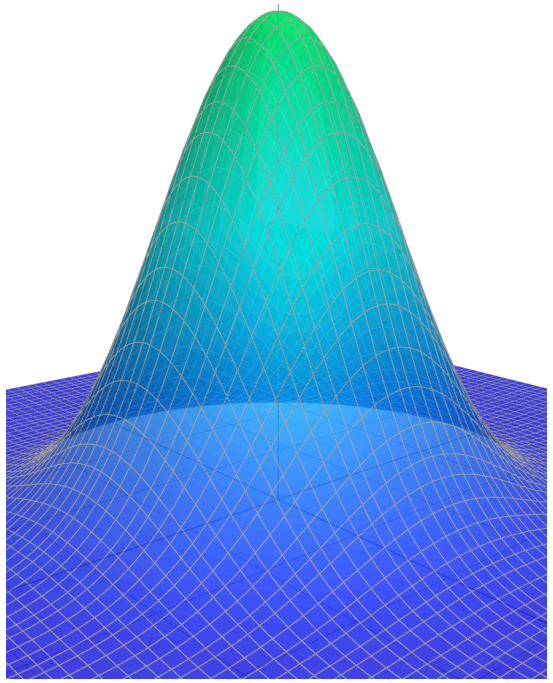
Exponential families

Statistical Inference - PhD EcoStatData

Tommaso Rigon
Università degli Studi di Milano-Bicocca

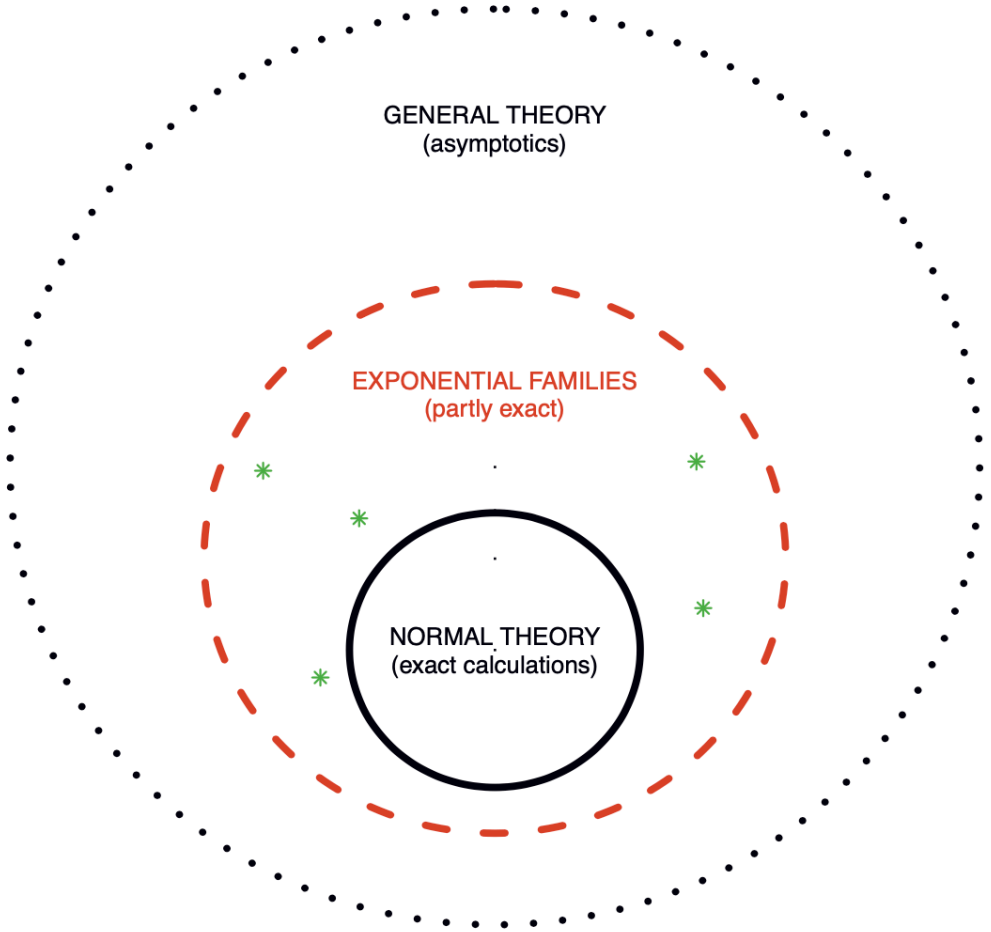
Homepage



- This unit will cover the following **topics**:
  - One-parameter and multiparameter exponential families
  - Likelihood, inference, sufficiency and completeness
- The **prime role** of **exponential families** in the theory of statistical inference was first emphasized by Fisher (1934).
- Most **well-known distributions**—such as Gaussian, Poisson, Binomial, and Gamma—are instances of exponential families.
- Exponential families are the distributions typically considered when presenting the usual "regularity conditions".

- With a few minor exceptions, this presentation will closely follow Chapters 5 and 6 of Pace and Salvan (1997).

Overview



- Figure 1 of Efron (2023). Three level of statistical modeling.

Home page

One-parameter exponential families

Exponential tilting

- Let Y be a **non-degenerate** random variable with **support** $\mathcal{Y} \subseteq \mathbb{R}$ and **density** $f_0(y)$ with respect to a dominating measure $\nu(\mathrm{d}y)$.
- We aim at building a **parametric family** $\mathcal{F} = \{f(;\theta) : \theta \in \Theta \subseteq \mathbb{R}\}$ with common support $\mathcal{Y}$ such that f_0 is a special case, namely $f_0 \in \mathcal{F}$.
- A strategy for doing this is called **exponential tilting**, namely we could set

$$f(y;\theta) \propto e^{\theta y} f_0(y).$$

  Thus, if $f(y;\theta)$ is generated via exponential tilting, then $f(y;0) = e^0 f_0(y) = f_0(y)$.
- Let us define the mapping $M_0 : \mathbb{R} \to (0, \infty]$

$$M_0(\theta) := \int_{\mathcal{Y}} e^{\theta y} f_0(y) \nu(\mathrm{d}y), \qquad \theta \in \mathbb{R}.$$

  If $M_0(\theta)$ is **finite** in a neighborhood of the origin, it is the **moment generating function** of Y.
- Moreover, we define the set $\tilde{\Theta} \subseteq \mathbb{R}$ as the set of all θ such that $M_0(\theta)$ is finite, i.e.

$$\tilde{\Theta} = \{\theta \in \mathbb{R} : M_0(\theta) < \infty\}.$$

Natural exponential family of order one

- The mapping $K(\theta) = K_0(\theta) = \log M_0(\theta)$ is the **cumulant generating function** of f_0. It is **finite** if and only if $M_0(\theta)$ is finite.

The parametric family generated via **exponential tilting** of f_0

$$\mathcal{F}^1_{\text{ne}} = \left\{ f(y;\theta) = \frac{e^{\theta y} f_0(y)}{M_0(\theta)} = f_0(y) \exp\{\theta y - K(\theta)\}, \quad y \in \mathcal{Y}, \theta \in \tilde{\Theta} \right\},$$

is called a **natural exponential family** of order one, and $\tilde{\Theta} = \{\theta \in \mathbb{R} : K(\theta) < \infty\}$ is the **natural parameter space**.

- The natural parameter space $\tilde{\Theta}$ is the **widest possible** and must be an **interval**; see exercises. The family $\mathcal{F}^1_{\text{ne}}$ is said to be **full**, whereas a subfamily of $\mathcal{F}^1_{\text{ne}}$ with $\Theta \subseteq \tilde{\Theta}$ is **non-full**.
- By definition, all the densities $f(y;\theta) \in \mathcal{F}^1_{\text{ne}}$ have the **same support**.

A natural exponential family of order one, $\mathcal{F}^1_{\text{ne}}$, is said to be **regular** if $\tilde{\Theta}$ is open.

Moment generating function

- In regular problems, the functions $M_0(\theta)$ and $K_0(\theta)$ associated to a r.v. Y with density f_0 are **finite** in a neighbor of the origin. A **sufficient** condition is that $\tilde{\Theta}$ is an **open set** (regular $\mathcal{F}^1_{\text{en}}$).

Suppose $M_0(t) < \infty$ for any $|t| < t_0$ and for some $t_0 > 0$. Then a standard result of probability theory (e.g. Billingsley (1995), Section 21) implies:

- The random variable Y has **finite moments** of all orders, i.e. $\mu_k = \mathbb{E}(Y^k) < \infty$ for all $k \geq 1$.
- The moments $(\mu_k)_{k\geq 1}$ and **moment generating function** $M_0(t)$ **uniquely characterize** the **law** of Y and f_0. Moreover, $M_0(t)$ admits a Taylor expansion around the origin:

$$M_0(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mu_k, \qquad |t| < t_0.$$

- The moments μ_k equal the kth derivative of $M_0(t)$ evaluated at the origin:

$$\mu_k = \mathbb{E}_\theta(Y^k) = \frac{\partial^k}{\partial t^k} M_0(t)\Big|_{t=0}, \qquad k \geq 1.$$

Cumulant generating function

Suppose $K_0(t) = \log M_0(t) < \infty$ for any $|t| < t_0$ and for some $t_0 > 0$. Then:

- K_0 **uniquely characterizes** the law of Y and it admits a **Taylor expansion**

$$K_0(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \cdots = \sum_{k=1}^{\infty} \frac{t^k}{k!} \kappa_k, \qquad |t| < t_0,$$

  where the coefficients $(\kappa_k)_{k \geq 1}$ are the **cumulants** of Y.

- The cumulants κ_k equal the kth derivative of $K_0(t)$ evaluated at the origin

$$\kappa_k = \frac{\partial^k}{\partial t^k} K_0(t) \Big|_{t=0}, \qquad k \geq 1.$$

  Moreover, it can be shown the following moment relationships hold:

$$\kappa_1 = \mathbb{E}_\theta(Y), \quad \kappa_2 = \mathrm{var}_\theta(Y), \quad \kappa_3 = \mathbb{E}_\theta\{(Y - \mu_1)^3\}, \quad \kappa_4 = \mathbb{E}_\theta\{(Y - \mu_1)^4\} - 3\mathrm{var}_\theta(Y)^2.$$

Refer to Pace and Salvan (1997), Section 3.2.5 for detailed derivations. Standardized cumulants $\kappa_3/\kappa_2^{3/2}$ and κ_4/κ_2^2 are the **skewness** and the (excess of) **kurtosis** of Y.

Home page

Example: uniform distribution 📖

- Let $Y \sim \text{Unif}(0,1)$ so that $f_0(y) = 1$ for $y \in [0,1]$. The **exponential tilting** of f_0 gives

$$f(y;\theta) \propto e^{\theta y} f_0(y) = e^{\theta y}, \qquad y \in [0,1], \quad \theta \in \mathbb{R}.$$

- The normalizing constant, that is, the **moment generating function**, is

$$M_0(\theta) = \mathbb{E}(e^{\theta Y}) = \int_0^1 e^{\theta y} \mathrm{d}y = \frac{e^\theta}{\theta}\Big|_0^1 = \frac{e^\theta - 1}{\theta}, \qquad \theta \neq 0.$$

  with $M_0(0) = 1$. Note that M_0 is continuous since $\lim_{\theta \to 0}(e^\theta - 1)/\theta = 1$.

- Consequently, we have $M_0(\theta) < \infty$ for all $\theta \in \mathbb{R}$ and the **natural parameter space** is $\tilde{\Theta} = \mathbb{R}$, which is an **open set**. The resulting density is

$$f(y;\theta) = \frac{\theta e^{\theta y}}{e^\theta - 1} = \exp\{\theta y - K(\theta)\}, \qquad y \in [0,1],$$

  where $K(\theta) = \log\{(e^\theta - 1)/\theta\}$.

- It holds in general that $\tilde{\Theta} = \mathbb{R}$ whenever f_0 has **bounded support**; thus, the family is **regular**.

Example: Poisson distribution 📖

- Let $Y \sim \text{Poisson}(1)$ so that $f_0(y) = e^{-1}/y!$ for $y \in \mathbb{N}$. The **exponential tilting** of f_0 gives

$$f(y;\theta) \propto e^{\theta y} f_0(y) = \frac{e^{\theta y} e^{-1}}{y!}, \qquad y \in \mathbb{N}, \quad \theta \in \mathbb{R}.$$

- The normalizing constant, that is, the **moment generating function**, is

$$M_0(\theta) = \mathbb{E}(e^{\theta Y}) = e^{-1} \sum_{k=0}^{\infty} \frac{e^{\theta k}}{k!} = \exp\{e^{\theta} - 1\}, \qquad \theta \in \mathbb{R}.$$

- Consequently, we have $M_0(\theta) < \infty$ for all $\theta \in \mathbb{R}$ and the **natural parameter space** is $\tilde{\Theta} = \mathbb{R}$, which is an **open set**. The resulting density is

$$f(y;\theta) = \frac{e^{\theta y} e^{-1}}{y!} \frac{e^{-e^{\theta}}}{e^{-1}} = \frac{e^{-1}}{y!} \exp\{\theta y - (e^{\theta} - 1)\} = \frac{\lambda^y e^{\lambda}}{y!}, \qquad y \in \mathbb{N},$$

  so that $K(\theta) = e^{\theta} - 1$ and having defined $\lambda = e^{\theta}$.

