# Exponential Families & From Prior Information to Prior Distribution

Timo Koski

20.01.2010

KTH Matematik

# These notes

The material in these notes was intially based on

- C.P. Roberts: *Bayesian Choice, Second Edition*, Springer-Verlag, Berlin , 2001.

Some auxuliary results required are quoted from

- M.J. Schervish: *Theory of Statistics* , Springer-Verlag, Berlin , 1995.

An idiosyncracy of Roberts is that $u \cdot x$ may designate both the product of real numbers $u$ and $x$ as well as the scalar product of vectors $u$ and $x$. Otherwise an effort has been made to unify the notation with the notes by Henrik. In addition, Roberts prefers to write $x$ for both outcome and random variable.

# Notation

An idiosyncracy of Roberts is that $u \cdot x$ may designate both the product of real numbers $u$ and $x$ as well as the scalar product of vectors $u$ and $x$. In addition, Roberts prefers to write $x$ for both outcome and random variable. Roberts deals with the *natural exponential family* to be introduced below. Otherwise an effort has been made to accomodate to the notation with the notes by Henrik.

# Parametric statistical model recalled

$x$ is an observation of a random variable $X$, $x \in \mathcal{X}$ (=sample space).

$$x \sim f_{X|\Theta}(x|\theta)$$

$f_{X|\Theta}(x|\theta)$ is a probability density w.r.t. to a $\sigma$ finite measure $\nu$ on $\mathcal{X}$.
$f_{X|\Theta}(x|\theta)$ is a known function of $x$ and $\theta$.
$\theta$ is an unknown parameter $\in \Omega \subseteq$ a vector space of <u>finite dimension</u>.

KTH Matematik

# Exponential Families

The family of distributions $\mu_{X|\Theta}$ with densities w.r.t. to a $\sigma$-finite measure $\nu$ on $\mathcal{X}$ defined by

$$\frac{d\mu_{X|\Theta=\theta}}{d\nu}(x \mid \theta) = f_{X|\Theta}(x|\theta) = C(\theta)h(x)e^{R(\theta)\cdot T(x)}$$

is called an exponential family (of dimension $k$), where

- $C(\theta)$ and $h(x)$ are measurable functions from $\Omega$ and $\mathcal{X}$ to $R_+$,
- $R(\theta)$ and $T(x)$ are measurable functions from $\Omega$ and $\mathcal{X}$ to $R^k$,
- $R(\theta) \cdot T(x)$ is a scalar product in $R^k$, i.e.,

$$R(\theta) \cdot T(x) = \sum_{i=1}^{k} R_i(\theta) \cdot T_i(x)$$

# Exponential Families

- The family of distributions $\mu_{X|\Theta}$ has densities w.r.t. to a $\sigma$-finite measure $\nu$ on $\mathcal{X}$, if for all $\theta \in \Omega$, $\mu_{X|\Theta=\theta} << \nu$. If there is another $\sigma$-finite measure, say $\nu\prime$ such that for all $\theta$, $\mu_{X|\Theta=\theta} << \nu\prime$, then there exists a representation as above. The dimension $k$ may depend on the dominating measure.
- If $\theta_0 \in\Omega$, then $\mu_{X|\Theta=\theta} << \mu_{X|\Theta=\theta_0}$ and the density of $\mu_{X|\Theta=\theta}$ w.r.t. $\mu_{X|\Theta=\theta_0}$ on $\mathcal{X}$ is

$$\frac{d\mu_{X|\Theta=\theta}}{d\mu_{X|\Theta=\theta_0}}(x \mid \theta) = \frac{C(\theta)}{C(\theta_0)} e^{(R(\theta)-R(\theta_0))\cdot T(x)}$$

Hence, e.g., the family $U(0,\theta), \theta \in \Omega = (0,\infty)$ cannot be an exponential family.

KTH Matematik

# EXAMPLES OF EXPONENTIAL FAMILIES: $Be(\theta)$

$\Omega = (0,1)$, $\nu$ is the counting measure[1].

$$f_{X|\Theta}(x|\theta) = \theta^x \cdot (1-\theta)^{1-x}. \quad x = 0,1$$

We write

$$f_{X|\Theta}(x|\theta) = C(\theta) e^{R(\theta)\cdot x},$$

where

$$C(\theta) = e^{\log 1-\theta}, T(x) = x, R(\theta) = \log \frac{\theta}{1-\theta}, h(x) = 1.$$

[1] $\mathcal{X}=$ positive integers, $\nu(A) =$ the number of elements in $A \subseteq \mathcal{X}$

# EXAMPLES OF EXPONENTIAL FAMILIES: $N(\mu, \sigma^2)$

$x^{(n)} = (x_1, x_2, \ldots, x_n)$, $x_i$ I.I.D. $\sim N(\mu, \sigma^2)$.

$$f\left(x^{(n)} | \mu, \sigma^2\right) = \frac{1}{\sigma^n \left(2\pi\right)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2}$$

$$= \frac{1}{\left(2\pi\right)^{n/2}} \sigma^{-n} e^{-\frac{n\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} n\overline{x}}.$$

