Variational Inference: The Exponential Families

Chunpai Wang

June 05, 2016

1 Exponential Families

The exponential family of distribution includes Gaussian,Binomial, Multinomial, Poisson, Gamma, Von Mises, and Beta etc. Given a parameter vector η, the exponential family of probability distributions have the following general form:

$$p(x|\eta) = h(x)\exp\{\eta^\top \phi(x) - A(\eta)\} \tag{1}$$

where

- the parameter η often referred to as the canonical parameter or natural parameters.
- $\phi(X)$ is referred to as a sufficient statistic, and sometimes can be understood as potential function, which is only a function of our data X.
- the function $A(\eta)$ is known as the cumulant function, which can be viewed as the logarithm of a normalization factor. In other words, it is used to make the integral of $p(x|\eta) = 1.0$. This is a function of parameter η.

$$\int p(x|\eta)dx = \int h(x)\frac{\exp(\eta^\top \phi(x))}{\exp A(\eta)}dx = 1.0 \tag{2}$$

we can get

$$\exp A(\eta) = \left(\int h(x)\exp(\eta^\top \phi(x))dx\right) \tag{3}$$

$$A(\eta) = \log\left(\int h(x)\exp(\eta^\top \phi(x))dx\right) \tag{4}$$

and $A(\eta)$ is called log-normalizer or log-partition function, which will be very important in variational inference.

$$\frac{\partial A}{\partial \eta} = \frac{\int h(x)\exp(\eta^\top \phi(x))\cdot\phi(x)dx}{\int h(x)\exp(\eta^\top \phi(x))dx} \tag{5}$$

$$= \frac{\int h(x)\exp(\eta^\top \phi(x))\cdot\phi(x)dx}{\exp(A(\eta))} \tag{6}$$

$$= \int h(x)\exp(\eta^\top \phi(x) - A(\eta))\cdot\phi(x)dx \tag{7}$$

$$= E(\phi(x)) \tag{8}$$

$$\frac{\partial^2 A}{\partial \eta} = \int h(x)\exp(\eta^\top \phi(x) - A(\eta))\cdot\phi(x)\cdot(\phi(x) - A'(\eta))dx \tag{9}$$

$$= \int p(x|\eta)\cdot\phi^2(x)dx - \int p(x|\eta)\cdot\phi(x)dx\cdot A'(\eta) \tag{10}$$

$$= \int p(x|\eta)\cdot\phi^2(x)dx - \int p(x|\eta)\cdot\phi(x)dx\cdot E(\phi(x)) \tag{11}$$

$$= E(\phi^2(x)) - E^2(\phi(x)) \tag{12}$$

$$= Var(\phi(x)) \tag{13}$$

- $A(\eta)$ is convex (second derivative is p.s.d)., because variance has all eigenvalues greater than 0.

We can rewrite some distributions into canonical exponential family form. The reason of exponential family distribution is helpful is because, when our conditional distribution can be expressed into the canonical exponential family form, it would be much easier to use the variational inference to get the result.

1.1 The Univariate Gaussian Distribution

The univariate Gaussian density can be written as follows (where the underlying measure is Lebesgue measure):

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{14}$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu^2 - \log\sigma\right\} \tag{15}$$

$$= h(x)\exp\{\eta^\top\phi(x) - A(\eta)\} \tag{16}$$

Then, we can write it in canonical exponential family form with

$$\eta = \begin{bmatrix}\mu/\sigma^2 \\ -1/2\sigma^2\end{bmatrix} \tag{17}$$

$$\phi(x) = \begin{bmatrix}x \\ x^2\end{bmatrix} \tag{18}$$

$$A(\eta) = \frac{\mu^2}{2\sigma^2} - \log\sigma = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \tag{19}$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \tag{20}$$

The univariate Gaussian distribution is a two-parameter distribution and that its sufficient statistic is a vector.

