Conjugate priors

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

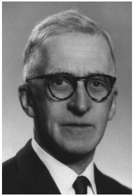
Outline

Introduction

- Exponential families (expfams) are a unifying generalization of many basic models, and they possess many nice properties.
- In Bayesian statistics, a key feature of expfams is that the posterior often has a nice form when using conjugate priors.
- Individually, expfams are often too simple for real applications.
- However, they can easily be combined to build complex hierarchical models that are amenable to inference with Markov chain Monte Carlo or variational inference.
- Examples of exponential families:
  - Bernoulli, binomial, Poisson, exponential, beta, gamma, inverse gamma, normal (Gaussian), multivariate Gaussian, log-normal, inverse Gaussian, multinomial, Dirichlet.

Introduction

Pitman



Koopman

Darmois

- The concept of exponential families was developed by E. J. G. Pitman (1897–1993), Bernard Koopman (1900–1981), and Georges Darmois (1888–1960).

Outline

One-parameter exponential families

- A *one-parameter exponential family* is a collection of distributions indexed by $\theta \in \Theta$, with p.d.f.s/p.m.f.s of the form
$$p(x|\theta) = \exp\big(\varphi(\theta)t(x) - \kappa(\theta)\big)h(x)$$
for some functions $\varphi(\theta), t(x), \kappa(\theta)$, and $h(x)$.

- $\kappa(\theta)$ is a log-normalization constant: since $\int p(x|\theta)dx = 1$,
$$\kappa(\theta) = \log \int \exp\big(\varphi(\theta)t(x)\big)h(x)\,dx.$$

- $t(x)$ is called the *sufficient statistic*.

Examples of one-parameter expfams

- The $\text{Exp}(\theta)$ distributions form an exponential family, since the p.d.f.s are

$$p(x|\theta) = \theta e^{-\theta x} \mathrm{I}(x > 0) = \exp\big(\varphi(\theta) t(x) - \kappa(\theta)\big) h(x)$$

for $\theta \in \Theta = (0, \infty)$, where $t(x) = -x$, $\varphi(\theta) = \theta$, $\kappa(\theta) = -\log\theta$, and $h(x) = \mathrm{I}(x > 0)$.

- The $\text{Poisson}(\theta)$ distributions form an exponential family, since the p.m.f.s are

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \mathrm{I}(x \in S) = ??? \quad \text{(Whiteboard)}$$

for $\theta \in \Theta = ???$, where $S = \{0, 1, 2, \ldots\}$, $t(x) = ???$, $\varphi(\theta) = ???$, $\kappa(\theta) = ???$, and $h(x) = ???$.

Examples of one-parameter expfams

- The $\mathrm{Exp}(\theta)$ distributions form an exponential family, since the p.d.f.s are

$$p(x|\theta) = \theta e^{-\theta x}\mathrm{I}(x > 0) = \exp\big(\varphi(\theta)t(x) - \kappa(\theta)\big)h(x)$$

  for $\theta \in \Theta = (0, \infty)$, where $t(x) = -x$, $\varphi(\theta) = \theta$, $\kappa(\theta) = -\log\theta$, and $h(x) = \mathrm{I}(x > 0)$.

- The $\mathrm{Poisson}(\theta)$ distributions form an exponential family, since the p.m.f.s are

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}\mathrm{I}(x \in S) = \exp\big(\varphi(\theta)t(x) - \kappa(\theta)\big)h(x)$$

  for $\theta \in \Theta = (0, \infty)$, where $S = \{0, 1, 2, \ldots\}$, $t(x) = x$, $\varphi(\theta) = \log\theta$, $\kappa(\theta) = \theta$, and $h(x) = \mathrm{I}(x \in S)/x!$.

Outline

Conjugate priors

- Consider some family of distributions $\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\}$.

- A family of priors $\{p_\alpha(\theta) : \alpha \in H\}$ is *conjugate* for $\mathcal{M}$ if for any α and any data, the resulting posterior equals $p_{\alpha'}(\theta)$ for some $\alpha' \in H$.