KTH Matematik

# EXAMPLES OF EXPONENTIAL FAMILIES: $N(\mu, \sigma^2)$

$\Omega = R \times (0, \infty)$, $\nu$ = Lebesgue measure on $R^n$. ,

$$f(x^{(n)} | \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \sigma^{-n} e^{-\frac{n\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} n\overline{x}}.$$

$$C(\theta) = \sigma^{-n} e^{-\frac{n\mu^2}{2\sigma^2}}, h(x) = \frac{1}{(2\pi)^{n/2}}.$$

$$R(\theta) \cdot T\left(x^{(n)}\right) = R_1(\theta) T_1\left(x^{(n)}\right) + R_2(\theta) T_2\left(x^{(n)}\right)$$

$$T_1\left(x^{(n)}\right) = \sum_{i=1}^{n} x_i^2, T_2\left(x^{(n)}\right) = n\overline{x}$$

$$R_1(\theta) = -\frac{1}{2\sigma^2}, R_2(\theta) = \frac{\mu}{\sigma^2}$$

# SOME COMMON EXPONENTIAL FAMILIES

Poisson $\mathrm{Po}(\theta)$, Gamma $\mathrm{Ga}(p, \theta)$, Binomial $\mathrm{Bin}(n, \theta)$, Negative Binomial $\mathrm{Neg}(m, \theta)$, Multinomial, Inverse Gaussian, Weibull (with known shape parameter)

# Inverse Gaussian

$\mu > 0$, $\lambda > 0$, $0 < x < \infty$, $\theta = (\mu, \lambda)$

$$f(x \mid \mu, \lambda) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}$$

The inverse Gaussian distribution is a two-parameter exponential family with natural parameters $-\lambda/(2\mu^2)$ and $-\lambda/2$, and $T_1(X) = X$ and $T_2(X) = 1/X$.

# Natural Parameter, Natural parameter space

Clearly the density

$$f_{X|\Theta}(x \mid \theta) = C(\theta)h(x)e^{R(\theta)\cdot T(x)}$$

depends only on

$$R = (R_1(\theta), R_2(\theta), \ldots, R_k(\theta)) .$$

We call $R$ the **natural parameter**.

$$\mathcal{N} = \mathcal{N}(\nu) := \left\{ R \in R^k | \int_{\mathcal{X}} h(x)e^{R\cdot T(x)}\nu(dx) < \infty \right\}$$

$\mathcal{N}$ is called the natural parameter space, we assume that $\mathcal{N} = \Omega$.

# Density of $T$

From Schervish p. 103:

**Lemma**

*If $X$ has an exponential family distribution, then $T(X)$ has an exponential family distribution, and there exists a measure $\nu_T$ such that*

$$\frac{d\mu_{T|\Theta}}{d\nu_T}(t) = C(\theta)e^{t\cdot\theta}$$

This will be discussed further in the lecture on sufficient statistics. A proof is found on p. 17 in notes by Henrik.

# Natural Exponential Families (1)

Thus, $\Theta \subseteq R^k$ and $\mathcal{X} \subseteq R^k$, we can make a change of variable & relabel: $R(\theta) \leftrightarrow \theta$ and $T(x) \leftrightarrow x$.

$$f_{X|\Theta}(x \mid \theta) = C(\theta)h(x)e^{\theta \cdot x}$$

and the family is said to be a *natural exponential family* . Here $\theta \cdot x$ is inner product on $R^k$.

# Exponential Families: the natural parameter space

Thus

$$1 = C(\theta) \int_{\mathcal{X}} h(x) e^{\theta \cdot x} \nu(dx).$$

$$C(\theta) = \frac{1}{\int_{\mathcal{X}} h(x) e^{\theta \cdot x} \nu(dx)}.$$

$$\mathcal{N}(\nu) := \left\{ \theta \in R^k | \int_{\mathcal{X}} h(x) e^{\theta \cdot x} \nu(dx) < \infty \right\}$$

$\mathcal{N} = \mathcal{N}(\nu)$ is called the natural parameter space, possibly $\neq \Omega$.

# The natural parameter space

$$\mathcal{N} := \left\{ \theta | \int_{\mathcal{X}} h(x) e^{\theta \cdot x} \nu(dx) < \infty \right\}$$

- An application of convexity of $\exp(\cdot)$ yields that $\mathcal{N}$ is convex (as shown below).
- WE ASSUME that $\mathcal{N}$ *is an open set in* $R^k$. Then we are dealing with a **regular exponential family**[2]

[2] O. Barndorff-Nielsen: *Information and Exponential families in Statistical Theory*, J. Wiley, 1978

# Natural Exponential Families (2)

### Sats

$\frac{1}{C(\theta)}$ *is a convex function.*

**Proof:** $\theta_1$ and $\theta_2$ are two points in $\mathcal{N}$ and $0 \leq \lambda \leq 1$. Then, since the exponential function is convex,

$$\frac{1}{C(\lambda\theta_1 + (1-\lambda)\theta_2)} = \int_{\mathcal{X}} h(x) e^{(\lambda\theta_1 + (1-\lambda)\theta_2)\cdot x} \nu(dx)$$

$$\leq \int_{\mathcal{X}} h(x) \left( \lambda e^{\theta_1 \cdot x} + (1-\lambda) e^{\theta_2 \cdot x} \right) \nu(dx) = \lambda \frac{1}{C(\theta_1)} + (1-\lambda) \frac{1}{C(\theta_2)}.$$