- In other words, the tilted density is again a Poisson distribution with mean e^{θ}.

Example: exponential family generated by a Gaussian 📖

- Let $Y \sim \mathrm{N}(0,1)$ so that $f_0(y) = 1/(\sqrt{2\pi})e^{-y^2/2}$ for $y \in \mathbb{R}$. The **exponential tilting** of f_0 gives

$$f(y;\theta) \propto e^{\theta y} f_0(y) = \frac{1}{\sqrt{2\pi}} e^{\theta y - y^2/2}, \qquad y, \theta \in \mathbb{R}.$$

- The normalizing constant, that is, the **moment generating function**, is

$$M_0(\theta) = \mathbb{E}(e^{\theta Y}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\theta y - y^2/2} \mathrm{d}y = e^{\theta^2/2}, \qquad \theta \in \mathbb{R}.$$

- Consequently, we have $M_0(\theta) < \infty$ for all $\theta \in \mathbb{R}$ and the **natural parameter space** is $\tilde{\Theta} = \mathbb{R}$, which is an **open set**. The resulting density is

$$f(y;\theta) = \frac{1}{\sqrt{2\pi}} e^{\theta y} e^{-y^2/2} e^{-\theta^2/2} = \frac{e^{-y^2/2}}{\sqrt{2\pi}} \exp\{\theta y - \theta^2/2\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2}, \qquad y \in \mathbb{R},$$

  so that $K(\theta) = \theta^2/2$.

- In other words, the tilted density is again a Gaussian distribution with mean θ.

Closure under exponential tilting 📖

- Let $\mathcal{F}_{\text{ne}}^1$ be an exponential family with **parameter** ψ and **natural parameter space** $\tilde{\Psi}$, with density $f(y;\psi) = f_0(y)\exp\{\psi y - K(\psi)\}$. The **exponential tilting** of $f(y;\psi)$ gives

$$f(y;\theta,\psi) \propto e^{\theta y} f(y;\psi) \propto f_0(y)\exp\{(\theta+\psi)y\},$$

  and the **normalizing constant** of $f_0(y)\exp\{(\theta+\psi)y\}$ is therefore

$$\int_{\mathcal{Y}} f_0(y)\exp\{(\theta+\psi)y\}\,\nu(\mathrm{d}y) = M_0(\theta+\psi).$$

- Thus, for any θ and ψ such that $M_0(\theta+\psi) < \infty$, the corresponding density is

$$f(y;\theta,\psi) = f_0(y)\exp\{(\theta+\psi)y - K(\theta+\psi)\},$$

  which is again a **member** of the **exponential family** $\mathcal{F}_{\text{ne}}^1$, with updated parameter $\theta+\psi$.

> Exponential families are **closed** under **exponential tilting**, and $\mathcal{F}_{\text{ne}}^1$ can be thought of as being generated by any of its members.

Moments and cumulants

- The functions $M_0(\theta)$ and $K(\theta) = K_0(\theta)$ of a $\mathcal{F}_{\text{en}}^1$, refer to the **baseline** density $f_0(y)$. Indeed, for any fixed θ, the **moment generating function** of $f(y;\theta) \in \mathcal{F}_{\text{en}}^1$ is

$$M_\theta(t) := \int_{\mathcal{Y}} e^{ty} f(y;\theta)\, \nu(\mathrm{d}y) = \frac{1}{M_0(\theta)} \int_{\mathcal{Y}} e^{(t+\theta)y} f_0(y)\, \nu(\mathrm{d}y) = \frac{M_0(t+\theta)}{M_0(\theta)}, \quad t+\theta \in \tilde{\Theta}.$$

- Consequently, the **cumulant generating function** of $f(y;\theta)$ relates to K_0 as follows:

$$K_\theta(t) = \log M_\theta(t) = K_0(t+\theta) - K_0(\theta), \quad t+\theta \in \tilde{\Theta}.$$

If $\mathcal{F}_{\text{en}}^1$ is a **regular** family, then $\tilde{\Theta}$ is an **open set**, and $\tilde{\Theta} = \operatorname{int} \tilde{\Theta}$, meaning that θ is always an **inner point** of $\tilde{\Theta}$. Therefore, there exists a t_0 such that $t+\theta \in \tilde{\Theta}$ for all $|t| < t_0$ implying that both M_θ and K_θ are **well-defined**.

If $\mathcal{F}_{\text{en}}^1$ is **not** regular, then for $M_\theta(t)$ and $K_\theta(t)$ to be well-defined, we require that θ is **not** a **boundary point**; that is, $\theta \in \operatorname{int} \tilde{\Theta}$, meaning it belongs to the interior of $\tilde{\Theta}$.

Textbooks sometimes suppress additive constants in defining $K_0(\theta)$, e.g. using e^θ instead of $e^\theta - 1$. This is inconsequential (constants cancel in $K_\theta(t)$) but somewhat misleading.

Mean value mapping I

- **Moments** and **cumulants** exist for every $\theta \in \operatorname{int} \tilde{\Theta}$. In particular, the cumulants are

$$\kappa_k = \frac{\partial^k}{\partial t^k} K_\theta(t)\Big|_{t=0} = \frac{\partial^k}{\partial t^k} \left[K(t+\theta) - K(\theta)\right]\Big|_{t=0} = \frac{\partial^k}{\partial \theta^k} K(\theta), \qquad k \geq 1.$$

Let $Y \sim f(y;\theta)$, with $f(y;\theta) \in \mathcal{F}_{\text{en}}^1$. The first two moments of Y are obtained as:

$$\mu(\theta) := \mathbb{E}_\theta(Y) = \frac{\partial}{\partial \theta} K(\theta), \qquad \operatorname{var}_\theta(Y) = \frac{\partial}{\partial \theta} \mu(\theta) = \frac{\partial^2}{\partial \theta^2} K(\theta),$$

We call $\mu : \operatorname{int} \tilde{\Theta} \to \mathbb{R}$ the **mean value mapping**.

- If f_0 is non-degenerate, then $\operatorname{var}_\theta(Y) > 0$, implying that $K(\theta)$ is a **convex function**, and $\mu(\theta)$ is a **smooth** and **monotone increasing**, namely is a **one-to-one** map.
- Thus, if $\mathcal{F}_{\text{en}}^1$ is a **regular** exponential family, then $\tilde{\Theta} = \operatorname{int} \tilde{\Theta}$ and $\mu(\theta)$ is a **reparametrization**.

Mean value mapping II

The mean value mapping has range $\mathcal{M} = \text{Range}(\mu) = \{\mu(\theta) : \theta \in \text{int}\,\tilde{\Theta}\}$. The set $\mathcal{M} \subseteq \mathbb{R}$ is called **mean space** or **expectation space**.

Let $C = C(\mathcal{Y})$ be the **closed convex hull** of the sample space $\mathcal{Y}$, which is the **smallest closed convex set** $C \subseteq \mathbb{R}$ **containing** $\mathcal{Y}$, namely:

$$C(\mathcal{Y}) = \{y \in \mathbb{R} : y = \lambda y_1 + (1-\lambda)y_2, \quad 0 \leq \lambda \leq 1, \quad y_1, y_2 \in \mathcal{Y}\}.$$

- Hence, if $\mathcal{Y} = \{0, 1, \ldots, N\}$, then $C = [0, N]$. If $\mathcal{Y} = \mathbb{N}$, then $C = \mathbb{R}^+$. If $\mathcal{Y} = \mathbb{R}$, then $C = \mathbb{R}$.
- Because of the properties of expectations, $\mu(\theta) \in \text{int}\, C(\mathcal{Y})$ for all $\theta \in \text{int}\,\tilde{\Theta}$, namely

$$\mathcal{M} \subseteq \text{int}\, C(\mathcal{Y}).$$

  Indeed, $\text{int}\, C(\mathcal{Y})$ is an **open interval** whose extremes are the **infimum** and **supremum** of $\mathcal{Y}$.

Both definitions naturally generalize to the multivariate case when $C, \mathcal{Y} \subseteq \mathbb{R}^p$, for $p > 1$.

Mean value mapping III 📖

- In a regular exponential family, the mean value mapping $\mu(\theta)$ is a **reparametrization**, meaning that for each $\theta \in \tilde{\Theta}$, there exists a **unique** mean $\mu \in \mathcal{M}$ such that $\mu = \mu(\theta)$.
- Moreover, in regular families, a much stronger result holds: for each value of $y \in \operatorname{int} C(\mathcal{Y})$, there exists a **unique** $\theta \in \tilde{\Theta}$ such that $\mu(\theta) = y$.

Theorem (Pace and Salvan (1997), Theorem 5.1)

If $\mathcal{F}_{\text{en}}^1$ is regular, then $\Theta = \operatorname{int} \tilde{\Theta} = \tilde{\Theta}$ and $\mathcal{M} = \operatorname{int} C$.

- This establishes a **duality** between the expectation space $\mathcal{M}$ and the sample space. Any value in $\operatorname{int} C$ can be "reached", that is, there exists a distribution $f(y;\theta)$ with that mean.
- This correspondence is crucial in maximum likelihood estimation and inference.

This theorem can actually be strengthened: a necessary and sufficient condition for $\mathcal{M} = \operatorname{int} C$ is that the family $\mathcal{F}_{\text{en}}^1$ is **steep** (a regular family is also steep); see Pace and Salvan (1997).

A non regular and non steep exponential family

- Let us a consider an exponential family $\mathcal{F}_{\text{en}}^1$ generated by the density

$$f_0(y) = c\frac{e^{-|y|}}{1+y^4}, \qquad y \in \mathbb{R}.$$

for some normalizing constant $c > 0$. The **exponential tilting** of f_0 gives

$$f(y;\theta) \propto e^{\theta y} f_0(y) \propto \frac{e^{-|y|+\theta y}}{1+y^4}, \qquad y \in \mathbb{R}, \quad \theta \in \tilde{\Theta}.$$

- The function $M_0(\theta)$ is unavailable in closed form, however $\tilde{\Theta} = [-1, 1]$ since

$$M_0(\theta) < \infty, \qquad \theta \in [-1, 1].$$

- Since $\tilde{\Theta}$ is a **closed set**, the exponential family is **not regular** (and is not steep either). In fact, one can show that $\lim_{\theta \to 1} \mu(\theta) = a < \infty$, implying that

$$\mathcal{M} = (-a, a), \qquad \text{whereas} \qquad \text{int}\, C = \mathbb{R}.$$

- In other words, there are no values of θ such that $\mu(\theta) = y$ for any $y > a$, which implies, for instance, that the method of moments will encounter difficulties in estimating θ.

Home page

Variance function I 📖

Let $Y \sim f(y;\theta)$, with $f(y;\theta) \in \mathcal{F}^1_{\text{en}}$ and let $\theta(\mu)$ be the **inverse map** of $\mu(\theta)$. The variance of Y can be expressed as a function of μ:

$$V(\mu) := \text{var}_{\theta(\mu)}(Y) = \frac{\partial^2}{\partial \theta^2} K(\theta)\Big|_{\theta=\theta(\mu)}.$$

The function $V : \mathcal{M} \to \mathbb{R}^+$ is called the **variance function** of the exponential family $\mathcal{F}^1_{\text{en}}$.

- The importance of the variance function $V(\mu)$ is related to the following **characterization** result due to Morris (1982).

Theorem (Pace and Salvan (1997), Theorem 5.2)

If Y has a density that belongs to a $\mathcal{F}^1_{\text{en}}$, then the **pair** $(\mathcal{M}, V(\mu))$ **uniquely** determine the natural parameter space $\tilde{\Theta}$ and the cumulant generating function $K(\theta)$, and hence also $f(y;\theta)$.

Variance function II 📖

- The characterization theorem of Morris (1982) is **constructive** in nature, as its **proof** provides a practical way of determining $K(\theta)$ from $(\mathcal{M}, V(\mu))$. In particular, the function $K(\cdot)$ must satisfy

$$K\left(\int_{\mu_0}^{\mu} \frac{1}{V(m)}\mathrm{d}m\right) = \int_{\mu_0}^{\mu} \frac{m}{V(m)}\mathrm{d}m,$$

  where μ_0 is an arbitrary point in $\mathcal{M}$.

- For example, let $\mathcal{M} = (0, \infty)$ and $V(\mu) = \mu^2$. Then, choosing $\mu_0 = 1$ gives

$$K\left(1 - \frac{1}{\mu}\right) = \log \mu,$$

  and therefore $\theta(\mu) = 1 - 1/\mu$, giving $\tilde{\Theta} = (-\infty, 1)$ and $\mu(\theta) = (1-\theta)^{-1}$. Hence we obtain $K(\theta) = -\log(1-\theta)$, which corresponds to the exponential density $f_0(y) = e^{-y}$, for $y > 0$.

> In order to identify $\mathcal{F}^1_{\text{en}}$ **both** $\mathcal{M}$ and $V(\mu)$ must be known.