- Example: $\{\text{Beta}(\theta|a, b) : a, b > 0\}$ is a conjugate prior family for $\{\text{Bernoulli}(\theta) : \theta \in (0, 1)\}$ since the posterior is

$$p(\theta|x_{1:n}) = \text{Beta}(\theta \mid a + \sum x_i,\ b + n - \sum x_i).$$

- Example: $\{\text{Gamma}(\theta|a, b) : a, b > 0\}$ is a conjugate prior family for $\{\text{Exp}(\theta) : \theta > 0\}$ since the posterior is

$$p(\theta|x_{1:n}) = ??? \quad \text{(Whiteboard)}.$$

Conjugate priors

- Consider some family of distributions $\mathcal{M} = \{p(x|\theta) : \theta \in \Theta\}$.

- A family of priors $\{p_\alpha(\theta) : \alpha \in H\}$ is *conjugate* for $\mathcal{M}$ if for any α and any data, the resulting posterior equals $p_{\alpha'}(\theta)$ for some $\alpha' \in H$.

- Example: $\{\mathrm{Beta}(\theta|a, b) : a, b > 0\}$ is a conjugate prior family for $\{\mathrm{Bernoulli}(\theta) : \theta \in (0, 1)\}$ since the posterior is

$$p(\theta|x_{1:n}) = \mathrm{Beta}(\theta \mid a + \textstyle\sum x_i,\, b + n - \sum x_i).$$

- Example: $\{\mathrm{Gamma}(\theta|a, b) : a, b > 0\}$ is a conjugate prior family for $\{\mathrm{Exp}(\theta) : \theta > 0\}$ since the posterior is

$$p(\theta|x_{1:n}) = \mathrm{Gamma}(\theta \mid a + n,\, b + \textstyle\sum x_i).$$

Conjugate priors for exponential families

- Under general conditions, for any exponential family there is a family of conjugate priors with p.d.f.

$$p_{n_0,t_0}(\theta) \propto \exp\big(n_0 t_0 \varphi(\theta) - n_0 \kappa(\theta)\big) \mathrm{I}(\theta \in \Theta)$$

  for all $n_0 > 0$ and $t_0 \in \mathbb{R}$ for which this is normalizable.

- The resulting posterior is $p_{n',t'}(\theta)$ where $n' = ???$ and

$$t' = ???$$

(Whiteboard)

Conjugate priors for exponential families

- Under general conditions, for any exponential family there is a family of conjugate priors with p.d.f.

$$p_{n_0,t_0}(\theta) \propto \exp\big(n_0 t_0 \varphi(\theta) - n_0 \kappa(\theta)\big) \mathrm{I}(\theta \in \Theta)$$

for all $n_0 > 0$ and $t_0 \in \mathbb{R}$ for which this is normalizable.

- The resulting posterior is $p_{n',t'}(\theta)$ where $n' = n_0 + n$ and

$$t' = \frac{n_0 t_0 + \sum_{i=1}^n t(x_i)}{n_0 + n} = \frac{n_0}{n_0 + n} t_0 + \frac{n}{n_0 + n} \frac{1}{n} \sum_{i=1}^n t(x_i).$$

- Note that t' is a convex combination of t_0 and $\frac{1}{n} \sum t(x_i)$.
- This helps interpret and select the hyperparameters t_0, n_0:
  - t_0 represents a prior "guess" at the expected value of $t(x)$, and
  - n_0 represents the prior "number of samples" (roughly speaking, how certain we are about t_0).

Conjugate priors for exponential families

- In most cases, it is probably just as easy to guess a conjugate prior and verify it, rather than use this general construction.
- Also, in some cases, this construction is not the most convenient, for example, for the $\mathcal{N}(\mu, \sigma^2)$ model.
- Rather, the purpose of showing this construction is to provide:
  - intuition for how to derive conjugate priors and posteriors,
  - understanding of how to interpret prior parameters, and
  - a theoretical result on existence of conjugate priors for exponential families.