# Natural Exponential Families (3)

### Följdsats

*$\mathcal{N}$ is a convex set.*

**Proof:** $\theta_1$ and $\theta_2$ are two points in $\mathcal{N}$ and $0 \leq \lambda \leq 1$. Then $\frac{1}{C(\theta_1)} < \infty$ and $\frac{1}{C(\theta_2)} < \infty$, and since $\frac{1}{C(\theta)}$ is convex, we get that $\lambda\theta_1 + (1-\lambda)\theta_2 \in \mathcal{N}$. □

# Natural Exponential Families (4)

$$f_{X|\Theta}(x \mid \theta) = h(x)e^{\theta \cdot x - \psi(\theta)}$$

where

$$\psi(\theta) = -\log C(\theta).$$

The function $\psi(\theta)$ is called the **cumulant function**.

# Natural Exponential Families (5)

## Proposition

*The moment generating function of a natural exponential family is*

$$M(u) = E_\theta\left[e^{uX}\right] = \frac{C(\theta)}{C(\theta + u)}$$

**Proof:**

$$E_\theta\left[e^{uX}\right] = \int_{\mathcal{X}} e^{ux} f_{X|\Theta}\left(x \mid \theta\right) \nu(dx) = \int_{\mathcal{X}} e^{ux} h(x) e^{\theta \cdot x - \psi(\theta)} \nu(dx)$$

$$= e^{-\psi(\theta)} \int_{\mathcal{X}} h(x) e^{(u+\theta)\cdot x} \nu(dx) = \frac{C(\theta)}{C(\theta + u)}.$$

# Natural Exponential Families: Poisson Distribution

$$f\left(x \mid \lambda\right) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, 2, \ldots,$$

- $$f\left(x \mid \lambda\right) = \frac{1}{x!}e^{\theta x - e^{\theta}}$$
  $$\psi\left(\theta\right) = e^{\theta}, \theta = \log\lambda, h(x) = \frac{1}{x!}$$
- Moment generating function: $M(u) = \frac{C(\theta)}{C(\theta+u)}$,

$$C\left(\theta\right) = e^{-\psi(\theta)} = e^{-e^{\theta}} = e^{-\lambda}, C\left(\theta + u\right) = e^{-\psi(\theta+u)} = e^{-\lambda e^u}$$

I.e., $M(u) = e^{-\lambda}e^{\lambda e^u} = e^{\lambda(e^u - 1)}$

# Mean in a Natural Exponential Family

If $E_\theta [X]$ denotes the mean (vector) of $X \sim f_{X|\Theta}(x|\theta)$ in a natural family, then[3]

$$E_\theta [X] = \int_{\mathcal{X}} x f(x \mid \theta)\, dx = \nabla_\theta \psi(\theta) .$$

where $\theta \in \mathrm{int}(\mathcal{N})$ and $\mathcal{X} \subseteq R^k$.

**Proof**:

$$\int_{\mathcal{X}} x f(x \mid \theta)\, dx = e^{-\psi(\theta)} \int_{\mathcal{X}} h(x) x e^{\theta \cdot x} dx.$$

[3] $\nabla_\theta \psi(\theta) = \left( \frac{\partial}{\partial \theta_1} \psi(\theta), \frac{\partial}{\partial \theta_2} \psi(\theta), \ldots, \frac{\partial}{\partial \theta_k} \psi(\theta) \right)^T$

# Mean in a Natural Exponential Family

$$e^{-\psi(\theta)} \int_{\mathcal{X}} h(x) x e^{\theta \cdot x} dx = e^{-\psi(\theta)} \int_{\mathcal{X}} h(x) \nabla_\theta e^{\theta \cdot x} dx$$

$$= {}^{4} e^{-\psi(\theta)} \nabla_\theta \int_{\mathcal{X}} h(x) e^{\theta \cdot x} dx = e^{-\psi(\theta)} \nabla_\theta \frac{1}{C(\theta)} =$$

$$= e^{-\psi(\theta)} \frac{(-\nabla_\theta C(\theta))}{C(\theta)^2}$$

$$= C(\theta) \frac{(-\nabla_\theta C(\theta))}{C(\theta)^2} = \frac{(-\nabla_\theta C(\theta))}{C(\theta)}$$

$$= \nabla_\theta (-\log C(\theta)) = \nabla_\theta \psi(\theta) .$$

□

[4] It is permissible to interchange integration and derivation, Schervish Thm 2.64. p. 105

# Mean in a Natural Exponential Family : Poisson Distribution

$$f\left(x \mid \lambda\right) = \frac{1}{x!} e^{\theta x - e^{\theta}}$$

$$\psi\left(\theta\right) = e^{\theta}$$

$$E_{\theta}\left[X\right] = \frac{d}{d\theta}\psi\left(\theta\right) = e^{\theta} = \lambda.$$

KTH Matematik

# Uncertainty

Uncertainty about the unknown $\theta$ is modeled by a probability distribution $\pi(\theta)$, and $\pi_{\Theta|X}(\theta|x)$ expresses the uncertainty about the unknown $\theta$ after the observation of $x$.

We use probability as tool for all parts of our analysis. This is coherence.