Well-known exponential families

Notation	$\mathrm{N}(\psi, 1)$	$\mathrm{Poisson}(\psi)$	$\mathrm{Bin}(N, \psi)$	$\mathrm{Gamma}(\nu, \psi), \nu > 0$
$\mathcal{Y}$	$\mathbb{R}$	$\mathbb{N}$	$\{0, 1, \ldots, N\}$	$(0, \infty)$
Natural param.				
$\theta(\psi)$	ψ	$\log \psi$	$\log\{\psi/(1-\psi)\}$	$-\psi$
$f_0(y)$	$(\sqrt{2\pi})^{-1} e^{-\frac{1}{2}y^2}$	$e^{-1}/y!$	$\binom{N}{y}\left(\frac{1}{2}\right)^N$	$y^{\nu-1}e^{-y}/\Gamma(\nu)$
$K(\theta)$	$\theta^2/2$	$e^\theta - 1$	$N\log(1+e^\theta) - N\log 2$	$-\nu\log(1-\theta)$
$\tilde{\Theta}$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$(-\infty, 0)$
Mean param.				
$\mu(\theta)$	θ	e^θ	$Ne^\theta/(1+e^\theta)$	$-\nu/\theta$
$\mathcal{M}$	$\mathbb{R}$	$(0, \infty)$	$(0, N)$	$(0, \infty)$
$V(\mu)$	1	μ	$\mu(1-\mu/N)$	μ^2/ν

Quadratic variance functions

- There is more in Morris (1982)'s paper. Specifically, he focused on a subclass of **quadratic** variance functions, which can be written as

$$V(\mu) = a + b\mu + c\mu^2,$$

for some known constants a, b, and c.

- Morris (1982) showed that, up to transformations such as convolution, there exist only **six families** within $\mathcal{F}_{\text{en}}^1$ that possess a **quadratic variance** function. These are: (i) the normal, (ii) the Poisson, (iii) the gamma, (iv) the binomial, (v) the negative binomial, and (vi) a sixth family.
- The sixth (less known) distribution is called the **generalized hyperbolic secant**, and it has density

$$f(y;\theta) = \frac{\exp\{\theta y - \log\cos\theta\}}{2\cosh(\pi y/2)}, \qquad y \in \mathbb{R}, \quad \theta \in (-\pi/2, \pi/2),$$

with **mean** function $\mu(\theta) = \tan\theta$ and **variance** function $V(\mu) = \csc^2(\theta) = 1 + \mu^2$, and $\mathcal{M} = \mathbb{R}$. It is also a **regular** exponential family.

A general definition of exponential families I

Let $h(y) > 0$, $s(y)$, be real-valued functions not depending on ψ and let $\theta(\psi), G(\psi)$ be real-valued functions not depending on y. The parametric family

$$\mathcal{F}_{\mathrm{e}}^{1} = \{f(y;\psi) = h(y)\exp\{\theta(\psi)s(y) - G(\psi)\}, \quad y \in \mathcal{Y} \subseteq \mathbb{R},\, \psi \in \Psi\},$$

is called a **exponential family** of order one, where the normalizing constant is

$$\exp G(\psi) = \int_{y} h(y)\exp\{\theta(\psi)s(y)\}\nu(\mathrm{d}y).$$

The family is **full** if the parameter space Ψ is the widest possible $\tilde{\Psi} = \{\psi \subseteq \mathbb{R} : G(\psi) < \infty\}$.

Suppose $f(y;\psi) \in \mathcal{F}_{\mathrm{e}}^{1}$. Then, the function $\theta(\psi)$ must be a **one-to-one** mapping, that is, a **reparametrization**, otherwise, the model would **not** be **identifiable**. Hence, we can write:

$$f(y;\psi) = h(y)\exp\{\theta(\psi)s(y) - \tilde{G}(\theta(\psi))\},$$

for some function $\tilde{G}(\cdot)$ such that $G(\psi) = \tilde{G}(\theta(\psi))$.

A general definition of exponential families II

- When $s(y)$ is an arbitrary function of y, then $\mathcal{F}_{\text{e}}^1$ is **broader** than $\mathcal{F}_{\text{en}}^1$.
- Without loss of generality, we can focus on the natural parametrization $\theta \in \Theta$ and a density baseline $h(y) = f_0(y)$, meaning that $f(y;\theta) \in \mathcal{F}_{\text{e}}^1$ can be written as

$$f(y;\theta) = f_0(y)\exp\{\theta s(y) - K(\theta)\},$$

  because the general case would be a **reparametrization** of this one.
- Let $Y \sim f(y;\theta)$, with $f(y;\theta) \in \mathcal{F}_{\text{e}}^1$. Then, the random variable $S = s(Y)$ has density

$$f_S(s;\psi) = \tilde{f}_0(s)\exp\{\theta s - K(\theta)\},$$

  for some baseline density $\tilde{f}_0(s)$, namely $f_S(s;\psi) \in \mathcal{F}_{\text{en}}^1$. If in addition $s(y)$ is a **one-to-one** invertible mapping, this means $Y = s^{-1}(S)$ is just a transformation of an $\mathcal{F}_{\text{en}}^1$.

> A full exponential family $\mathcal{F}_{\text{e}}^1$ is, technically, a broader definition, but in practice it leads to a **reparametrization** of a natural exponential family $\mathcal{F}_{\text{en}}^1$ in a **transformed space** $s(Y)$.

Multiparameter exponential families

Natural exponential families of order p

- Let Y be a **non-degenerate** random variable with **support** $\mathcal{Y} \subseteq \mathbb{R}^p$ and **density** $f_0(y)$ with respect to a dominating measure $\nu(\mathrm{d}y)$.
- Let us define the mapping $M_0 : \mathbb{R}^p \to (0, \infty]$

$$M_0(\theta) := \int_{\mathcal{Y}} e^{\theta^T y} f_0(y) \nu(\mathrm{d}y), \qquad \theta \in \mathbb{R}^p.$$

The parametric family generated via **exponential tilting** of a density f_0

$$\mathcal{F}^p_{\mathrm{ne}} = \left\{ f(y;\theta) = \frac{e^{\theta^T y} f_0(y)}{M_0(\theta)} = f_0(y) \exp\{\theta^T y - K(\theta)\}, \quad y \in \mathcal{Y} \subseteq \mathbb{R}^p,\, \theta \in \tilde{\Theta} \right\},$$

is called a **natural exponential family** of order one, $K(\theta) = \log M_0(\theta)$ and $\tilde{\Theta} = \{\theta \in \mathbb{R}^p : K(\theta) < \infty\}$ is the **natural parameter space**.

- The family $\mathcal{F}^p_{\mathrm{ne}}$ is said to be **full**, whereas a subfamily of $\mathcal{F}^p_{\mathrm{ne}}$ with $\Theta \subseteq \tilde{\Theta}$ is **non-full**. Moreover, the family $\mathcal{F}^p_{\mathrm{ne}}$ is said to be **regular** if $\tilde{\Theta}$ is an open set.

Example: multinomial distribution I 📖

- Let $Y = (Y_1, \dots, Y_{p-1}) \sim \text{Multinom}(N; 1/p, \dots, 1/p)$ be a **multinomial** random vector with **uniform probabilities**, so that its density f_0 is

$$f_0(y) = \frac{N!}{y_1! \cdots y_p!} \left(\frac{1}{p}\right)^N, \qquad y = (y_1, \dots, y_{p-1}) \in \mathcal{Y} \subseteq \mathbb{R}^{p-1},$$

  where $\mathcal{Y} = \{(y_1, \dots, y_{p-1}) \in \{0, \dots, N\}^{p-1} : \sum_{j=1}^{p-1} y_j \leq N\}$, having set $y_p := N - \sum_{j=1}^{p-1} y_j$.

- The **exponential tilting** of f_0 yields

$$f(y; \theta) \propto f_0(y) e^{\theta^T y} = \frac{N!}{y_1! \cdots y_p!} \left(\frac{1}{p}\right)^N e^{\theta_1 y_1 + \cdots + \theta_{p-1} y_{p-1}}, \qquad y \in \mathcal{Y},\ \theta \in \mathbb{R}^{p-1}.$$

- As a consequence of the **multinomial theorem**, the normalizing constant, that is, the **moment generating function**, is

$$M_0(\theta) = \mathbb{E}\left(e^{\theta^T Y}\right) = \left(\frac{1}{p}\right)^N (1 + e^{\theta_1} + \cdots + e^{\theta_{p-1}})^N.$$

  Thus $M_0(\theta) < \infty$ for all $\theta \in \mathbb{R}^{p-1}$ and the **natural parameter space** is the **open set** $\tilde{\Theta} = \mathbb{R}^{p-1}$.

Example: multinomial distribution II 📖

- The resulting **tilted** density is

$$f(y;\theta) = f_0(y)e^{\theta^T y - K(\theta)} = \frac{N!}{y_1!\cdots y_p!}\frac{e^{\theta_1 y_1 + \cdots + \theta_{p-1}y_{p-1}}}{(1+e^{\theta_1}+\cdots+e^{\theta_{p-1}})^N},$$

  where $K(\theta) = \log M_0(\theta) = N\log(1+e^{\theta_1}+\cdots+e^{\theta_{p-1}}) - N\log p$.

- In other words, the tilted density is again a **multinomial** distribution with parameters N and **probabilities** $\pi_j = e^{\theta_j}/(1+e^{\theta_1}+\cdots+e^{\theta_{p-1}})$. In fact, we can write:

$$\begin{aligned} f(y;\theta) &= \frac{N!}{y_1!\cdots y_p!}\frac{e^{\theta_1 y_1}\cdots e^{\theta_p y_p}}{(\sum_{j=1}^p e^{\theta_j})^{y_1}\cdots(\sum_{j=1}^p e^{\theta_j})^{y_p}} = \frac{N!}{y_1!\cdots y_p!}\prod_{j=1}^p\left(\frac{e^{\theta_j}}{\sum_{k=1}^p e^{\theta_k}}\right)^{y_j} \\ &= \frac{N!}{y_1!\cdots y_p!}\prod_{j=1}^p \pi_j^{y_j}. \end{aligned}$$

  where we defined $\theta_p := 0$, so that $\sum_{j=1}^p e^{\theta_j} = 1 + e^{\theta_1} + \cdots + e^{\theta_{p-1}}$, recalling that $\sum_{j=1}^p y_j = N$.

- The tilted density belongs to a regular **natural exponential family** $\mathcal{F}_{\text{en}}^{p-1}$ of **order** $p-1$.

Example: independent exponential families 📖

- Let $Y = (Y_1, \ldots, Y_p)$ be a random vector of **independent** random variables, each belonging to a **full natural exponential family** $\mathcal{F}_{\text{en}}^1$ of order 1, with density

$$f(y_j; \theta_j) = f_j(y_j) \exp\{\theta_j y_j - K_j(\theta_j)\}, \qquad \theta_j \in \tilde{\Theta}_j.$$

- Let $\theta = (\theta_1, \ldots, \theta_p)$. Because of independence, the **joint distribution** of Y is

$$\begin{aligned} f(y; \theta) &= \prod_{j=1}^{p} f(y_j; \theta_j) = \prod_{j=1}^{p} f_j(y_j) \exp\{\theta_j y_j - K_j(\theta_j)\} \\ &= \left[\prod_{j=1}^{p} f_j(y_j)\right] \exp\left\{\sum_{j=1}^{p} \theta_j y_j - \sum_{j=1}^{p} K_j(\theta_j)\right\} \\ &= f_0(y) \exp\{\theta^T y - K(\theta)\}, \end{aligned}$$

  where $f_0(y) = \prod_{j=1}^{p} f_j(y_j)$, $K(\theta) = \sum_{j=1}^{p} K_j(\theta_j)$, and the **natural parameter space** is

$$\tilde{\Theta} = \tilde{\Theta}_1 \times \cdots \times \tilde{\Theta}_p.$$

- Thus, $f(y; \theta)$ is an $\mathcal{F}_{\text{en}}^p$, in which $K(\theta)$ is a **separable** function.

Mean value mapping and other properties

- Let $Y \sim f(y;\theta)$, with $f(y;\theta) \in \mathcal{F}^p_{\text{en}}$. The cumulant generating function is

$$K_\theta(t) = \log M_\theta(t) = K_0(t+\theta) - K_0(\theta), \qquad t+\theta \in \tilde{\Theta}.$$

  In particular, the first two moments of Y are obtained as:

$$\mu(\theta) := \mathbb{E}_\theta(Y) = \frac{\partial}{\partial \theta} K(\theta), \qquad \text{var}_\theta(Y) = \frac{\partial}{\partial \theta^\top} \mu(\theta) = \frac{\partial^2}{\partial \theta \partial \theta^\top} K(\theta),$$

- If f_0 is non-degenerate, then the **covariance matrix** $\text{var}_\theta(Y)$ is **positive definite**, implying that $K(\theta)$ is a **convex function**, and $\mu(\theta)$ is a **smooth one-to-one** map.
- The definitions of mean value mapping $\mu(\theta)$, its range $\mathcal{M}$, the convex hull $C(\mathcal{Y})$ of the sample space, and the variance function $V(\mu)$ also naturally extend to the multi-parameter setting.
- Refer to Jorgensen (1987) for an extension of the results of Morris (1982) about $V(\mu)$.