Outline

Multi-parameter exponential families

- The generalization to more than one parameter is straightforward.
- An *exponential family* is a collection of distributions indexed by $\theta \in \Theta$, with p.d.f.s/p.m.f.s of the form

$$p(x|\theta) = \exp\big(\varphi(\theta)^{\mathtt{T}} t(x) - \kappa(\theta)\big) h(x)$$

  for some vector-valued functions

$$\varphi(\theta) = \begin{pmatrix} \varphi_1(\theta) \\ \vdots \\ \varphi_k(\theta) \end{pmatrix} \quad \text{and} \quad t(x) = \begin{pmatrix} t_1(x) \\ \vdots \\ t_k(x) \end{pmatrix}$$

  and some real-valued functions $\kappa(\theta)$ and $h(x)$.
- As before, $\kappa(\theta)$ is the log-normalization constant.
- Conjugate priors can be constructed in the same way as the one-parameter case, except that now $t_0 \in \mathbb{R}^k$ and $t' \in \mathbb{R}^k$.

Example of multi-param exponential family

- The $\mathrm{Gamma}(a, b)$ distributions, with $a, b > 0$, are an exponential family:

$$\begin{aligned}\mathrm{Gamma}(x|a,b) &= \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \mathrm{I}(x > 0) \\ &= ??? \quad \text{(Whiteboard)}\end{aligned}$$

where $\theta = ???$, $\varphi(\theta) = ???$, $t(x) = ???$, and $h(x) = ???$.

Example of multi-param exponential family

- The $\text{Gamma}(a, b)$ distributions, with $a, b > 0$, are an exponential family:

$$\begin{aligned}\text{Gamma}(x|a, b) &= \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \mathrm{I}(x > 0) \\ &= \exp\big(\varphi(\theta)^{\mathtt{T}} t(x) - \kappa(\theta)\big) h(x)\end{aligned}$$

  where $\theta = (a, b)^{\mathtt{T}}$, $\varphi(\theta) = (-b, a-1)^{\mathtt{T}}$, $t(x) = (x, \log x)^{\mathtt{T}}$, and $h(x) = \mathrm{I}(x > 0)$.

Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/o9tzjuC3BXqkV4hZ8

(Three people per room, randomly assigned. 15 minutes.)

Outline

Constructing new conjugate priors by reweighting

- Technically, for any model $\mathcal{M}$, there exists a conjugate prior family—namely, the set of all distributions on Θ.
- However, usually people only consider conjugate priors that are computationally tractable or closed form.

- Reweighting is a one way of constructing new conjugate priors from existing ones:
  - Suppose $\{p_\alpha(\theta) : \alpha \in H\}$ is conjugate for a model $\mathcal{M}$.
  - Let $g(\theta)$ be any nonnegative function, and define $z(\alpha) = \int p_\alpha(\theta) g(\theta) d\theta$.
  - If $0 < z(\alpha) < \infty$ for all $\alpha \in H$, then

$$\left\{p_\alpha(\theta) g(\theta)/z(\alpha) : \alpha \in H\right\}$$

    is also a conjugate prior family.

- A useful special case is to take $g(\theta) = \mathrm{I}(\theta \in A)$ for some A.
- If $p_\alpha(\theta)$ is computationally nice and $g(\theta)$ is well-chosen, then the reweighted family is often computationally nice too.

Constructing new conjugate priors by mixing

- Mixtures are another way of making new conjugate priors:
  If $\{p_\alpha(\theta) : \alpha \in H\}$ is a conjugate prior family for $\mathcal{M}$, then

$$\Big\{\sum_{i=1}^k \pi_i p_{\alpha_i}(\theta) \;:\; \alpha_1,\ldots,\alpha_k \in H,\, \pi \in \Delta_k\Big\}$$

  is also conjugate for $\mathcal{M}$, where

$$\Delta_k = \Big\{\pi \in \mathbb{R}^k \;:\; \pi_1,\ldots,\pi_k \geq 0,\, \sum_{i=1}^k \pi_i = 1\Big\}.$$

- In other words, finite mixtures of conjugate priors are conjugate priors.

- If $p_\alpha(\theta)$ is computationally nice, then the mixture family will usually be nice as well, in terms of posterior computation.

- By reweighting and mixing, we can construct very flexible classes of computationally nice conjugate priors.