Mathematically: the unknown $\theta$ becomes an outcome of a random variable, i.e., $(X, \Theta)$ will have a joint distribution. For the precise formulation of this see the notes by Henrik.

# Bayesian Parametric Statistical Model

A Bayesian parametric statistical model consists of

- a parametric model

$$x \sim f_{X|\Theta}(x|\theta)$$

- a prior density (an improper density can be used)

$$\theta \sim \pi(\theta)$$

The quantity of interest: posterior distribution

$$\theta|x \sim \pi_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) \cdot \pi(\theta)$$

# Bayes' rule: parametric model

$$\pi_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x \mid \theta) \cdot \pi(\theta)}{\int_{\Theta} f_{X|\Theta}(x \mid \theta) \cdot \pi(\theta)\, d\theta},$$

Terminology for Bayes' Rule:

- $\pi(\theta)$ : **prior density** on $\Omega$; here w.r.t. the Lebesgue measure.
- $\pi_{\Theta|X}(\theta|x)$ : **posterior density** on $\Omega$, here w.r.t. the Lebesgue measure.
- $m(x) = \int_{\Theta} f_{X|\Theta}(x \mid \theta) \cdot \pi(\theta)\, d\theta$ : **marginal distribution** of $x$, also known as the prior predictive distribution of $x$.

# Q: How do we choose $\pi(\theta)$ ?

- Assessment (by Questionnaries)
- Conjugate prior
- Non-informative prior
  - Laplace's prior
  - Jeffreys' prior
- Maximum entropy prior

# Assessment of prior knowledge

(One form of) Bayesian statistics relies upon a **personalistic theory of probability** for quantification of prior knowledge. In such a theory

- probability measures the confidence that a particular individual (assessor) has in the truth of a particular proposition
- no attempt is made to specify which assessments are correct
- personal probabilities should satisfy certain postulates of coherence.

R.L.Winkler in

- Robert L. Winkler: The Assessment of Prior Distributions in Bayesian Analysis
  Journal of the American Statistical Association, Vol. 62, No. 319. (Sep., 1967), pp. 776-800.)

devises questionnaires (or interviews) to elicit information to write down a prior distribution. Students of Univ. of Chigago were asked to, e.g., assess the uncertainty about the probability of a randomly chosen student of Univ. of Chigago being Roman Catholic using a probability distribution. The assessment was done by four different methods, like giving fractiles, making bets, assessing impact of additional data, drawing graphs. One interesting finding is that the assessments by the same person using different methods may be conflicting.

# Diffuse/Non-diffuse prior distributions by assessment

The priors in Winkler's study are not diffuse: the students of Univ. of Chigago have, since they have been around, an idea about the number of Roman Catholics at the campus of of Univ. of Chigago.

# Choice of prior distributions by assessment: Eliciting probabilities

More on this:

- R.L. Keeney & D. von Winterfeldt: Eliciting Probabilities from Experts in Complex Technical Problems. *IEEE Transactions on Engineering Management*, Vol. 38, 1991, pp.191−201.
- K.M. Chaloner & G.T. Duncan: Assessment of a Beta Distribution: PM Elicitation. *The Statistician*, 32, 1983, pp. 174−180

One more point ⇒

# Assessing Priors: Conjugate Prior

The interviews of Winkler were mathematically speaking all concerned with assessing the prior of $\theta$ in a Bernoulli $\mathrm{Be}(\theta) -$ I.I.D. process. Winkler claims a sensitivity analysis (loc.cit p. 791) showing that the prior distributions assessed by the interviews yielded posterior distributions that were ‘only little’ different (by a test of goodness-of-fit) from those obtained from Beta densities on $\theta$. Beta densities are conjugate priors.

An intuitive way of understanding conjugate priors is that with conjugate priors the prior knowledge can be translated into equivalent sample information. A formal definition of conjugate priors follows.

# Conjugate Prior

## Definition

*Let $\mathcal{F}$ be a class of probability densities $f_{X|\Theta}(x \mid \theta)$. A family of probability distributions $\Pi$ on $\Theta$ is said to be* **conjugate** *or* **closed under sampling** *for $\mathcal{F}$, if for every prior $\pi \in \Pi$, the posterior distribution $\pi_{\Theta|X}(\theta|x)$ also belongs to $\Pi$ for every $f \in \mathcal{F}$.*

□

# Conjugate Family of Priors

A conjugate family is usually associated with a particular sampling distribution that is even characteristic of conjugate priors: exponential families.

# Conjugate Priors for Exponential Families: An Intuitive Example

$x^{(n)} = (x_1, x_2, \ldots, x_n)$. $x_i \sim \mathrm{Po}(\lambda)$, I.I.D.,

$$f\left(x^{(n)} \mid \lambda\right) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

The likelihood is

$$L\left(\lambda; x^{(n)}\right) \propto e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}$$

This suggests the conjugate density as the density of the Gamma distribution, which is of the form

$$\pi\left(\lambda\right) \propto e^{-\beta\lambda} \lambda^{\alpha-1}$$

and hence

$$\pi\left(\lambda | x^{(n)}\right) \propto e^{-\lambda(\beta+n)} \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1}$$

