Exponential Families

David M. Blei

1 Introduction

- We discuss the **exponential family**, a very flexible family of distributions.

- Most distributions that you have heard of are in the exponential family.
  - Bernoulli, Gaussian, Multinomial, Dirichlet, Gamma, Poisson, Beta

2 Set-up

- An exponential family distribution has the following form,

$$p(x \mid \eta) = h(x) \exp\{\eta^\top t(x) - a(\eta)\} \tag{1}$$

- The different parts of this equation are
  - The natural parameter η
  - The sufficient statistic $t(x)$
  - The underlying measure $h(x)$, e.g., counting measure or Lebesgue measure
  - The log normalizer $a(\eta)$,

$$a(\eta) = \log \int h(x) \exp\{\eta^\top t(x)\}. \tag{2}$$

    Here we integrate the unnormalized density over the sample space. This ensures that the density integrates to one.

- The statistic $t(x)$ is called *sufficient* because the likelihood for η only depends on x through $t(x)$.

- The exponential family has fundamental connections to the world of graphical models. For our purposes, we'll use exponential families as components in directed graphical models, e.g., in the mixtures of Gaussians.

3 The Gaussian distribution

- As a running example, consider the Gaussian distribution.

- The familiar form of the univariate Gaussian is

$$p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{3}$$

- We put it in exponential family form by expanding the square

$$p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma \right\} \tag{4}$$

- We see that

$$\begin{aligned}
\eta &= \langle \mu/\sigma^2, -1/2\sigma^2 \rangle && (5)\\
t(x) &= \langle x, x^2 \rangle && (6)\\
a(\eta) &= \mu^2/2\sigma^2 + \log\sigma && (7)\\
&= -\eta_1^2/4\eta_2 - (1/2)\log(-2\eta_2) && (8)\\
h(x) &= 1/\sqrt{2\pi} && (9)
\end{aligned}$$

- If you are new to this, work it out for others on the list.

4 Moments

- The derivatives of the log normalizer gives the moments of the sufficient statistics,

$$\begin{aligned}
\frac{d}{d\eta} a(\eta) &= \frac{d}{d\eta}\left(\log \int \exp\left\{ \eta^\top t(x) \right\} h(x) dx \right) && (10)\\
&= \frac{\int t(x) \exp\left\{ \eta^\top t(x) \right\} h(x) dx}{\int \exp\left\{ \eta^\top t(x) \right\} h(x) dx} && (11)\\
&= \int t(x) \exp\left\{ \eta^\top t(x) - a(\eta) \right\} h(x) dx && (12)\\
&= \mathrm{E}\left[t(X) \right] && (13)
\end{aligned}$$

- The next derivatives are higher moments. The second derivative is the variance, etc.

- Let's go back to the Gaussian example.

- The derivative with respect to η_1 is

$$\begin{aligned} \frac{da(\eta)}{d\eta_1} &= -\frac{\eta_1}{2\eta_2} && (14)\\ &= \mu && (15)\\ &= \mathrm{E}[X] && (16) \end{aligned}$$

- The derivative with respect to η_2 is

$$\begin{aligned} \frac{da(\eta)}{d\eta_2} &= \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} && (17)\\ &= \sigma^2 + \mu^2 && (18)\\ &= \mathrm{E}[X^2] && (19) \end{aligned}$$

- This means that the variance is

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 && (20)\\ &= -\frac{1}{2\eta_2} && (21) \end{aligned}$$

- In a **minimal exponential family**, the components of the sufficient statistics $t(x)$ are linearly independent.

- In a minimal exponential family, the mean $\mu := \mathrm{E}[t(X)]$ is another parameterization of the distribution. That is, there is a 1-1 mapping between η and μ.
  - The function $a(\eta)$ is convex. (It is log-sum-exponential.)
  - Thus there is a 1-1 mapping between its argument and its derivative.
  - Thus there is a 1-1 mapping between η and $\mathrm{E}[t(X)]$.

- Side note: the MLE of an exponential family matches the mean parameters with the empirical statistics of the data.
  - Assume $x_{1:n}$ are from an exponential family.
  - Find $\hat{\eta}$ that maximizes the likelihood of x.
  - This is the η such that $\mathrm{E}[t(X)] = (1/n)\sum_i t(x_i)$.

5 Conjugacy

- Consider the following set up:

$$\begin{aligned} \eta &\sim F(\cdot \mid \lambda) && (22)\\ x_i &\sim G(\cdot \mid \eta) \quad \text{for } i \in \{1, \ldots, n\}. && (23) \end{aligned}$$

- This is a classical Bayesian data analysis setting. And, this is used as a component in more complicated models, e.g., in hierarchical models.

- The posterior distribution of η given the data $x_{1:n}$ is

$$p(\eta \mid x_{1:n}, \lambda) \propto F(\eta \mid \lambda) \prod_{i=1}^{n} G(x_i \mid \eta). \tag{24}$$

  When this distribution is in the same family as F, i.e., when its parameters are part of the parameter-space defined by λ, then we say that F and G make a **conjugate pair**.

- For example,
  - A Gaussian likelihood with fixed variance, and a Gaussian prior on the mean
  - A multinomial likelihood and a Dirichlet prior on the probabilities
  - A Bernoulli likelihood and a beta prior on the bias
  - A Poisson likelihood and a gamma prior on the rate

  In all these settings, the conditional distribution of the parameter given the data is in the same family as the prior.