Outline

Table of conjugate priors

- Wikipedia has a nice table for reference — but double-check it for correctness. (`https://en.wikipedia.org/wiki/Conjugate_prior`)

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^{n} x_i,\ \beta+n-\sum_{i=1}^{n} x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$p(\tilde{x}=1)=\frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^{n} x_i,\ \beta+\sum_{i=1}^{n} N_i-\sum_{i=1}^{n} x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\mid\alpha',\beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^{n} x_i,\ \beta+rn$	$\alpha-1$ total successes, $\beta-1$ failures[note 1] (i.e., $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	$\operatorname{BetaNegBin}(\tilde{x}\mid\alpha',\beta')$ (beta-negative binomial)
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k+\sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta+1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x}\mid k',\theta')$ (negative binomial)
			$\alpha,\ \beta$[note 3]	$\alpha+\sum_{i=1}^{n} x_i,\ \beta+n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x}\mid\alpha',\frac{1}{1+\beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+(c_1,\ldots,c_k)$, where c_i is the number of observations in category i	α_i-1 occurrences of category i[note 1]	$p(\tilde{x}=i)=\frac{\alpha_i{}'}{\sum_i \alpha_i{}'} = \frac{\alpha_i+c_i}{\sum_i \alpha_i+n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+\sum_{i=1}^{n}\mathbf{x}_i$	α_i-1 occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\mid\boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total	M (number of target	Beta-	$n=N, \alpha, \beta$	$\alpha+\sum_{i=1}^{n} x_i,\ \beta+\sum_{i=1}^{n} N_i-\sum_{i=1}^{n} x_i$	$\alpha-1$ successes, $\beta-1$	

Some expfam forms are more convenient than others

- There are multiple ways of putting a distribution in expfam form, some of which may be more useful than others.

- Example: We can write $\mathcal{N}(\mu, \sigma^2)$ as a two-param expfam,

$$\begin{aligned}\mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\tfrac{1}{2\sigma^2}(x-\mu)^2\big) \\ &= \exp\Big(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\Big) \\ &= \exp\big(\theta^{\mathtt{T}} t(x) - \kappa(\theta)\big)\end{aligned}$$

  where $\theta = \Big(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\Big)^{\mathtt{T}}$ and $t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$.

- However, we usually prefer to keep μ and σ^2 separate to facilitate prior specification.

Not all conjugate priors are useful

- We mentioned that for any model, the set of all distributions on Θ is a conjugate prior family — albeit a useless one.

- Even when a parametric conjugate prior exists, it is not always very useful.

- Example: Damsleth (1975) showed that for $\beta, \nu > 0$,

$$p_{\beta,\nu}(a) \propto \frac{\beta^a}{\Gamma(a)^\nu} \, \mathrm{I}(a > 0)$$

  is a conjugate prior on the shape parameter a of a Gamma distribution, $\mathrm{Gamma}(x|a, b) = \dfrac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$.

- However, this family is difficult to use, computationally — it does not seem to permit closed-form calculation of samples, moments, or the normalization constant.

Auxiliary variable trick for mixtures of expfams

- Mixtures of expfams are often computationally tractable, for instance:
  - t-distribution (a continuous mixture of normals with common mean), useful for robustness to outliers,
  - mixture of Gaussians (a discrete mixture of normals with different means), useful for handling heterogeneity, or
  - models where expectation–maximization would be useful.
- Auxiliary variable trick (surprisingly powerful!):
  - Suppose the likelihood is $p(x|\theta) = \sum_z p(x|z,\theta)p(z|\theta)$.
  - Suppose $p(x|z,\theta)$ is easy to work with (e.g., an expfam).
  - Sample from the joint posterior on z and θ, i.e., $p(z,\theta|x)$.
  - Note that if $(Z, \boldsymbol{\theta}) \sim p(z,\theta|x)$, then $\boldsymbol{\theta} \sim p(\theta|x)$.
  - So just keep the θ part of each sample and discard the z part.

References and supplements

- Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. The Annals of Statistics, 7(2), 269-281.
- E. Damsleth (1975). Conjugate Classes for Gamma Distributions. Scandinavian Journal of Statistics, 2(2), 80-84.
- Hoffman-Jorgensen, J. (1994). Probability with a view towards statistics. CRC Press.

Individual activity: Exit ticket

Answer these questions individually:
https://forms.gle/KHYKT28AaVSFp5jy7