# Conjugate Family of Priors for Exponential Families

## Proposition

*For the natural exponential family*

$$f_{X|\Theta}(x \mid \theta) = h(x)e^{\theta \cdot x - \psi(\theta)}$$

*the conjugate family*[a] *is given by*

$$\pi(\theta) = \psi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta \cdot \mu - \lambda\psi(\theta)}$$

*and the posterior is*

$$\psi(\theta|\mu + x, \lambda + 1).$$

[a] (if this is a probabilty density, c.f. below)

# Conjugate Priors for Exponential Families: Proof

*Proof:* By Bayes' rule

$$\pi\left(\theta|x\right) = \frac{f\left(x \mid \theta\right)\pi\left(\theta\right)}{m(x)}$$

We have

$$\begin{aligned} f\left(x \mid \theta\right)\pi\left(\theta\right) &= h(x)e^{\theta\cdot x-\psi(\theta)}\psi\left(\theta|\mu,\lambda\right) \\ &= h(x)K\left(\mu,\lambda\right)e^{\theta\cdot(x+\mu)-(1+\lambda)\psi(\theta)} \end{aligned}$$

# Conjugate Priors for Exponential Families: Proof

$$m(x) = \int_{\Theta} f\left(x \mid \theta\right) \pi\left(\theta\right) d\theta =$$

$$= h(x) K\left(\mu, \lambda\right) \int_{\Theta} e^{\theta \cdot (x+\mu) - (1+\lambda)\psi(\theta)} d\theta$$

$$= h(x) K\left(\mu, \lambda\right) K\left(x + \mu, \lambda + 1\right)^{-1}.$$

# Conjugate Priors for Exponential Families: Proof

$$\pi(\theta|x) = \frac{h(x)K(\mu,\lambda)\,e^{\theta\cdot(x+\mu)-(1+\lambda)\psi(\theta)}}{h(x)K(\mu,\lambda)\,K(x+\mu,\lambda+1)^{-1}}$$

$$= K(x+\mu,\lambda+1)\,e^{\theta\cdot(x+\mu)-(1+\lambda)\psi(\theta)},$$

which shows that the posterior belongs to the same family as the prior and that

$$\pi(\theta|x) = \psi(\theta|\mu+x,\lambda+1)$$

as claimed. □

# Conjugate Priors for Exponential Families

If $\lambda > 0$ and $\frac{\mu}{\lambda} \in \mathrm{Int}(\mathcal{N})$, then

$$\pi(\theta) = \psi(\theta|\mu, \lambda) = K(\mu, \lambda)\, e^{\theta \cdot \mu - \lambda \psi(\theta)}$$

is a probability density on $\Theta$ (proof is an exercise for the reader), which is presupposed in the proof above.
The parameters of the prior, $\lambda$ and $\mu$, are called **hyperparameters**.

# Mean for Exponential Families

We have the following properties:

- if $\pi(\theta) = K(x_o, \lambda)\, e^{\theta \cdot x_o - \lambda \psi(\theta)}$ then

$$\xi(\theta) = \int_{\Theta} E_{\theta}[x]\, \pi(\theta)\, d\theta = \frac{x_o}{\lambda}$$

This has been proved by Diaconis and Ylvisaker[5]. The proof is not summarized here.

[5] P. Diaconis & D. Ylvisaker: Conjugate Priors for Expoenntial Families. *The Annals of Statistics*, vol. 7, 1979, pp. 269−281.

# Posterior Means with Conjugate Priors for Exponential Families

- if $\pi\left(\theta\right) = K\left(\mu, \lambda\right) e^{\theta \cdot \mu - \lambda \psi(\theta)}$ then

$$\int_{\Theta} E_{\theta}\left[x\right] \pi\left(\theta | x^{(n)}\right) d\theta = \frac{\mu + n\overline{x}}{\lambda + n}$$

This follows from the preceding, as shown by Diaconis and Ylvisaker (1979). In fact Diaconis and Ylvisaker prove that this is a characterization of conjugate priors for regular exponential families.

# Mean of a Predictive Distribution

$$\int_{\Theta} E_{\theta}\left[x\right]\pi\left(\theta|x^{(n)}\right)d\theta = \int_{\Theta}\int_{\mathcal{X}} xf\left(x|\theta\right)\nu(dx)\pi\left(\theta|x^{(n)}\right)d\theta$$

(by Fubini's theorem)

$$= \int_{\mathcal{X}} x\int_{\Theta} f\left(x|\theta\right)\pi\left(\theta|x^{(n)}\right)d\theta\nu(dx)$$

(by definition in lecture 1)

$$= \int_{\mathcal{X}} xg(x|x^{(n)})\nu(dx)$$

the mean of the posterior predictive distribution (see the notes by Henrik pp. 10−11).

# Mean of a Predictive Distribution

Hence if conjugate priors for exponential families are used, then

$$\int_{\mathcal{X}} xg(x|x^{(n)})\nu(dx) = \frac{\mu + n\overline{x}}{\lambda + n}$$

is the mean of the corresponding predictive distribution. This suggests $\mu$ and $\lambda$ as 'virtual observations'.