Theorem (Pace and Salvan (1997), Theorem 5.3)

If $\mathcal{F}^p_{\text{en}}$ is regular, then $\mathcal{M} = \text{int}\, C$.

Independence of the components

Theorem (Pace and Salvan (1997), Theorem 5.4)

If the natural observations of an $\mathcal{F}_{\text{en}}^{p}$ are **independent** for some $\theta_0 \in \tilde{\Theta}$, then this is also true for every $\theta \in \tilde{\Theta}$.

- This theorem essentially establishes that if the baseline density $f_0(\cdot)$ has independent components, then the exponential tilting preserves independence.

Theorem (Pace and Salvan (1997), Theorem 5.5)

If, for every $\theta \in \tilde{\Theta}$, the natural observations of a **regular** $\mathcal{F}_{\text{en}}^{p}$ are **uncorrelated**, then they are also **independent**.

- This generalizes a well-known fact about multivariate Gaussians, which are in fact an $\mathcal{F}_{\text{en}}^{p}$.
- In practice, if the Hessian matrix of $K(\theta)$ is **diagonal**, then the natural observations are **independent**. This occurs whenever $K(\theta)$ is **separable**.

Marginal and conditional distributions

- Consider a $\mathcal{F}_{\text{en}}^{p}$ family, so that $f(y;\theta) = f_0(y)\exp\{\theta^T y - K(\theta)\}$.
- Let $y = (t,u)$ be a **partition** of the natural observations y, where t has k components and u has $p-k$ components. Let us partition θ accordingly, so that $\theta = (\tau,\zeta)$ and

$$f(y;\tau,\zeta) = f_0(y)\exp\{\tau^T t + \zeta^T u - K(\tau,\zeta)\}, \qquad (\tau,\zeta)\in\tilde{\Theta}.$$

Theorem (Pace and Salvan (1997), Theorem 5.6)

i. The family of **marginal** distributions of U is an $\mathcal{F}_{\text{en}}^{p-k}$ for every fixed value of τ and

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

ii. The family of **conditional** distributions of T given $U=u$ is an $\mathcal{F}_{\text{en}}^{k}$ and the conditional densities do not depend on ζ, that is

$$f_{T|U=u}(t;u,\tau) = h_u(t)\exp\{\tau^T t - K_u(\tau)\}, \quad \exp K_u(\tau) = \mathbb{E}_0\left(e^{\tau^T T} \mid U=u\right).$$

Conditional likelihoods

- The former result on marginal and conditional laws is not just an elegant probabilistic fact. Indeed, it has meaningful inferential applications.
- Often, we can split the parameter vector θ into a **parameter of interest** τ and a **nuisance parameter** ζ. We are not interested in learning ζ.

> The main idea relies on noticing that $f_{T|U=u}(t; u, \tau) = h_u(t) \exp\{\tau^T t - K_u(\tau)\}$ does not involve ζ and therefore we could define a **conditional likelihood** based on $f_{T|U=u}$.

- A practical drawback of this approach is that the conditional cumulant generating function $K_u(\tau)$ is not always available in closed form, albeit with notable exceptions.
- The approach is valid, in the sense that a likelihood based on $f_{T|U=u}$ is a **genuine likelihood**. On the other hand, note that the full likelihood would be based on

$$f(y; \tau, \zeta) = f_U(u; \tau, \zeta)\, f_{T|U=u}(t; u, \tau),$$

and thus the conditional likelihood is **discarding information**, that is, it neglects $f_U(u; \tau, \zeta)$.

A general definition of exponential families I

Let $s_1(y), \ldots, s_p(y)$ and $h(y) > 0$ be real-valued functions not depending on the parameter ψ, and let $\theta_1(\psi), \ldots, \theta_p(\psi)$, $G(\psi)$ be real-valued functions not depending on y. The family

$$\mathcal{F}_{\mathrm{e}}^p = \left\{ f(y;\psi) = h(y) \exp\{\theta(\psi)^T s(y) - G(\psi)\}, \quad y \in \mathcal{Y} \subseteq \mathbb{R}^p,\ \psi \in \Psi \subseteq \mathbb{R}^q \right\},$$

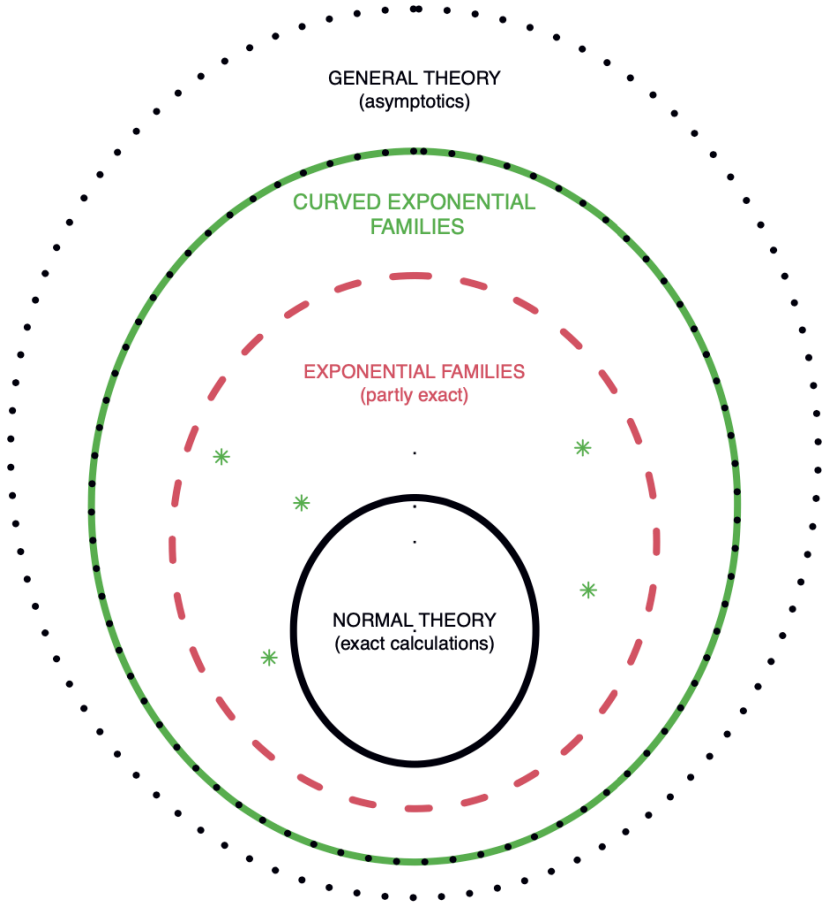
is called an **exponential family** of order p, where the normalizing constant is

$$\exp G(\psi) = \int_{\mathcal{Y}} h(y) \exp\{\theta(\psi)^T s(y)\} \nu(\mathrm{d}y).$$

The notation $\mathcal{F}_{\mathrm{e}}^p$ is understood to indicate a **minimal representation**, i.e., such that there is **no linear dependence** between $1, s_1(y), \ldots, s_p(y)$ or, equivalently, between $1, \theta_1(\psi), \ldots, \theta_p(\psi)$.

- If $q > p$, then ψ is **not identifiable** and this possibility should be discarded.
- If $q = p$, then $\theta(\psi)$ must be a one-to-one mapping, i.e., a **reparametrization**, otherwise the model is again **not identifiable**.
- If $q < p$, we have a (p, q)-**curved exponential family**, which corresponds to a **restriction** of the natural parameter space.

Curved exponential families



- Figure 4.1 of Efron (2023), Chapter 4. Three levels of statistical modeling, now with a fourth level added representing curved exponential families.

Home page

A general definition of exponential families II

- We refer to Efron (2023), Chapter 4, for a detailed discussion on curved exponential families. From now on, we will focus on the $p = q$ case.
- Without loss of generality, we can focus on the **natural parametrization** $\theta \in \Theta \subseteq \mathbb{R}^p$ and baseline density $h(y) = f_0(y)$, meaning that $f(y;\theta) \in \mathcal{F}_{\text{e}}^p$ can be written as

$$f(y;\theta) = f_0(y) \exp\{\theta^T s(y) - K(\theta)\},$$

  because the general case would be a **reparametrization** of this one.
- Let $Y \sim f(y;\theta)$, with $f(y;\theta) \in \mathcal{F}_{\text{e}}^p$. Then, the **random vector** $S = s(Y) = (s_1(Y), \ldots, s_p(Y))$ has density

$$f_S(s;\theta) = \tilde{f}_0(s) \exp\{\theta^T s - K(\theta)\},$$

  for some baseline density $\tilde{f}_0(s)$, namely $f_S(s;\theta) \in \mathcal{F}_{\text{en}}^p$. If in addition $s(y)$ is a **one-to-one** invertible mapping, this means $Y = s^{-1}(S)$ is just a transformation of an $\mathcal{F}_{\text{en}}^p$.

As in the single parameter case, a full exponential family $\mathcal{F}_{\text{e}}^p$ with $p = q$ in practice leads to a **reparametrization** of a natural exponential family $\mathcal{F}_{\text{en}}^p$ in a **transformed space** $s(Y)$.

Example: gamma distribution 📖

- The family $\text{Gamma}(\nu, \lambda)$ with $\nu, \lambda > 0$ is an $\mathcal{F}_e^2$. In fact, its **density** is

$$\begin{aligned} f(y; \nu, \lambda) &= \frac{\lambda^\nu}{\Gamma(\nu)} y^{\nu-1} e^{-\lambda y} = \frac{1}{y} \exp\{\nu \log y - \lambda y - \log \Gamma(\nu) + \nu \log \lambda\} \\ &= h(y) \exp\{\theta(\psi)^T s(y) - G(\psi)\}. \end{aligned}$$

  where $h(y) = y^{-1}$, the **sufficient statistic** $s(y) = (s_1(y), s_2(y)) = (\log y, y)$, whereas the **natural parameters** and the cumulant generating function are

$$\theta(\psi) = (\theta_1(\psi), \theta_2(\psi)) = (\nu, -\lambda), \qquad G(\psi) = \log \Gamma(\nu) - \nu \log \lambda,$$

  having set $\psi = (\nu, \lambda)$.

- As previously shown, this implies that the family

$$f(s; \theta) = \tilde{h}(s) \exp\{\theta^T s - \log \Gamma(\theta_1) + \theta_1 \log(-\theta_2)\}, \qquad \theta \in \tilde{\Theta},$$

  is a regular **natural exponential family** of order 2, with some function $\tilde{h}(s)$.

Example: von Mises distribution I

- Let Y be a random variable describing an **angle**, so that $\mathcal{Y} = (0, 2\pi)$, and let us consider the **uniform density** on the **circle**, namely

$$f_0(y) = \frac{1}{2\pi}, \qquad y \in (0, 2\pi).$$

- We define a tilted density $f(y;\theta) \in \mathcal{F}_e^2$ by considering $s(y) = (\cos y, \sin y)$, i.e., the **cartesian coordinates** of y. This choice of $s(y)$ ensures the appealing property $f(y;\theta) = f(y + 2k\pi;\theta)$.
- More precisely, let $\theta = (\theta_1, \theta_2)$ and define the parametric family of densities

$$f(y;\theta) = f_0(y)\exp\{\theta^T s(y) - K(\theta)\}, \qquad \theta \in \tilde{\Theta},$$

where $h(y) = 1/2\pi$. The normalizing constant has a "closed form"

$$\exp K(\theta) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{\theta_1\cos(y) + \theta_2\sin(y)\}\mathrm{d}y = \mathcal{A}_0(||\theta||_2),$$

where $\mathcal{A}_\nu(\cdot)$ is known as the **modified Bessel function** of the first kind and order ν.

- It is easy to check that $K(\theta) < \infty$ for all values of $\theta \in \mathbb{R}^2$; therefore, $\tilde{\Theta} = \mathbb{R}^2$. This completes the definition of what is known as the **von Mises** distribution.

Example: von Mises distribution II

- Instead of the **natural parametrization**, it is often convenient to consider a **reparametrization** $\psi = (\tau, \gamma)$, defined through the one-to-one mapping

$$\theta(\psi) = (\tau \cos\gamma, \tau \sin\gamma), \qquad \psi \in \tilde{\Psi} = (0, \infty) \times (0, 2\pi).$$

- Using this parametrization, thanks to well-known **trigonometric** relationships, we obtain the more familiar formulation of the von Mises distribution, which is

$$f(y; \psi) = h(y) \exp\{\theta(\psi) s(y) - G(\psi)\} = \frac{1}{2\pi \mathcal{A}_0(\tau)} e^{\tau \cos(y-\gamma)}, \qquad y \in (0, 2\pi),$$

  so that $\gamma \in (0, 2\pi)$ can be interpreted as the **location** and $\tau > 0$ as the **precision**.

- We also note that the distribution of $s(Y)$ is a **regular natural exponential family** of order 2, with density

$$f_S(s; \theta) = \frac{1}{2\pi} \exp\{\theta^T s - \log \mathcal{A}_0(||\theta||_2)\}, \qquad s \in \mathcal{S} = \{(s_1, s_2) \in \mathbb{R}^2 : s_1^2 + s_2^2 = 1\},$$

  clarifying that $S = s(Y)$ is a random vector taking values on a **circle** with unit radius.