1.2 Parameter Estimation of η with Maximal Likelihood

We will try to use MLE to estimate the parameter η of canonical exponential family form, and we will show why the canonical form is helpful and why $\phi(x)$ is called sufficient statistics. Given the dataset $X = \{x_1, ..., x_n\}$, the log-likelihood can be expressed as

$$\arg\max_\eta \left[\log p(X|\eta)\right] = \arg\max_\eta \left[\log\prod_{i=1}^n p(x_i|\eta)\right] \tag{21}$$

$$= \arg\max_\eta \left[\log\prod_{i=1}^n \left(h(x_i)\exp\{\eta^\top\phi(x_i) - A(\eta)\}\right)\right] \tag{22}$$

$$= \arg\max_\eta \left[\log\left\{\left(\prod_{i=1}^n h(x_i)\right)\exp\left\{\eta^\top\left(\sum_{i=1}^n\phi(x_i)\right) - n\cdot A(\eta)\right\}\right\}\right] \tag{23}$$

$$= \arg\max_\eta \left\{\eta^\top\left(\sum_{i=1}^n\phi(x_i)\right) - n\cdot A(\eta)\right\} \tag{24}$$

$$\tag{25}$$

We take the derivative w.r.t η and set it to zero, and we get

$$\sum_{i=1}^n \phi(x_i) - n\cdot A'(\eta) = 0 \tag{26}$$

$$A'(\eta) = \frac{\sum_{i=1}^n\phi(x_i)}{n} \tag{27}$$

We can see that the estimator of η is solely related to our data X and $\phi(x_i)$, and that is the reason $\phi(x_i)$ called the sufficient statistics. For example, in the univariate Gaussian we have

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \tag{28}$$

thus

$$\begin{bmatrix} \frac{A'(\eta)}{\eta_1} \\ \frac{A'(\eta)}{\eta_2} \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \end{bmatrix} = \frac{\sum_{i=1}^n \phi(x_i)}{n} = \begin{bmatrix} \frac{\sum_i x_i}{n} \\ \frac{\sum_i x_i^2}{n} \end{bmatrix} \tag{29}$$

2 Conjugate Prior

$$p(\eta|X) \propto p(X|\eta) * p(\eta|\alpha) \tag{30}$$

If the likelihood is in exponential family, and the prior is also in exponential family with sufficient statistics in a specific form, then we get the conjugacy and the posterior is also in the exponential family.

$$\begin{aligned} p(\eta|X) &\propto p(X|\eta) * p(\eta|\alpha) && (31)\\ &= h(X)\exp\{\eta^\top \phi(X) - A(\eta)\} * h(\eta)\exp\{\alpha^\top \phi(\eta) - A(\alpha)\} && (32)\\ &= h(\eta)\exp\{\eta^\top \phi(X) - A(\eta) + \alpha^\top \phi(\eta)\} * h(X)\exp\{A(\alpha)\} && (33)\\ &\propto h(\eta)\exp\{\eta^\top \phi(X) - A(\eta) + \alpha^\top \phi(\eta)\} && (34) \end{aligned}$$

Now, we assume η is one dimensional for simplicity, and let

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \tag{35}$$

$$\phi(\eta) = \begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix} \tag{36}$$

then we will get

$$\begin{aligned} p(\eta|X) &\propto h(\eta)\exp\{\eta^\top \phi(X) - A(\eta) + \alpha^\top \phi(\eta)\} && (37)\\ &= h(\eta)\exp\{\eta * \phi(X) - A(\eta) + \alpha_1 * \eta - \alpha_2 * A(\eta)\} && (38)\\ &= h(\eta)\exp\{(\phi(X) + \alpha_1) * \eta - (1 + \alpha_2) * A(\eta)\} && (39)\\ &= h(\eta)\exp\{\hat{\eta}^\top \phi(\eta)\} && (40)\\ & && (41) \end{aligned}$$

which is in the exponential family, where

$$\hat{\eta} = \begin{bmatrix} \phi(X) + \alpha_1 \\ (1 + \alpha_2) \end{bmatrix} \tag{42}$$

and

$$\phi(\eta) = \begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix} \tag{43}$$

We can see that the prior and posterior are in the same exponential family form, and only the parameters are different. Thus, they are conjugate if only we set the sufficient statistics of the prior to the assumed form.

3 Another Property of Exponential Family Distribution

Another property is

$$\frac{\partial A(\eta)}{\partial \eta} = E_{p(X|\eta)}\left[\phi(X)\right] \tag{44}$$

we can see that left hand side is a form about η, and the right hand side is

$$\int p(X|\eta)\phi(X)dX \tag{45}$$

is also a function of η after we marginalize the X. Now, we will prove the property. First, we have

$$p(X|\eta) = h(X)\exp\{\eta^\top \phi(X) - A(\eta)\} \tag{46}$$

If we take the integral, we get

$$\int h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}dX = 1.0 \tag{47}$$