# Laplace's Prior

P.S. Laplace[6] formulated the principle of insufficient reason to choose a prior as a uniform prior. There are drawbacks in this. Consider Laplace's prior for $\theta \in [0,1]$

$$\pi(\theta) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

Then consider

$$\phi = \theta^2.$$

[6] http://www-groups.dcs.st-and.ac.uk/∼history/Mathematicians/Laplace.html

# Laplace's Prior

We find the density of $\phi = \theta^2$. Take $0 < v < 1$.

$$F_\phi(v) = P\left(\phi \leq v\right) = P\left(\theta \leq \sqrt{v}\right) = \int_0^{\sqrt{v}} \pi\left(\theta\right) d\theta$$

$$= \sqrt{v}.$$

$$f_\phi(v) = \frac{d}{dv} F_\phi(v) = \frac{d}{dv}\sqrt{v} = \frac{1}{2}\frac{1}{\sqrt{v}}$$

which is no longer uniform. But how come we should have non-uniform prior density for $\theta^2$ when there is full ignorance about $\theta$ ?

# Invariant Prior

We want to use a method (M) for choosing a prior density with the following property:
If $\psi = g(\theta)$, $g$ a monotone map, we have used the method (M) to find $\pi$, then the density of $\psi$ given by the method (M) is

$$\pi_{\Psi}(\psi) = \pi\left(g^{-1}(\psi)\right) \cdot \mid \frac{d}{d\psi} g^{-1}(\psi) \mid,$$

which is the standard probability calculus rule for change of variable in a probability density.

# Invariant Prior: Jeffreys' Prior

We shall now describe one method (M), i.e., Jeffreys' prior.
In order to introduce Jeffreys' prior we need first to define Fisher information, which will be needed even for purposes other than choice of prior.

# Fisher Information of $X$

A parametric model $x \sim f(x|\theta)$, where $f(x|\theta)$ is differentiable w.r.t to $\theta \in R$, we define $I(\theta)$, *Fisher information* of $x$, as

$$I(\theta) = \int_{\mathcal{X}} \left( \frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 f(x|\theta) \, \nu(dx)$$

Conditions for existence of $I(\theta)$ are given in Schervish (1995), p. 111.

# Fisher Information of $x$: An Example

$$I(\theta) = E_\theta \left[ \left( \frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \right]$$

Example:

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\theta)^2/2\sigma^2},$$

$\sigma$ is known.

$$\frac{\partial \log f(x|\theta)}{\partial \theta} = \frac{(x-\theta)}{\sigma^2}$$

$$I(\theta) = E\left[ \frac{(x-\theta)^2}{\sigma^4} \right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

KTH Matematik

# Fisher Information of $x$, $\theta \in R^k$

$x \sim f(x|\theta)$, where $f(x|\theta)$ is differentiable w.r.t to $\theta \in R^k$, we define $I(\theta)$, *Fisher information* of $x$, as the matrix

$$I(\theta) = \left(I_{ij}(\theta)\right)_{i,j=1}^{k,k}$$

$$I_{ij}(\theta) = \mathrm{Cov}_\theta\left(\frac{\partial \log f(x|\theta)}{\partial \theta_i}, \frac{\partial \log f(x|\theta)}{\partial \theta_j}\right)$$

# Fisher Information of $x^{(n)}$

Same parametric model $x_i \sim f(x|\theta)$, I.I.D., $x^{(n)} = (x_1, x_2, \ldots, x_n)$.

$$f\left(x^{(n)}|\theta\right) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \ldots \cdot f(x_n|\theta)$$

*Fisher information* of $x^{(n)}$ is

$$I_{x^{(n)}}(\theta) = \int_{\mathcal{X}} \left( \frac{\partial \log f\left(x^{(n)}|\theta\right)}{\partial\theta} \right)^2 f\left(x^{(n)}|\theta\right) \nu\left(dx^{(n)}\right)$$

$$= n \cdot I(\theta).$$

KTH Matematik

# Fisher Information of $x$: another form

A parametric model $x \sim f(x|\theta)$, where $f(x|\theta)$ is twice differentiable w.r.t to $\theta \in R$. If we can write

$$\frac{d}{d\theta}\int_{\mathcal{X}}\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right) f(x|\theta)\,\nu(dx) =$$

$$= \int_{\mathcal{X}}\frac{\partial}{\partial\theta}\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right) f(x|\theta)\,\nu(dx),$$

then

$$I(\theta) = -\int_{\mathcal{X}}\left(\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right) f(x|\theta)\,\nu(dx)$$

# Fisher Information of $x$, $\theta \in R^k$

$x \sim f(x|\theta)$, where $f(x|\theta)$ is differentiable w.r.t to $\theta \in R^k$, then under some conditions

$$I(\theta) = \left[ \left( -E_\theta \left( \frac{\partial^2 \log f(x|\theta)}{\partial\theta_i \partial\theta_j} \right) \right)_{ij} \right]_{i,j=1}^{k,k}$$

# Fisher Information of $x$: Natural Exponential Family

For a natural exponential family

$$f\left(x \mid \theta\right) = h(x)e^{\theta \cdot x - \psi(\theta)}$$

$$\frac{\partial^2 \log f\left(x|\theta\right)}{\partial\theta_i \partial\theta_j} = -\frac{\partial^2 \psi\left(\theta\right)}{\partial\theta_i \partial\theta_j}$$

so no expectation needs to be computed to obtain $I(\theta)$.