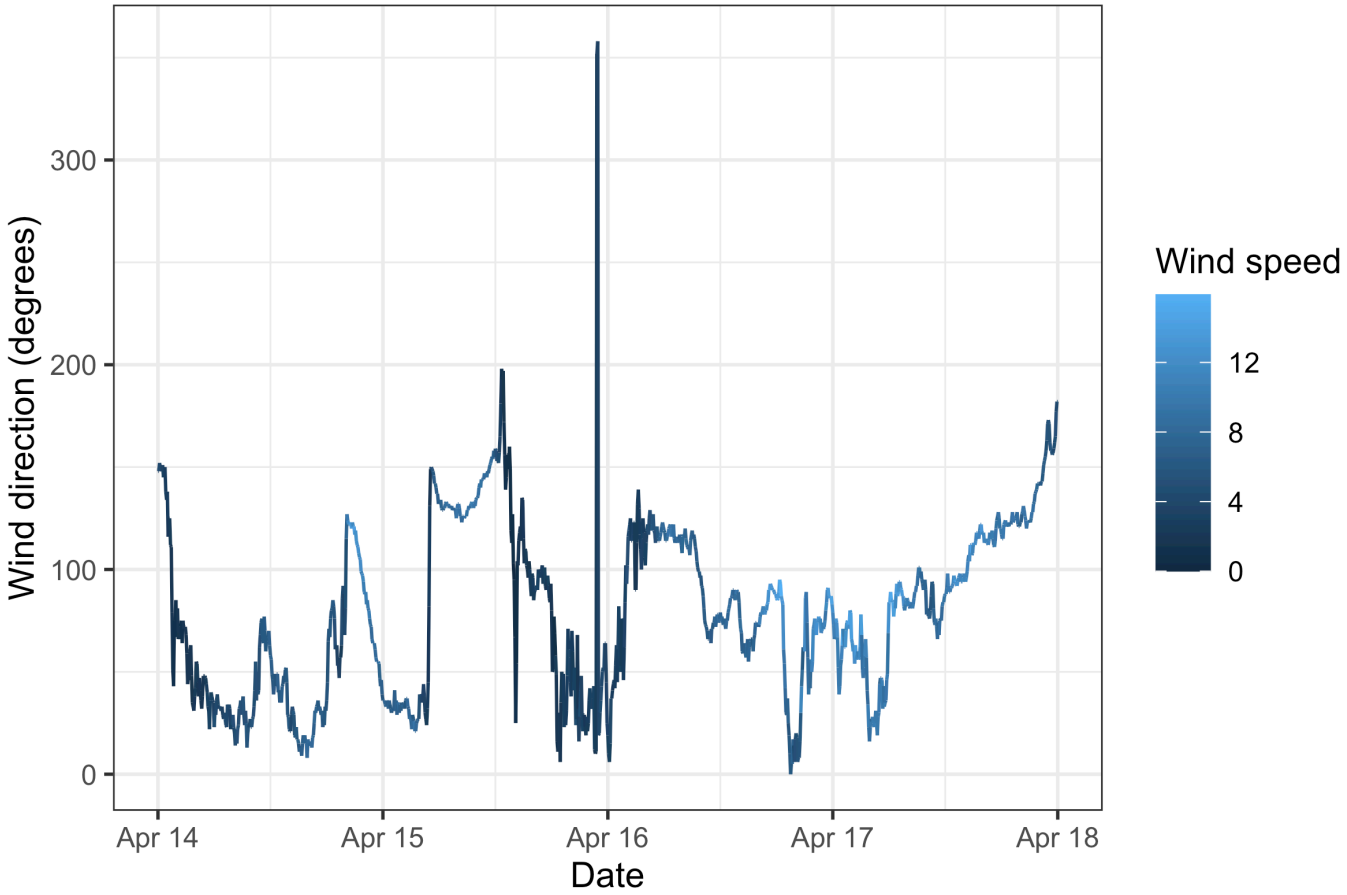
Example: wind direction in Venice I

- The von Mises distribution is sometimes regarded as the "Gaussian distribution for **circular data**". To provide a concrete example, let us consider the **wind directions** measured from the San Giorgio meteorological station, in Venice.
- Measurements are recorded every **5 minutes**, from 14-04-2025 to 18-04-2025, for a total of $n = 1153$. The variable `wind_dir` is recorded in **degrees**, i.e., between 0 and 360.

```
# A tibble: 10 × 3
   date                wind_dir `Wind speed`
   <dttm>                 <dbl>        <dbl>
 1 2025-04-14 00:00:00      148          4.6
 2 2025-04-14 00:05:00      148          4.4
 3 2025-04-14 00:10:00      152          4.1
 4 2025-04-14 00:15:00      150          4.1
 5 2025-04-14 00:20:00      150          4
 6 2025-04-14 00:25:00      148          3.8
 7 2025-04-14 00:30:00      151          3.3
 8 2025-04-14 00:35:00      145          3
 9 2025-04-14 00:40:00      148          3.5
10 2025-04-14 00:45:00      150          2.9
```

The dataset is available here. The original source is the webpage of Venice municipality.

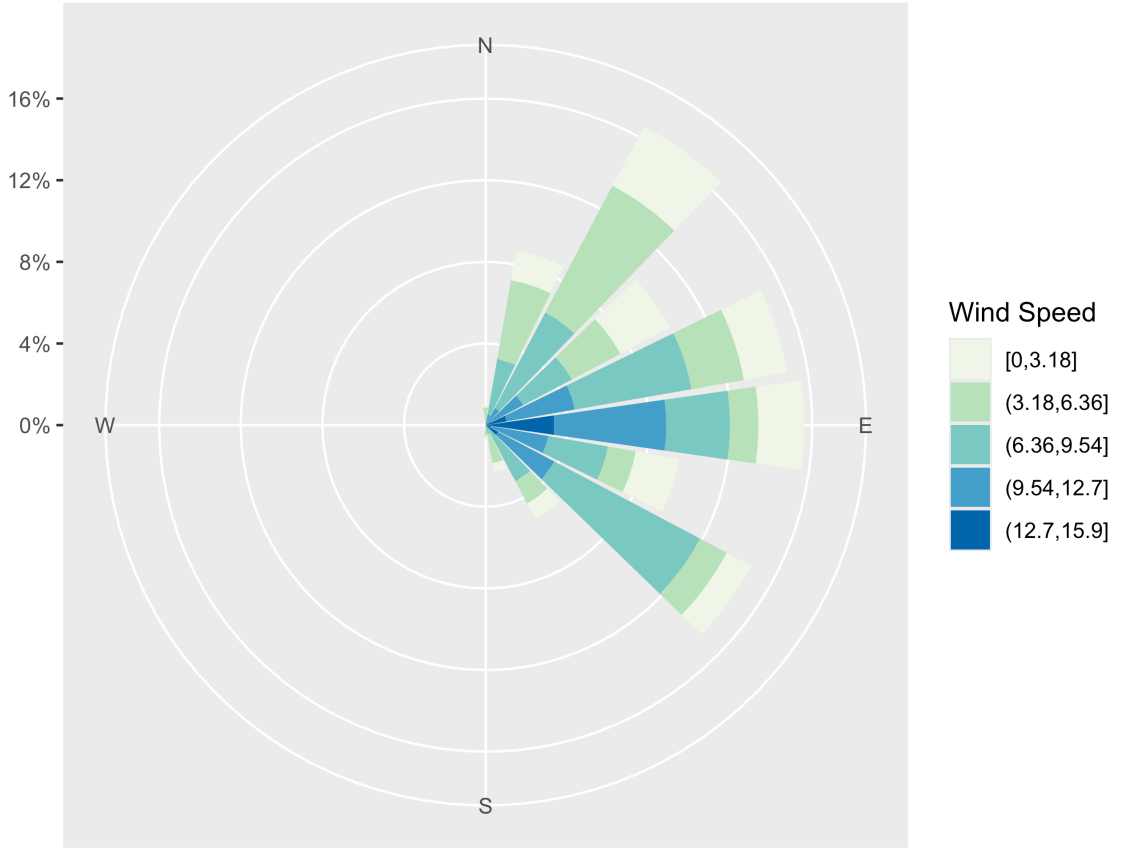
Example: wind direction in Venice II



- This is a somewhat **misleading** graphical representation of **wind directions** evolving over time. Indeed, the "spikes" are not real: the angles 1 and 359 are, in fact, very close.

Home page

Example: wind direction in Venice III



- A better graphical representation of **wind directions** and **wind speed**, using Cartesian coordinates. From this wind rose, it is clear the winds were coming mostly from the east.

Home page

Inference

Independent sampling, sufficiency and completeness

- Let $Y_1, \ldots, Y_n$ be iid random vectors with density $f(y;\theta)$, where $f(y;\theta) \in \mathcal{F}_{\text{e}}^p$ and, without loss of generality, we let $f(y;\theta) = f_0(y)\exp\{\theta^T s(y) - K(\theta)\}$. The **likelihood** function is

$$L(\theta; \boldsymbol{y}) = \prod_{i=1}^n \exp\left\{\theta^T s(y_i) - K(\theta)\right\} = \exp\left\{\theta^T \sum_{i=1}^n s(y_i) - nK(\theta)\right\},$$

  from which we see that $s = \sum_{i=1}^n s(y_i) = (\sum_{i=1}^n s_1(y_i), \ldots, \sum_{i=1}^n s_p(y_i))$ is the **minimal sufficient statistic** as long as $n \geq p$, which has **fixed dimension** p whatever the sample size.

- Inference can therefore be based on the random vector $S = \sum_{i=1}^n s(Y_i)$, whose distribution is

$$f_S(s;\theta) = \tilde{f}_0(s)\exp\{\theta^T s - \tilde{K}(\theta)\},$$

  with $\tilde{K}(\theta) = nK(\theta)$ and for some density $\tilde{f}_0(s)$. In other words, $f_S(s;\theta) \in \mathcal{F}_{\text{en}}^p$.

Theorem (Pace and Salvan (1997), Theorem 5.7)

A sufficient statistic S with distribution $\mathcal{F}_{\text{en}}^p$ is **complete**, provided that $\text{int}\,\tilde{\Theta} \neq \emptyset$.

Sufficiency and completeness

- Thus, the log-likelihood function, after a reduction via **sufficiency**, is

$$\ell(\theta) = \ell(\theta; s) = \theta^T s - nK(\theta), \qquad \theta \in \tilde{\Theta},$$

  with $S = \sum_{i=1}^{n} s(Y_i)$ being distributed as a $\mathcal{F}_{\text{en}}^{p}$ with cumulant generating function $nK(\theta)$, whereas each $s(Y_i)$ is distributed as a $\mathcal{F}_{\text{en}}^{p}$ with cumulant generating function $K(\theta)$.

- The **completeness** of S in exponential families is a classical result that enables the usage of the Rao-Blackwell-Lehmann-Scheffé theorem for finding the UMVUE.
- Moreover, the existence of a **minimal sufficient** statistic that performs a **non-trivial dimensionality reduction**, from n to p and with $p \leq n$, is a major simplification.
- This only occurs in exponential families, except for non-regular cases.

Theorem (Koopman-Pitman-Darmois, Robert (1994), Theorem 3.3.3)

Under iid sampling, if a parametric family whose **support** does **not depend** on the **parameter** is such that there exists a **sufficient statistic** of **constant dimension** p, then the family is $\mathcal{F}_{\text{e}}^{p}$.

Likelihood quantities

- After a sufficiency reduction, we get $\ell(\theta) = \theta^T s - nK(\theta)$. Thus, the **score function** is

$$\ell^*(\theta) = s - n\frac{\partial}{\partial\theta}K(\theta) = s - n\mu(\theta),$$

  where $\mu(\theta) = \mathbb{E}_\theta(s(Y_1))$ is the **mean value mapping** of each $s(Y_i)$ and $n\mu(\theta) = \mathbb{E}(S)$.

- By direct calculation, we show that the **first Bartlett identity** holds, namely

$$\mathbb{E}_\theta(\ell^*(\theta; S)) = \mathbb{E}_\theta(S) - n\mu(\theta) = n\mu(\theta) - n\mu(\theta) = \mathbf{0}.$$

  The **Fisher information** is straightforward to compute, being equal to

$$I(\theta) = \mathbb{E}_\theta(\ell^*(\theta)\ell^*(\theta)^T) = \mathbb{E}_\theta\{(S - n\mu(\theta))(S - n\mu(\theta))^T\} = \mathrm{var}_\theta(S) = n\,\mathrm{var}_\theta(s(Y_1)).$$

- Moreover, the **observed information** is

$$\mathcal{I}(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\tilde{K}(\theta) = n\frac{\partial^2}{\partial\theta\partial\theta^T}K(\theta) = n\,\mathrm{var}_\theta(s(Y_1)),$$

  which proves the **second Bartlett identity** as an implication of the **remarkable** identity $\mathcal{I}(\theta) = I(\theta)$, stronger than the usual $I(\theta) = \mathbb{E}_\theta(\mathcal{I}(\theta))$. In fact, $\mathcal{I}(\theta)$ is **non-stochastic**.