Now, we take the derivative w.r.t η, we have

$$\frac{\partial \int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}dX}{\partial \eta} = 0 \tag{48}$$

$$\int_X \frac{\partial h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}}{\partial \eta}dX = 0 \tag{49}$$

$$\int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}[\phi(X) - A'(\eta)]dX = 0 \tag{50}$$

$$\int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}\phi(X)dX - \int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}A'(\eta)dX = 0 \tag{51}$$

$$\int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}\phi(X)dX = \int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}A'(\eta)dX \tag{52}$$

$$\int_X h(X)\exp\{\eta^\top \phi(X) - A(\eta)\}\phi(X)dX = A'(\eta) \tag{53}$$

$$E_{p(X|\eta)}\left[\phi(X)\right] = A'(\eta) \tag{54}$$

4 Exponential Family in Variational Inference

We will see that the exponential family distributions will make maximizing the ELBO easier. Recall that, the log-likelihood is equal to the ELBO + KL divergence.

$$F(q,\theta) = \int_y q(Y)\log\frac{p(X,Y|\theta)}{q(Y)}dy \tag{55}$$

$$= \int_y q(Y)\log\frac{p(Y|X,\theta)\cdot p(X|\theta)}{q(Y)}dy \tag{56}$$

$$= \int_y q(Y)\log\frac{p(Y|X,\theta)}{q(Y)}dy + \int_y q(Y)\log p(X|\theta)dy \tag{57}$$

$$= -KL(q(Y)\|p(Y|X,\theta)) + L(\theta|X) \tag{58}$$

$$= L(\theta|X) - KL(q(Y)\|p(Y|X,\theta)) \tag{59}$$

$$\tag{60}$$

In order to maximize the log-likelihood, we alternate to optimize the ELBO.

$$F(q,\theta) = \int_y q(Y)\log\frac{p(X,Y|\theta)}{q(Y)}dy \tag{61}$$

$$= \int_y q(Y)\log p(X,Y|\theta)dy - \int_y q(Y)\log q(Y)dy \tag{62}$$

$$= E_{q(Y)}\left[\log p(X,Y|\theta)\right] - E_{q(Y)}\left[\log p(Y)\right] \tag{63}$$

where the Y represent the latent random variables. To make the problem simple, we assume the hidden parameters are $Y = [\beta_1, \beta_2]$ and neglect the parameter θ in expression, then we have

$$F(q(\beta_1, \beta_2)) = E_{q(\beta_1,\beta_2)}\left[\log p(X, \beta_1, \beta_2)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1, \beta_2)\right] \tag{64}$$

which mean, we would like to use probability distribution q parameterized by β_1 and β_2 to approximate the posterior distribution $p(\beta_1, \beta_2|X)$, which is intractable. As we already known, the posterior distribution can be written as

$$p(\beta_1, \beta_2|X) = p(\beta_1|\beta_2, X)p(\beta_2|X) \tag{65}$$
$$= p(\beta_2|\beta_1, X)p(\beta_1|X) \tag{66}$$

Now, assume those **complete conditional distributions** $p(\beta_1|\beta_2, X)$ and $p(\beta_2|\beta_1, X)$ are in exponential family distributions, we can express them into the canonical exponential family form, which are

$$p(\beta_1|\beta_2, X) = h(\beta_1)\exp\left\{\phi(\beta_1)^\top \eta + A(\eta)\right\} \tag{67}$$
$$p(\beta_2|\beta_1, X) = h(\beta_2)\exp\left\{\phi(\beta_2)^\top \eta + A(\eta)\right\} \tag{68}$$

since the distribution $p(\beta_1|\beta_2, X)$ contains the information from β_2 and X, we assume the natural parameters is a function of β_2 and X, it follows that

$$p(\beta_1|\beta_2, X) = h(\beta_1)\exp\left\{\phi(\beta_1)^\top \eta(\beta_2, X) + A(\eta(\beta_2, X))\right\} \tag{69}$$
$$p(\beta_2|\beta_1, X) = h(\beta_2)\exp\left\{\phi(\beta_2)^\top \eta(\beta_1, X) + A(\eta(\beta_1, X))\right\} \tag{70}$$