# Jeffreys' Prior defined

$$\pi(\theta) := \frac{\sqrt{I(\theta)}}{\int_{\Theta}\sqrt{I(\theta)}d\theta}$$

assuming that the standardizing integral in the denominator exists. Otherwise the prior is improper.

Let $\psi = g\left(\theta\right)$, $g$ a monotone map. The prior $\pi(\theta)$ is Jeffreys' prior. Let us compute the prior density $\pi_{\Psi}(\psi)$ for $\psi$:

$$\pi_{\Psi}(\psi) = \pi\left(g^{-1}(\psi)\right) \cdot \mid \frac{d}{d\psi} g^{-1}\left(\psi\right) \mid$$

$$\propto \sqrt{E_{\theta}\left[\left(\frac{\partial \log f\left(X|\theta\right)}{\partial \theta}\right)^2\right]} \mid \frac{d}{d\psi} g^{-1}\left(\psi\right) \mid$$

$$= \sqrt{E_{g^{-1}((\psi)}\left[\left(\frac{\partial \log f\left(X|g^{-1}\left(\psi\right)\right)}{\partial \theta} \frac{d}{d\psi} g^{-1}\left(\psi\right)\right)^2\right]}$$

$$= \sqrt{E_{g^{-1}(\psi)}\left[\left(\frac{\partial \log f\left(X|g^{-1}\left(\psi\right)\right)}{\partial \psi}\right)^2\right]} = \sqrt{I\left(\psi\right)}$$

Hence the prior for $\psi$ is the Jeffreys$^{\prime}$ prior.

# $\mathcal{F}$ = Binomial Distribution & $\Pi$ =Conjugate Priors

We let $\Theta$ be a random variable, whose values are denoted by $\theta$, $\Omega = (0,1)$. We condition on $\Theta = \theta$, and consider $X$, which is the sum of $n$ I.I.D $\mathrm{Be}(\theta)$ R.V's. Hence for $x = 0, 1, 2, \ldots, n$,

$$f(x|\theta) = P\left(X = x \mid \Theta = \theta\right)$$

$$= \left( \begin{array}{c} n \\ x \end{array} \right) \theta^{x} \cdot (1-\theta)^{n-x},$$

(the Binomial distribution)

# Prior Density

Any function $\pi(\cdot)$ such that

$$\pi\left(\theta\right) \geq 0, \qquad 0 \leq \theta \leq 1,$$

$$\pi\left(\theta\right) = 0 \qquad \text{elsewhere},$$

and

$$\int_0^1 \pi\left(\theta\right) d\theta = 1,$$

can serve as prior distribution.

# Improper Prior Densities

Functions with the properties

$$\pi\left(\theta\right) \geq 0, \qquad 0 \leq \theta \leq 1,$$

$$\pi\left(\theta\right) = 0 \qquad \text{elsewhere},$$

and

$$\int_0^1 \pi\left(\theta\right) d\theta = \infty,$$

are also invoked as prior distributions, and are called improper priors.

# The Posterior Density

Bayes' rule

$$\pi\left(\theta \mid x\right) = \frac{f\left(x \mid \theta\right) \cdot \pi\left(\theta\right)}{\int_0^1 f\left(x \mid \theta\right) \cdot \pi\left(\theta\right) d\theta}, 0 \leq \theta \leq 1$$

and zero elsewhere. The marginal distribution of $x$ is

$$m(x) = \int_0^1 f\left(x \mid \theta\right) \cdot \pi\left(\theta\right) d\theta.$$

# The Posterior Density

Take $\theta \sim \mathrm{U}(0,1)$. i.e.,

$$\pi\left(\theta\right) = \begin{cases} 1 & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

# The marginal distribution of $X$: uniform prior

$$m(x) = \int_0^1 f\left(x \mid \theta\right) \cdot \pi\left(\theta\right) d\theta$$

$$= \left( \begin{array}{c} n \\ x \end{array} \right) \int_0^1 \theta^x \cdot \left(1 - \theta\right)^{n-x} d\theta,$$

where we used the Beta integral

$$= \left( \begin{array}{c} n \\ x \end{array} \right) \frac{x!(n-x)!}{(n+1)!}$$

# The Beta Density

$$\pi(\theta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} & 0 < \theta < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

is a probability density $\mathcal{B}\mathrm{e}(\alpha, \beta)$.

$$\int_0^1 \pi(\theta) d\theta = 1 \Leftrightarrow \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

# The Beta Integral

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Recall also that $\Gamma(x+1) = x!$, if $x$ is a positive integer. $\alpha = \beta = 1$ gives the distribution $U(0,1)$. We set

$$B(\alpha, \beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The Jeffreys prior for $\mathrm{Be}(\theta)$ is $\mathcal{B}\mathrm{e}(1/2, 1/2)$ (i.e., a choice of hyperparameters).