Existence of the maximum likelihood

- The maximum likelihood estimate $\hat{\theta}$, if it exists, is the **unique** solution of the **score equation**

$$s - n\mu(\theta) = \mathbf{0}, \qquad \text{so that} \qquad \hat{\theta} = \mu^{-1}\left(\frac{s}{n}\right) = \mu^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} s(y_i)\right).$$

  It is unique because $\ell(\theta)$ is **concave** in θ, namely its second derivative is

$$\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(\theta) = -\mathrm{var}_\theta(S) < 0, \qquad \theta \in \tilde{\Theta}.$$

Theorem (Pace and Salvan (1997), Theorem 5.8)

If $\mathcal{F}^p_{\text{en}}$ is regular, then the maximum likelihood estimate $\hat{\theta}$ exists and is the unique solution of $\ell^*(\theta) = \mathbf{0}$ if and only if $s \in \text{int}\, C(\mathcal{S})$, where $C(\mathcal{S})$ is the closed convex hull of the support of $\mathcal{S}$.

- As a corollary, if $\mathcal{F}^p_{\text{en}}$ is regular, the MLE exists and is unique **with probability one** if and only if the boundary of $C = C(\mathcal{S})$ has probability 0. This is often violated when S is **discrete**.

Likelihood quantities: mean parametrization 📖

- Let us consider the **mean parametrization** $\mu = \mu(\theta) = \mathbb{E}_\theta(s(Y_1))$, whose inverse is $\theta = \theta(\mu)$. The log-likelihood is:

$$\ell(\mu) = \ell(\theta(\mu)) = \theta(\mu)^T s - nK(\theta(\mu)), \qquad \mu \in \mathcal{M}.$$

- Hence, using the **chain rule** of differentiation, we obtain the **score**

$$\ell^*(\mu) = \left(\frac{\partial}{\partial \mu}\theta(\mu)\right)(s - n\mu) = \mathrm{var}_\mu(s(Y_1))^{-1}(s - n\mu),$$

  where the last step follows from the properties of the derivatives of inverse functions.

- Thus, the **observed information** matrix for the mean parametrization is

$$\mathcal{I}_\mu(\mu) = -\frac{\partial^2}{\partial \mu \partial \mu^T}\ell(\mu) = -\left(\frac{\partial^2}{\partial \mu \partial \mu^T}\theta(\mu)\right)(s - n\mu) + n\,\mathrm{var}_\mu(s(Y_1))^{-1},$$

  whereas the **Fisher information** matrix for μ is

$$I_\mu(\mu) = \mathbb{E}_\mu(\mathcal{I}_\mu(\mu)) = n\,\mathrm{var}_\mu(s(Y_1))^{-1} = n\,V(\mu)^{-1}.$$

Maximum likelihood: mean parametrization

- Thus, the maximum likelihood estimate of the **mean parametrization** $\hat{\mu} = \mu(\hat{\theta})$ is

$$\hat{\mu} = \frac{s}{n} = \frac{1}{n}\sum_{i=1}^{n} s(y_i).$$

  This means $\hat{\mu}$ is both the **maximum likelihood** and the **method of moments** estimate of μ.

- It is also an **unbiased estimator**, because by definition

$$\mathbb{E}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mu}(s(Y_i)) = \mathbb{E}_{\theta}(s(Y_1)) = \mu.$$

- Furthermore $\hat{\mu}$ is the **UMVUE** of μ. Indeed, we could first notice that $\hat{\mu}$ is a function of S, which is a **complete** sufficient statistic Alternatively, we could note that the variance of $\hat{\mu}$ is

$$\mathrm{var}_{\mu}(\hat{\mu}) = \frac{1}{n}\mathrm{var}_{\mu}(s(Y_1)) = \frac{1}{n}V(\mu) = \mathcal{I}_{\mu}(\mu)^{-1},$$

  which corresponds to the **Cramer-Rao** lower bound.

Example: binomial distribution 📖

- Let $Y_1, \ldots, Y_n$ be iid Bernoulli random variables with mean $\mu \in (0,1)$, that is $\text{pr}(Y_i = 1) = \mu$. Then, the log-likelihood function is

$$\ell(\mu) = \sum_{i=1}^{n} [y_i \log \mu + (1 - y_i) \log (1 - \mu)] = s \log \mu + (n - s) \log (1 - \mu),$$

  with $S = \sum_{i=1}^{n} Y_i$ being the **minimal sufficient** statistic and the **natural parametrization** is $\theta(\mu) = \log \mu/(1 - \mu)$. Note that $S \sim \text{Binom}(n, \mu)$.

- The **variance function** is $V(\mu) = \text{var}_\mu(Y_i) = \mu(1 - \mu)$, so that the **score function** becomes

$$\ell^*(\mu) = \frac{s}{\mu} - \frac{n - s}{1 - \mu} = \frac{1}{V(\mu)}(s - n\mu),$$

  leading to the well-known UMVUE maximum likelihood estimator $\hat{\mu} = s/n$.

- Finally, the **observed information** and the **Fisher information** equal, respectively

$$\mathcal{I}_\mu(\mu) = \frac{s}{\mu^2} - \frac{n - s}{(1 - \mu)^2}, \qquad I_\mu(\mu) = \mathbb{E}_\mu(\mathcal{I}_\mu(\mu)) = \frac{n}{\mu(1 - \mu)} = \frac{n}{V(\mu)}.$$

Example: von Mises distribution III 📖

- Let $Y_1, \dots, Y_n$ be iid random variables from a **Von-Mises** distribution with density $f(y; \psi) = (2\pi \mathcal{A}_0(\tau))^{-1} \exp\{\tau \cos(y - \gamma)\}$, with $y \in (0, 2\pi)$, therefore the **log-likelihood** is

$$\ell(\psi) = \tau \sum_{i=1}^{n} \cos(y_i - \gamma) - n \log \mathcal{A}_0(\tau).$$

- The Jacobian of the log-likelihood is

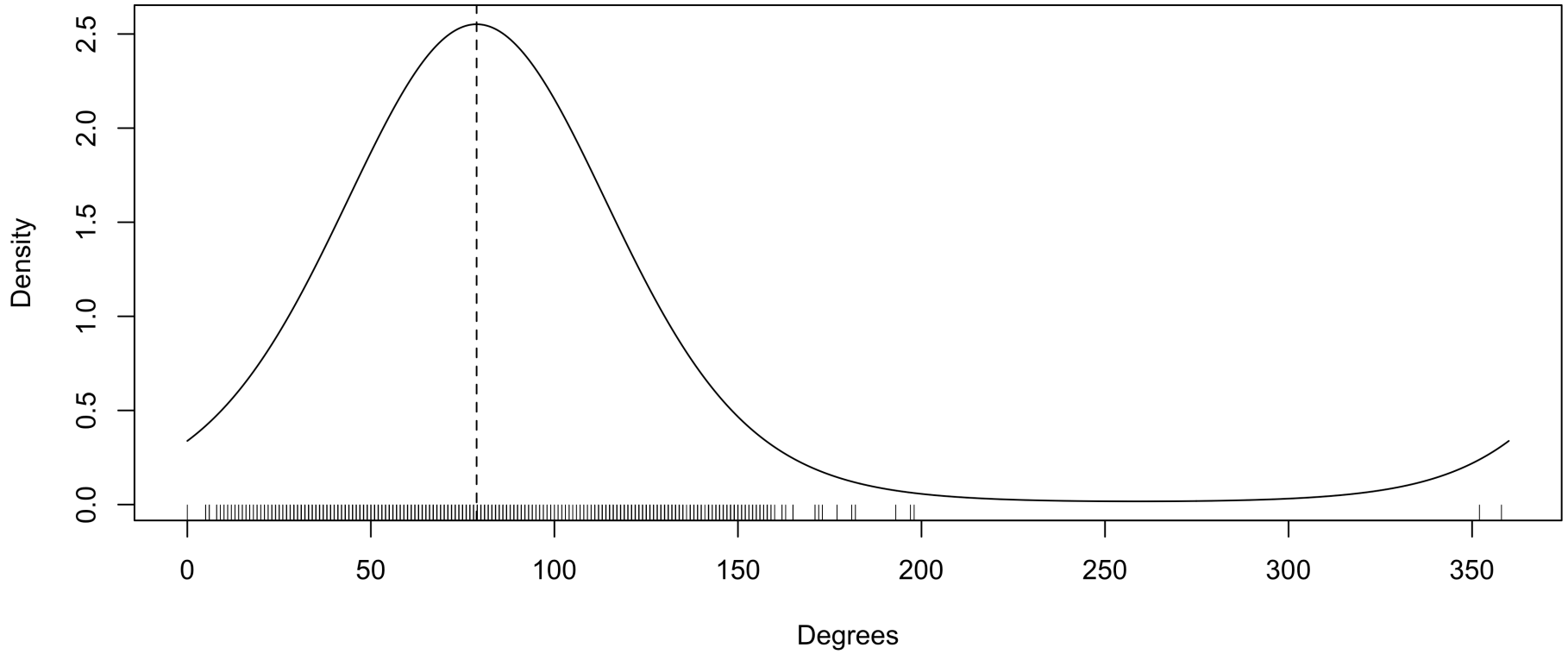
$$\frac{\partial}{\partial \gamma} \ell(\psi) = \tau \sum_{i=1}^{n} \sin(y_i - \gamma), \qquad \frac{\partial}{\partial \tau} \ell(\psi) = \sum_{i=1}^{n} \cos(y_i - \gamma) - n \frac{\mathcal{A}_1(\tau)}{\mathcal{A}_0(\tau)}.$$

- Thus, the **maximum likelihood estimate** $(\hat{\gamma}, \hat{\tau})$ is the solution of the following equations

$$\tan(\hat{\gamma}) = \frac{\sum_{i=1}^{n} \sin y_i}{\sum_{i=1}^{n} \cos y_i}, \qquad \frac{1}{n} \sum_{i=1}^{n} \cos(y_i - \hat{\gamma}) = \frac{\mathcal{A}_1(\hat{\tau})}{\mathcal{A}_0(\hat{\tau})}.$$

  The estimate for τ can be obtained **numerically** e.g. using the `circular::A1inv` function.

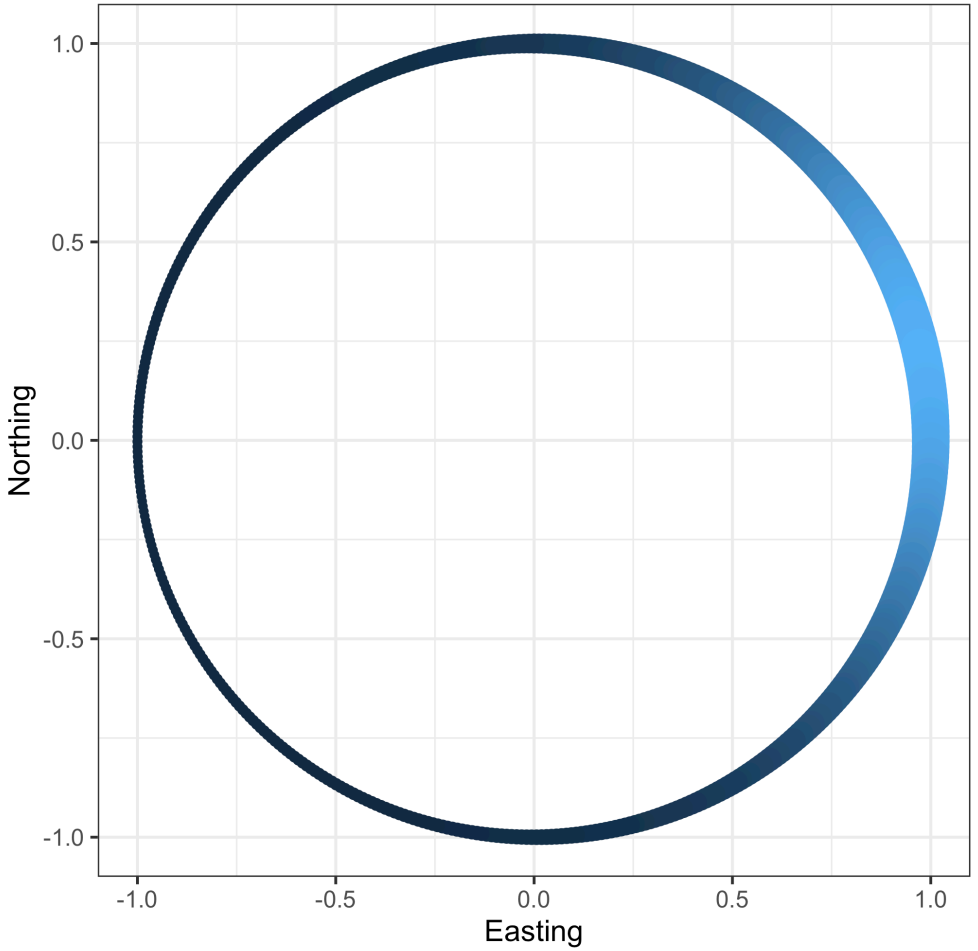
Example: wind direction in Venice IV



- The **estimated** values are $\hat{\gamma} = 1.375$ (corresponding to about 79 degrees) and $\hat{\tau} = 2.51$.

