential Families.

Sources

- Nielsen, Frank; Garcia, Vincent (2009). "Statistical exponential families: A digest with flash cards". arXiv:0911.4863 (https://arxiv.org/abs/0911.4863). Bibcode:2009arXiv0911.4863N (http://adsabs.harvard.edu/abs/2009arXiv0911.4863N).
- Nielsen, Frank; Nock, Richard (2010). *Entropies and cross-entropies of exponential families* (https://web.archive.org/web/20190331194854/https://www.lix.polytechnique.fr/~nielsen/EntropyEF-ICIP2010.pdf) (PDF). IEEE International Conference on Image Processing. doi:10.1109/ICIP.2010.5652054 (https://doi.org/10.1109%2FICIP.2010.5652054). Archived from the original (https://www.lix.polytechnique.fr/~nielsen/EntropyEF-ICIP2010.pdf) (PDF) on 2019-03-31.

Further reading

- Fahrmeir, Ludwig; Tutz, G. (1994). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer. pp. 18–22, 345–349. ISBN 0-387-94233-5.
- Keener, Robert W. (2006). *Theoretical Statistics: Topics for a Core Course*. Springer. pp. 27–28, 32–33. ISBN 978-0-387-93838-7.
- Lehmann, E. L.; Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). sec. 1.5. ISBN 0-387-98502-6.

External links

- A primer on the exponential family of distributions (http://www.casact.org/pubs/dpp/dpp04/04dpp117.pdf)
- Exponential family of distributions (http://jeff560.tripod.com/e.html) on the Earliest known uses of some of the words of mathematics (http://jeff560.tripod.com/mathword.html)
- jMEF: A Java library for exponential families (https://vincentfpgarcia.github.com/jMEF/)