- Suppose the data come from an exponential family. Every exponential family has a conjugate prior (in theory),

$$\begin{aligned} p(x_i \mid \eta) &= h_\ell(x) \exp\{\eta^\top t(x_i) - a_\ell(\eta)\} & (25)\\ p(\eta \mid \lambda) &= h_c(\eta) \exp\{\lambda_1^\top \eta + \lambda_2^\top (-a_\ell(\eta)) - a_c(\lambda)\}. & (26) \end{aligned}$$

  - The natural parameter $\lambda = \langle \lambda_1, \lambda_2 \rangle$ has dimension $\dim(\eta) + 1$.
  - The sufficient statistics are $\langle \eta, -a(\eta) \rangle$.
  - The other terms depend on the form of the exponential family. For example, when η are multinomial parameters then the other terms help define a Dirichlet.

- Let's compute the posterior,

$$\begin{aligned} p(\eta \mid x_{1:n}, \lambda) &\propto p(\eta \mid \lambda) \prod_{i=1}^{n} p(x_i \mid \eta) & (27)\\ &= h(\eta) \exp\{\lambda_1^\top \eta + \lambda_2(-a(\eta)) - a_c(\lambda)\} & (28)\\ &\quad \cdot \quad (\textstyle\prod_{i=1}^{n} h(x_i)) \exp\{\eta^\top \sum_{i=1}^{n} t(x_i) - n a_x(\eta)\} & (29)\\ &\propto h(\eta) \exp\{(\lambda_1 + \textstyle\sum t(x_i))^\top \eta + (\lambda_2 + n)(-a(\eta))\}. & (30) \end{aligned}$$

This is the same exponential family as the prior, with parameters

$$\hat{\lambda}_1 = \lambda_1 + \sum_{i=1}^{n} t(x_i) \tag{31}$$

$$\hat{\lambda}_2 = \lambda_2 + n. \tag{32}$$

6 Example: Data from a unit variance Gaussian

- Suppose the data x_i come from a unit variance Gaussian

$$p(x \,|\, \mu) = \frac{1}{\sqrt{2\pi}} \exp\{-(x-\mu)^2/2\}. \tag{33}$$

- This is a simpler exponential family than the previous Gaussian

$$p(x \,|\, \mu) = \frac{\exp\{-x^2/2\}}{\sqrt{2\pi}} \exp\{\mu x - \mu^2/2\}. \tag{34}$$

  In this case

$$\eta = \mu \tag{35}$$

$$t(x) = x \tag{36}$$

$$h(x) = \frac{\exp\{-x^2/2\}}{\sqrt{2\pi}} \tag{37}$$

$$a(\eta) = \mu^2/2 = \eta^2/2. \tag{38}$$

- We are interested in the conjugate prior. (State the end result on the next page.)

- Consider a model with an unknown mean. What is the conjugate prior? It is

$$p(\eta \,|\, \lambda) = h(\eta) \exp\{\lambda_1 \eta + \lambda_2(-\eta^2/2) - a_c(\lambda)\} \tag{39}$$

- Set $\lambda_1^* = \lambda_1$ and $\lambda_2^* = -\lambda_2/2$. This means the sufficient statistics are $\langle \eta, \eta^2 \rangle$.

- This is a **Gaussian distribution**. We now know the conjugate prior.

- Now consider the posterior,

$$\hat{\lambda}_1 = \lambda_1 + \textstyle\sum_{i=1}^{n} x_i \tag{40}$$

$$\hat{\lambda}_2 = \lambda_2 + n \tag{41}$$

$$\hat{\lambda}_2^* = \frac{-(\lambda_2 + n)}{2}. \tag{42}$$

- Let's map this back to traditional Gaussian parameters.
  - The mean is
$$\mathrm{E}[\mu \mid x_{1:n}, \lambda] = \frac{\lambda_1 + \sum_{i=1}^{n} x_i}{\lambda_2 + n} \tag{43}$$
  - The variance is
$$\mathrm{Var}(\mu \mid x_{1:n}, \lambda) = \frac{1}{\lambda_2 + n} \tag{44}$$

- Finally, for closure, let's parameterize everything in the mean parameterization.
  - Consider a prior mean and prior variance $\{\mu_0, \sigma_0^2\}$.
  - We know that
$$\begin{aligned} \lambda_1 &= \mu_0/\sigma_0^2 && (45)\\ \lambda_2 &= -1/2\sigma_0^2 && (46)\\ \lambda_2^* &= 1/\sigma_0^2. && (47) \end{aligned}$$
    The expression λ_2^* is also called the **precision**.
  - So the posterior mean is
$$\mathrm{E}[\mu \mid x_{1:n}, \mu_0, \sigma_0^2] = \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^{n} x_i}{1/\sigma_0^2 + n} \tag{48}$$
  - The posterior variance is
$$\mathrm{Var}(\mu \mid x_{1:n}, \mu_0, \sigma_0^2) = \frac{1}{1/\sigma_0^2 + n} \tag{49}$$

- Intuitively, when we haven't seen any data then our estimate of the mean is the *prior mean*. As we see more data, our estimate of the mean moves towards the *sample mean*.

  Before seeing data, our "confidence" about the estimate is the prior variance. As we see more data, the confidence decreases.