Home page

Example: wind direction in Venice V



Home page

Asymptotic theory: remarks

- Let us consider an iid sample from a model such that the minimal sufficient statistic belongs to a **regular exponential family** $\mathcal{F}_{\text{en}}^{p}$, with natural parameter $\theta \in \tilde{\Theta}$.
- It is straightforward to verify that the **regularity conditions A1–A6** from Unit A are **all satisfied**. Thus, Theorem 5.1 of Lehmann and Casella (1998) applies directly.
- We also proved that, if the score function has a root, then the maximum likelihood estimate $\hat{\theta}$ exists and is the **unique solution** of $\ell^*(\theta) = \mathbf{0}$, where $\ell^*(\theta) = s - n\mu(\theta)$.
- The maximum likelihood estimate may fail to exist if s lies on the boundary of $C(\mathcal{S})$. However, as $n \to \infty$, the probability that s lies on the boundary of $C(\mathcal{S})$ tends to zero.
- Indeed, by the law of large numbers, S/n converges almost surely to $\mu(\theta) \in \mathcal{M} = \operatorname{int} \mathcal{C}(S)$, implying that a unique root of the score function eventually exists with probability one.

> If the observations are iid from a **regular** exponential family, the maximum likelihood estimator $\hat{\theta}$ is consistent and asymptotically normal for θ. By the **continuous mapping theorem**, this implies that $\hat{\mu} = \mu(\hat{\theta})$, or any other **smooth reparametrization**, is a consistent estimator of μ.

Wald inequality: a direct proof 📖

- Let us recall that **Wald inequality** states that

$$\mathbb{E}_{\theta_0}\left(\ell(\theta;\boldsymbol{Y})\right) < \mathbb{E}_{\theta_0}\left(\ell(\theta_0;\boldsymbol{Y})\right), \qquad \theta \neq \theta_0,$$

  and the proof relies on the Kullback-Leibler divergence.

- Let us focus on the **univariate** case $\Theta \subseteq \mathbb{R}$. It is instructive to provide a **direct proof** for exponential families, recalling that $\ell(\theta_0 \boldsymbol{Y}) = \theta S - nK(\theta)$.

- In the first place, note that

$$\mathbb{E}_{\theta_0}(\ell(\theta;\boldsymbol{Y})) = n\left[\theta\mu(\theta_0) - K(\theta)\right],$$

  implying that Wald inequality holds true if and only if

$$\mu(\theta_0)\left(\theta_0 - \theta\right) > K(\theta_0) - K(\theta), \qquad \theta \neq \theta_0.$$

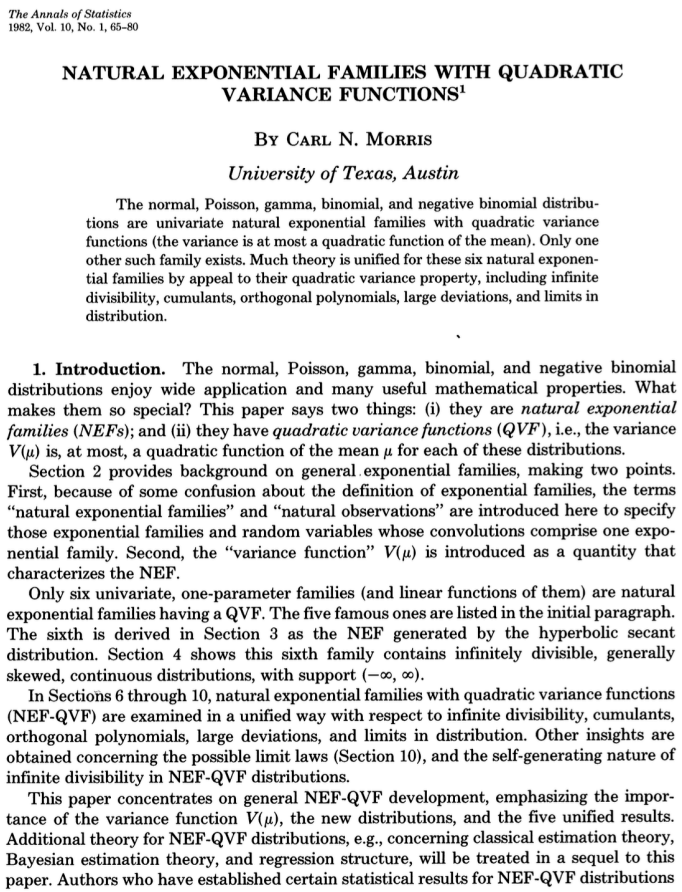
- This is indeed the case, the above being a characterization of **convexity** for $K(\cdot)$, which we previously show having $\partial^2/\partial\theta^2 K(\theta) > 0$ for all $\theta \in \tilde{\Theta}$. Moreover, recall that $\mu(\theta) = \partial/\partial\theta K(\theta)$.

References and study material

Main references

- Pace and Salvan (1997)
  - **Chapter 5** (*Exponential families*)
  - **Chapter 6** (*Exponential dispersion families*)
- Davison (2003)
  - **Chapter 5** (*Models*)
- Efron and Hastie (2016)
  - **Chapter 5** (*Parametric models and exponential families*)
- Efron (2023)
  - **Chapter 1** (*One-parameter exponential families*)
  - **Chapter 2** (*Multiparameter exponential families*)

Morris (1982)



The Annals of Statistics
1982, Vol. 10, No. 1, 65–80

NATURAL EXPONENTIAL FAMILIES WITH QUADRATIC VARIANCE FUNCTIONS[1]

BY CARL N. MORRIS

University of Texas, Austin

The normal, Poisson, gamma, binomial, and negative binomial distributions are univariate natural exponential families with quadratic variance functions (the variance is at most a quadratic function of the mean). Only one other such family exists. Much theory is unified for these six natural exponential families by appeal to their quadratic variance property, including infinite divisibility, cumulants, orthogonal polynomials, large deviations, and limits in distribution.

1. Introduction. The normal, Poisson, gamma, binomial, and negative binomial distributions enjoy wide application and many useful mathematical properties. What makes them so special? This paper says two things: (i) they are *natural exponential families* (*NEFs*); and (ii) they have *quadratic variance functions* (*QVF*), i.e., the variance $V(\mu)$ is, at most, a quadratic function of the mean μ for each of these distributions.

Section 2 provides background on general exponential families, making two points. First, because of some confusion about the definition of exponential families, the terms "natural exponential families" and "natural observations" are introduced here to specify those exponential families and random variables whose convolutions comprise one exponential family. Second, the "variance function" $V(\mu)$ is introduced as a quantity that characterizes the NEF.

Only six univariate, one-parameter families (and linear functions of them) are natural exponential families having a QVF. The five famous ones are listed in the initial paragraph. The sixth is derived in Section 3 as the NEF generated by the hyperbolic secant distribution. Section 4 shows this sixth family contains infinitely divisible, generally skewed, continuous distributions, with support $(-\infty, \infty)$.

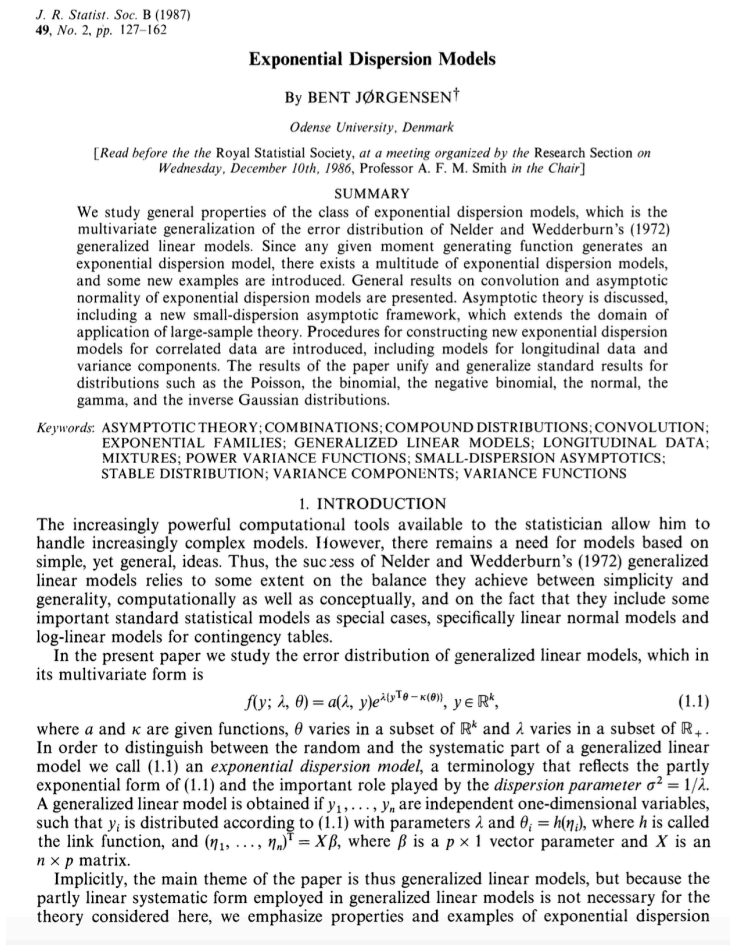
In Sections 6 through 10, natural exponential families with quadratic variance functions (NEF-QVF) are examined in a unified way with respect to infinite divisibility, cumulants, orthogonal polynomials, large deviations, and limits in distribution. Other insights are obtained concerning the possible limit laws (Section 10), and the self-generating nature of infinite divisibility in NEF-QVF distributions.

This paper concentrates on general NEF-QVF development, emphasizing the importance of the variance function $V(\mu)$, the new distributions, and the five unified results. Additional theory for NEF-QVF distributions, e.g., concerning classical estimation theory, Bayesian estimation theory, and regression structure, will be treated in a sequel to this paper. Authors who have established certain statistical results for NEF-QVF distributions

- Morris (1982, AoS) is a **seminal** paper in the field of **exponential families**.
- It is a must-read, as it encompasses and overviews many of the results discussed in this unit.
- It also shows that exponential families with quadratic variance are **infinitely divisible**, provided that $c \geq 0$.
- The paper covers several **advanced topics**, including:
  - orthogonal polynomials;
  - limiting results;
  - large deviations;
  - ...and more.

Home page

Jorgensen (1987)



J. R. Statist. Soc. B (1987)
49, *No.* 2, *pp.* 127–162

Exponential Dispersion Models

By BENT JØRGENSEN†

Odense University, Denmark

[*Read before the the* Royal Statistial Society, *at a meeting organized by the* Research Section *on Wednesday, December 10th, 1986*, Professor A. F. M. Smith *in the Chair*]

SUMMARY

We study general properties of the class of exponential dispersion models, which is the multivariate generalization of the error distribution of Nelder and Wedderburn's (1972) generalized linear models. Since any given moment generating function generates an exponential dispersion model, there exists a multitude of exponential dispersion models, and some new examples are introduced. General results on convolution and asymptotic normality of exponential dispersion models are presented. Asymptotic theory is discussed, including a new small-dispersion asymptotic framework, which extends the domain of application of large-sample theory. Procedures for constructing new exponential dispersion models for correlated data are introduced, including models for longitudinal data and variance components. The results of the paper unify and generalize standard results for distributions such as the Poisson, the binomial, the negative binomial, the normal, the gamma, and the inverse Gaussian distributions.

Keywords: ASYMPTOTIC THEORY; COMBINATIONS; COMPOUND DISTRIBUTIONS; CONVOLUTION; EXPONENTIAL FAMILIES; GENERALIZED LINEAR MODELS; LONGITUDINAL DATA; MIXTURES; POWER VARIANCE FUNCTIONS; SMALL-DISPERSION ASYMPTOTICS; STABLE DISTRIBUTION; VARIANCE COMPONENTS; VARIANCE FUNCTIONS

1. INTRODUCTION

The increasingly powerful computational tools available to the statistician allow him to handle increasingly complex models. However, there remains a need for models based on simple, yet general, ideas. Thus, the success of Nelder and Wedderburn's (1972) generalized linear models relies to some extent on the balance they achieve between simplicity and generality, computationally as well as conceptually, and on the fact that they include some important standard statistical models as special cases, specifically linear normal models and log-linear models for contingency tables.

In the present paper we study the error distribution of generalized linear models, which in its multivariate form is

$$f(y;\lambda,\theta)=a(\lambda,y)e^{\lambda\{y^{\mathrm{T}}\theta-\kappa(\theta)\}},\ y\in\mathbb{R}^k, \tag{1.1}$$

where a and κ are given functions, θ varies in a subset of $\mathbb{R}^k$ and λ varies in a subset of $\mathbb{R}_+$. In order to distinguish between the random and the systematic part of a generalized linear model we call (1.1) an *exponential dispersion model*, a terminology that reflects the partly exponential form of (1.1) and the important role played by the *dispersion parameter* $\sigma^2=1/\lambda$. A generalized linear model is obtained if $y_1,\ldots,y_n$ are independent one-dimensional variables, such that y_i is distributed according to (1.1) with parameters λ and $\theta_i=h(\eta_i)$, where h is called the link function, and $(\eta_1,\ldots,\eta_n)^{\mathrm{T}}=X\beta$, where β is a $p\times 1$ vector parameter and X is an $n\times p$ matrix.