Now, in order to approximate the posterior distribution $p(\beta_1, \beta_2|X)$, we assume our suggested distributions are also in exponential family, and we also assume these two parameters are **independent**, and the mean-field family is fully factorized, and each are controlled by its own free parameters λ_1 and λ_2

$$p(\beta_1, \beta_2|X) \approx q(\beta_1, \beta_2) \tag{71}$$
$$= q(\beta_1) \cdot q(\beta_2) \tag{72}$$
$$= q(\beta_1|\lambda_1) \cdot q(\beta_2|\lambda_2) \tag{73}$$
$$= h(\beta_1)\exp\left\{\phi(\beta_1)^\top \lambda_1 - A(\lambda_1)\right\} \cdot h(\beta_2)\exp\left\{\phi(\beta_2)^\top \lambda_2 - A(\lambda_2)\right\} \tag{74}$$

We can see that our approximation q control the ELBO, and the distribution q is determined by the parameters λ_1, λ_2, thus we can view the ELBO as a function of parameters λ_1, λ_2

$$\mathcal{L}(\lambda_1, \lambda_2) = E_{q(\beta_1,\beta_2)}\left[\log p(X, \beta_1, \beta_2)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1, \beta_2)\right] \tag{75}$$

Now, we can use the coordinate ascent method to optimize the ELBO over the λ_1, λ_2 alternatively. We can write down the closed form solution of λ_1 and λ_2 to optimize the ELBO alternatively.

$$\mathcal{L}(\lambda_1, \lambda_2) = E_{q(\beta_1,\beta_2)}\left[\log p(X, \beta_1, \beta_2)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1, \beta_2)\right] \tag{76}$$
$$= E_{q(\beta_1,\beta_2)}\left[\log p(\beta_1|\beta_2, X) + \log p(\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_2)\right] \tag{77}$$

$$\arg\max_{\lambda_1} \mathcal{L}(\lambda_1, \lambda_2) \tag{78}$$
$$= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\log p(\beta_1|\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1)\right] \tag{79}$$
$$= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1)\exp\left\{\phi(\beta_1)^\top \eta(\beta_2, X) - A(\eta(\beta_2, X))\right\}\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1)\right] \tag{80}$$
$$= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1)\right] E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \eta(\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[A(\eta(\beta_2, X))\right] - E_{q(\beta_1,\beta_2)}\left[\log q(\beta_1|\lambda_1)\right] \tag{81}$$
$$= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1)\right] + E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \eta(\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[A(\eta(\beta_2, X))\right] \tag{82}$$
$$- E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1) \cdot \exp\left\{\phi(\beta_1)^\top \lambda_1 - A(\lambda_1)\right\}\right] \tag{83}$$
$$= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1)\right] + E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \eta(\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[A(\eta(\beta_2, X))\right] \tag{84}$$
$$- E_{q(\beta_1,\beta_2)}\left[\log h(\beta_1)\right] - E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \lambda_1\right] + E_{q(\beta_1,\beta_2)}\left[A(\lambda_1)\right] \tag{85}$$

Because we assume $q(\beta_1, \beta_2) = q(\beta_1) \cdot q(\beta_2)$, and by the fact that

$$E_{q(\beta_1,\beta_2)}\left[f(\beta_1)f(\beta_2)\right] = E_{q(\beta_1)}\left[f(\beta_1)\right] \cdot E_{q(\beta_2)}\left[f(\beta_2)\right] \tag{86}$$

we can rewrite the formula above as

$$\begin{aligned}
&\arg\max_{\lambda_1} \mathcal{L}(\lambda_1, \lambda_2) && (87)\\
&= \arg\max_{\lambda_1} E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \eta(\beta_2, X)\right] - E_{q(\beta_1,\beta_2)}\left[A(\eta(\beta_2, X))\right] && (88)\\
&\qquad - E_{q(\beta_1,\beta_2)}\left[\phi(\beta_1)^\top \lambda_1\right] + E_{q(\beta_1,\beta_2)}\left[A(\lambda_1)\right] && (89)\\
&= \arg\max_{\lambda_1} E_{q(\beta_1)q(\beta_2)}\left[\phi(\beta_1)^\top \eta(\beta_2, X)\right] - E_{q(\beta_1)q(\beta_2)}\left[A(\eta(\beta_2, X))\right] && (90)\\
&\qquad - E_{q(\beta_1)q(\beta_2)}\left[\phi(\beta_1)^\top \lambda_1\right] + E_{q(\beta_1)q(\beta_2)}\left[A(\lambda_1)\right] && (91)\\
&= \arg\max_{\lambda_1} E_{q(\beta_1)}\left[\phi(\beta_1)\right] \cdot E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - \underbrace{E_{q(\beta_2)}\left[A(\eta(\beta_2, X))\right]}_{\text{constant w.r.t } \lambda_1} && (92)\\
&\qquad - E_{q(\beta_1)}\left[\phi(\beta_1)^\top \lambda_1\right] + \underbrace{E_{q(\beta_1)}\left[A(\lambda_1)\right]}_{=\mathrm{A}(\lambda_1)} && (93)\\
&= \arg\max_{\lambda_1} E_{q(\beta_1)}\left[\phi(\beta_1)\right] \cdot E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - E_{q(\beta_1)}\left[\phi(\beta_1)^\top \lambda_1\right] + A(\lambda_1) && (94)
\end{aligned}$$