# The marginal distribution of $X$, $\theta \sim U(0,1)$

$$m(x) = \int_0^1 f\left(x \mid \theta\right) \cdot d\theta = \left( \begin{array}{c} n \\ x \end{array} \right) \frac{x!(n-x)!}{(n+1)!}$$

$$= \frac{n!}{x!(n-x)!} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{(n+1)}$$

There is an interpretation of Bayes' work claiming that the problem really attacked and solved by Bayes was: What should $\pi(\theta)$ be so that

$$\int_0^1 f\left(x \mid \theta\right) \cdot \pi(\theta) d\theta = \frac{1}{(n+1)}.$$

Chaloner and Duncan use predictive probabilities in this vein as the method of predictive modal (PM) elicitation the hyperparameters in a Beta density. K.M. Chaloner & G.T. Duncan: Assessment of a Beta Distribution: PM Elicitation. *The Statistician*, 32, 1983, pp. 174–180

# The Posterior Density

$$\pi\left(\theta \mid x\right) = \frac{\left(\begin{array}{c} n \\ x \end{array}\right) \theta^{x} \cdot \left(1-\theta\right)^{n-x}}{m(x)}$$

$$= \begin{cases} \frac{(n+1)!}{x!(n-x)!} \cdot \theta^{k} \left(1-\theta\right)^{n-k} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

# The Posterior Density

$$\frac{(n+1)!}{x!(n-x)!} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} = \frac{1}{B(x+1,n-x+1)}.$$

# The Posterior Density

$$\pi\left(\theta \mid x\right) = \begin{cases} \frac{1}{B(x+1,n-x+1)} \cdot \theta^{x}\left(1-\theta\right)^{n-x} & 0 \leq \theta \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is again a Beta density.

# The Posterior Density $\theta \sim \mathcal{B}\mathrm{e}(\alpha, \beta)$

$$\pi\left(\theta \mid x\right) = \begin{cases} \frac{1}{B(x+\alpha, n-x+\beta)} \cdot \theta^{x+\alpha-1}\left(1-\theta\right)^{\beta+n-x-1} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is Beta density $\mathcal{B}\mathrm{e}(\alpha + x, \beta + n - x)$ .

# The Beta-binomial distribution

Let $\pi(\theta)$ be the density of $\mathcal{B}\mathrm{e}(\alpha, \beta)$. Then for $x = 0, \ldots, n$

$$g(x) = \int_0^1 \binom{n}{x} \theta^x \cdot (1-\theta)^{n-x} \pi(\theta)\, d\theta$$

$$= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}$$

$$= \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}$$

is the Beta-binomial distribution.

# Kullback's Information Measure

Let $f\left(x\right)$ and $g\left(x\right)$ be two densities. Kullback's information measure $I\left(f;g\right)$ is defined as

$$I\left(f;g\right) := \int_{\mathcal{X}} f\left(x\right) \log \frac{f\left(x\right)}{g\left(x\right)} \nu(dx).$$

We intertpret $\log \frac{f(x)}{0} = \infty$, $0 \log 0 = 0$. It can be shown that $I\left(f;g\right) \geq 0$. Kullback's Information Measure does not require the same kind of conditions for existence as the Fisher information.

KTH Matematik

# Kullback's Information Measure: Two Normal Distributions

Let $f\left(x\right)$ and $g\left(x\right)$ be densities for $N\left(\theta_1;\sigma^2\right)$, $N\left(\theta_2;\sigma^2\right)$, respectively. Then

$$\log \frac{f\left(x\right)}{g\left(x\right)} = \frac{1}{2\sigma^2}\left[\left(x-\theta_2\right)^2-\left(x-\theta_1\right)^2\right]$$

$$\begin{aligned} I\left(f;g\right) &= \frac{1}{2\sigma^2}E_{\theta_1}\left[\left(x-\theta_2\right)^2-\left(x-\theta_1\right)^2\right] \\ &= \frac{1}{2\sigma^2}\left[E_{\theta_1}\left(x-\theta_2\right)^2-\sigma^2\right]. \end{aligned}$$

# Kullback's Information Measure: Two Normal Distributions

We have

$$E_{\theta_1}(x-\theta_2)^2 = E_{\theta_1}\left(x^2\right) - 2\theta_2 E_{\theta_1}(x) + \theta_2^2$$

$$= \sigma^2 + \theta_1^2 - 2\theta_2\theta_1 + \theta_2^2 = \sigma^2 + (\theta_1-\theta_2)^2 .$$

Then

$$I(f;g) = \frac{1}{2\sigma^2}\left[\sigma^2 + (\theta_1-\theta_2)^2 - \sigma^2\right] =$$

$$= \frac{1}{2\sigma^2}(\theta_1-\theta_2)^2 .$$

$$I(f;g) = \frac{1}{2\sigma^2}(\theta_1-\theta_2)^2$$

# Kullback's Information Measure: Natural exponential densities

Let $f_i(x) = h(x)e^{\theta_i \cdot x - \psi(\theta_i)}$, $i = 1, 2$. Then

$$I(f_1; f_2) = (\theta_1 - \theta_2) \cdot \nabla_\theta \psi(\theta_1) - (\psi(\theta_1) - \psi(\theta_2))$$

# Summary:

The fact that prior cannot be chosen uniquely is a serious objection to Bayesian statistics. Clearly, conjugate priors are perhaps mainly preferred for mathematical convenience. The question is, how much will the choice of prior influence the statistical conclusions and decisions ?

# Robustness and Sensitivity

There are robustness and sensitivity analyses of the impact of choice of prior on the posterior. Some of this (as known to the lecturer) requires mathematical tools that are not readily presentable here.