Implicitly, the main theme of the paper is thus generalized linear models, but because the partly linear systematic form employed in generalized linear models is not necessary for the theory considered here, we emphasize properties and examples of exponential dispersion

- Jorgensen (1987, JRSSB) is another **seminal** paper in the field of **exponential dispersion families**.
- It studies a multivariate extension of exponential dispersion models of Nelder and Wedderburn (1972).
- It characterizes the entire class in terms of variance function, extending Morris (1982).
- It also describes a notion of asymptotic normality called **small sample asymptotics**.
- It is a **read** paper and among the discussants we find, J.A. Nelder, A.C. Davison, C.N. Morris.

Home page

Diaconis and Ylvisaker (1979)

The Annals of Statistics
1979, Vol. 7, No. 2, 269–281

CONJUGATE PRIORS FOR EXPONENTIAL FAMILIES

BY PERSI DIACONIS[1] AND DONALD YLVISAKER[2]
Stanford University and The University of California, Los Angeles

Let X be a random vector distributed according to an exponential family with natural parameter $\theta \in \Theta$. We characterize conjugate prior measures on Θ through the property of linear posterior expectation of the mean parameter of X : $E\{E(X|\theta)|X = x\} = ax + b$. We also delineate which hyperparameters permit such conjugate priors to be proper.

1. Introduction. Modern Bayesian statistics is dominated by the notion of conjugate priors. The usual definition is that a family of priors is conjugate if it is closed under sampling (Lindley [1972], pages 22–23 or Raiffa and Schlaifer [1961], pages 43–57). Consider the following example: let S_n be the number of heads in n independent tosses of a coin with unknown parameter p. The accepted family of conjugate priors for p is the beta family with densities

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \qquad \alpha > 0, \beta > 0.$$

Let h be any positive bounded measurable function on the unit interval and observe that a prior density proportional to $h(p)f(p; \alpha, \beta)$ leads to a posterior density of p, given $S_n = x$, proportional to $h(p)f(p; \alpha + x, \beta + h - x)$. Thus, the family $\{h(\cdot)f(\cdot; \alpha, \beta)|\alpha > 0, \beta > 0, h$ positive, bounded, measurable$\}$ with each member normalized to be a prior density, is closed under sampling. Now beta priors have the additional property that the posterior expectation of the mean parameter p is a linear function of S_n. That is, there are numbers a_n, b_n such that

$$E[p|S_n = k] = \frac{\int_0^1 p^{k+1}(1-p)^{n-k} f(p; \alpha, \beta)\,dp}{\int_0^1 p^{k}(1-p)^{n-k} f(p; \alpha, \beta)\,dp} = a_n k + b_n$$

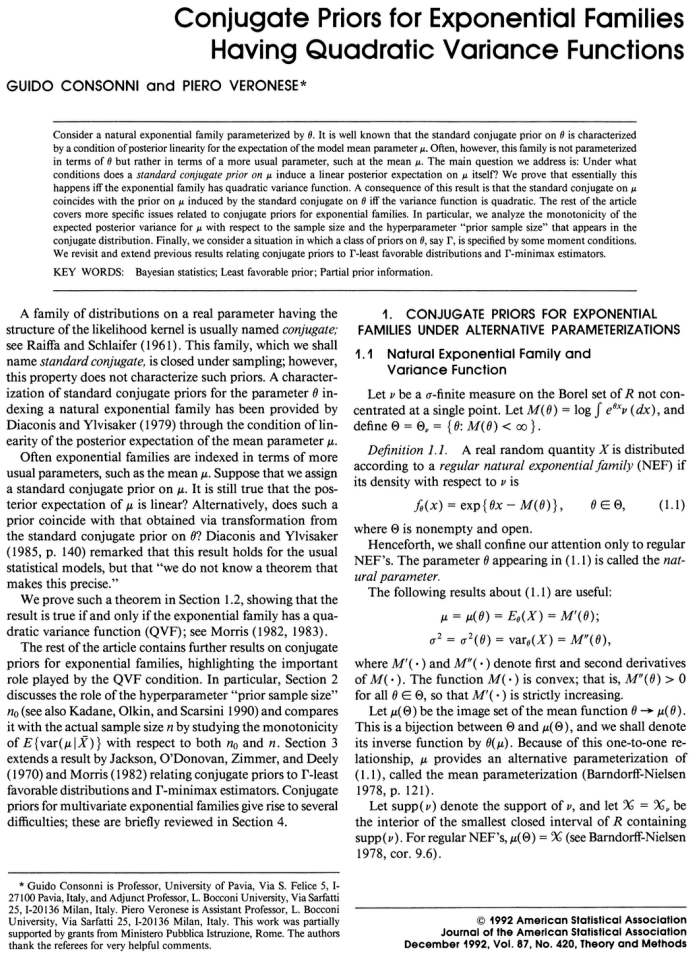
holds for $k = 0, 1, 2, \ldots, n$. A principal result of this paper is that, subject to regularity conditions, the conjugate priors typically used satisfy, and are characterized by, a similar relation of posterior linearity:

$$E\{E(X|\theta)|X = x\} = ax + b. \tag{1.1}$$

The regularity conditions assumed below allow such standard examples as the normal prior for normal location, the gamma prior for the Poisson, the inverse Wishart prior for normal covariance, and the beta prior for the negative binomial.

- Bayesian statistics also greatly benefits from the use of exponential families.
- Diaconis and Ylvisaker (1979, AoS) is a **seminal paper** on the topic of **conjugate priors**.
- Broadly speaking, conjugate priors always exist for exponential families.
- These are known as the Diaconis–Ylvisaker conjugate priors.
- Classical priors such as beta–Bernoulli and Poisson–gamma are special cases.
- The posterior expectation under the mean parametrization is a **linear combination** of the data and the prior mean.

Consonni and Veronese (1992)



Conjugate Priors for Exponential Families Having Quadratic Variance Functions

GUIDO CONSONNI and PIERO VERONESE*

Consider a natural exponential family parameterized by θ. It is well known that the standard conjugate prior on θ is characterized by a condition of posterior linearity for the expectation of the model mean parameter μ. Often, however, this family is not parameterized in terms of θ but rather in terms of a more usual parameter, such at the mean μ. The main question we address is: Under what conditions does a *standard conjugate prior on* μ induce a linear posterior expectation on μ itself? We prove that essentially this happens iff the exponential family has quadratic variance function. A consequence of this result is that the standard conjugate on μ coincides with the prior on μ induced by the standard conjugate on θ iff the variance function is quadratic. The rest of the article covers more specific issues related to conjugate priors for exponential families. In particular, we analyze the monotonicity of the expected posterior variance for μ with respect to the sample size and the hyperparameter "prior sample size" that appears in the conjugate distribution. Finally, we consider a situation in which a class of priors on θ, say Γ, is specified by some moment conditions. We revisit and extend previous results relating conjugate priors to Γ-least favorable distributions and Γ-minimax estimators.

KEY WORDS: Bayesian statistics; Least favorable prior; Partial prior information.

A family of distributions on a real parameter having the structure of the likelihood kernel is usually named *conjugate;* see Raiffa and Schlaifer (1961). This family, which we shall name *standard conjugate,* is closed under sampling; however, this property does not characterize such priors. A characterization of standard conjugate priors for the parameter θ indexing a natural exponential family has been provided by Diaconis and Ylvisaker (1979) through the condition of linearity of the posterior expectation of the mean parameter μ.

Often exponential families are indexed in terms of more usual parameters, such as the mean μ. Suppose that we assign a standard conjugate prior on μ. It is still true that the posterior expectation of μ is linear? Alternatively, does such a prior coincide with that obtained via transformation from the standard conjugate prior on θ? Diaconis and Ylvisaker (1985, p. 140) remarked that this result holds for the usual statistical models, but that "we do not know a theorem that makes this precise."

We prove such a theorem in Section 1.2, showing that the result is true if and only if the exponential family has a quadratic variance function (QVF); see Morris (1982, 1983).

The rest of the article contains further results on conjugate priors for exponential families, highlighting the important role played by the QVF condition. In particular, Section 2 discusses the role of the hyperparameter "prior sample size" n_0 (see also Kadane, Olkin, and Scarsini 1990) and compares it with the actual sample size n by studying the monotonicity of $E\{\text{var}(\mu|\bar{X})\}$ with respect to both n_0 and n. Section 3 extends a result by Jackson, O'Donovan, Zimmer, and Deely (1970) and Morris (1982) relating conjugate priors to Γ-least favorable distributions and Γ-minimax estimators. Conjugate priors for multivariate exponential families give rise to several difficulties; these are briefly reviewed in Section 4.

1. CONJUGATE PRIORS FOR EXPONENTIAL FAMILIES UNDER ALTERNATIVE PARAMETERIZATIONS

1.1 Natural Exponential Family and Variance Function

Let ν be a σ-finite measure on the Borel set of R not concentrated at a single point. Let $M(\theta) = \log \int e^{\theta x} \nu(dx)$, and define $\Theta = \Theta_\nu = \{\theta: M(\theta) < \infty\}$.

Definition 1.1. A real random quantity X is distributed according to a *regular natural exponential family* (NEF) if its density with respect to ν is

$$f_\theta(x) = \exp\{\theta x - M(\theta)\}, \qquad \theta \in \Theta, \qquad (1.1)$$

where Θ is nonempty and open.

Henceforth, we shall confine our attention only to regular NEF's. The parameter θ appearing in (1.1) is called the *natural parameter.*

The following results about (1.1) are useful:

$$\mu = \mu(\theta) = E_\theta(X) = M'(\theta);$$

$$\sigma^2 = \sigma^2(\theta) = \text{var}_\theta(X) = M''(\theta),$$

where $M'(\cdot)$ and $M''(\cdot)$ denote first and second derivatives of $M(\cdot)$. The function $M(\cdot)$ is convex; that is, $M''(\theta) > 0$ for all $\theta \in \Theta$, so that $M'(\cdot)$ is strictly increasing.

Let $\mu(\Theta)$ be the image set of the mean function $\theta \to \mu(\theta)$. This is a bijection between Θ and $\mu(\Theta)$, and we shall denote its inverse function by $\theta(\mu)$. Because of this one-to-one relationship, μ provides an alternative parameterization of (1.1), called the mean parameterization (Barndorff-Nielsen 1978, p. 121).

Let $\text{supp}(\nu)$ denote the support of ν, and let $\mathcal{X} = \mathcal{X}_\nu$ be the interior of the smallest closed interval of R containing $\text{supp}(\nu)$. For regular NEF's, $\mu(\Theta) = \mathcal{X}$ (see Barndorff-Nielsen 1978, cor. 9.6).

* Guido Consonni is Professor, University of Pavia, Via S. Felice 5, I-27100 Pavia, Italy, and Adjunct Professor, L. Bocconi University, Via Sarfatti 25, I-20136 Milan, Italy. Piero Veronese is Assistant Professor, L. Bocconi University, Via Sarfatti 25, I-20136 Milan, Italy. This work was partially supported by grants from Ministero Pubblica Istruzione, Rome. The authors thank the referees for very helpful comments.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Theory and Methods

- Consonni and Veronese (1992, JASA) is another **Bayesian** contribution which refines the results of Diaconis and Ylvisaker (1979).
- It investigates when a **conjugate prior** specified on the mean parameter μ of a natural exponential family leads to a linear posterior expectation of μ.
- The main result shows that this **posterior linearity** holds if and only if the **variance function** is **quadratic**.
- The paper also explores the **monotonicity** of the **posterior variance** of μ with respect to both the sample size and the prior sample size.

Home page

References

Billingsley, Patrick. 1995. *Probability And Measure*. Wiley.

Consonni, Guido, and Piero Veronese. 1992. "Conjugate Priors for Exponential Families Having Quadratic Functions." *Journal of the American Statistical Association* 87 (420): 1123–27.

Davison, A. C. 2003. *Statistical Models*. Cambridge University Press.

Diaconis, Persi, and Donald Ylvisaker. 1979. "Conjugate prior for exponential families." *The Annals of Statistics* 7 (2): 269–92.

Efron, Bradley. 2023. *Exponential Families in Theory and Practice*. Cambridge University Press.

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference*. Cambridge University Press.

Fisher, R. A. 1934. "Two new properties of mathematical likelihood." *Proceedings of the Royal Society of London. Series A* 144 (852): 285–307.

Jorgensen, Bert. 1987. "Exponential dispersion model." *Journal of the Royal Statistical Society. Series B: Methodological* 49 (2): 127–62.

Lehmann, E. L., and G. Casella. 1998. *Theory of Point Estimation, Second Edition*. Springer.

Morris, Carl N. 1982. "Natural Exponential Families with Quadratic Variance Functions." *Annals of Statistics* 10 (1): 65–80.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized linear models." *Journal of the Royal Statistical Society. Series A: Statistics in Society* 135 (3): 370–84.

Pace, Luigi, and Alessandra Salvan. 1997. *Principles of statistical inference from a Neo-Fisherian*