According to the property:

$$\frac{\partial A(\eta)}{\partial \eta} = E_{p(X|\eta)}\left[\phi(X)\right] \tag{95}$$

we have

$$\begin{aligned}
&\arg\max_{\lambda_1} \mathcal{L}(\lambda_1, \lambda_2) && (96)\\
&= \arg\max_{\lambda_1} E_{q(\beta_1)}\left[\phi(\beta_1)\right] \cdot E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - E_{q(\beta_1)}\left[\phi(\beta_1)^\top \lambda_1\right] + A(\lambda_1) && (97)\\
&= \arg\max_{\lambda_1} A'(\lambda_1)^\top E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - A(\lambda_1)^\top \lambda_1 + A(\lambda_1) && (98)
\end{aligned}$$

We can take the derivative w.r.t to λ_1 and set it to 0, we have

$$\begin{aligned}
\frac{\partial(A'(\lambda_1)^\top E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - A(\lambda_1)^\top \lambda_1 + A(\lambda_1))}{\partial \lambda_1} &= 0 && (99)\\
A''(\lambda_1)^\top E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - A'(\lambda_1)^\top \lambda_1 - A'(\lambda_1) + A'(\lambda_1) &= 0 && (100)\\
A''(\lambda_1)^\top E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - A'(\lambda_1)^\top \lambda_1 &= 0 && (101)\\
A''(\lambda_1)^\top \left(E_{q(\beta_2)}\left[\eta(\beta_2, X)\right] - \lambda_1\right) &= 0 && (102)
\end{aligned}$$

If $A''(\lambda_1) \neq 0$, then we get the closed form solution of λ_1

$$\lambda_1 = E_{q(\beta_2|\lambda_2)}\left[\eta(\beta_2, X)\right] \tag{103}$$

Respectively, we can also get the closed form solution for λ_2

$$\lambda_2 = E_{q(\beta_1|\lambda_1)}\left[\eta(\beta_1, X)\right] \tag{104}$$

So, initially we can set λ_1 to a random value, then we can calculate the $q(\beta_1|\lambda_1)$, and thus we can compute the optimal $\lambda_2 = E_{q(\beta_1|\lambda_1)}\left[\eta(\beta_1, X)\right]$, and continue to get the optimal λ_1 iteratively, until converged. In the end, we can get the optimal $q(\beta_1|\lambda_1) \cdot q(\beta_2|\lambda_2)$ to approximate the posterior distribution $p(\beta_1, \beta_2|X)$.

5 Exponential Family in Conditional Random Fields

As we already know in the unconditional model, [5]

$$\begin{aligned}
p(x|\theta) &= \exp(\phi(x)^\top \theta - g(\theta)) && (105)\\
g(\theta) &= \log \sum \exp(\phi(x)^\top \theta) && (106)\\
\frac{\partial g}{\partial \theta} &= E_{p(x|\theta)}(\phi(x)) && (107)
\end{aligned}$$

Note, that we omit the $h(x)$ for simple explanation; in the conditional model, we have some observed variables, thus

$$p(y|\theta, x) = \exp(\phi(x, y)^\top \theta - g(\theta|x)) \tag{108}$$

$$g(\theta|x) = \log \sum \exp(\phi(x, y)^\top \theta) \tag{109}$$

$$\frac{\partial g(\theta|x)}{\partial \theta} = E_{p(y|\theta,x)}(\phi(x, y)) \tag{110}$$

5.1 Estimation

References

[1] Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends® in Machine Learning 1.1–2 (2008): 1-305.

[2] The Exponential Family: Basics https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter8.pdf

[3] CMU Lecture https://www.youtube.com/watch?v=SWkhhn4s6Es

[4] Exponential family, Conjugacy, and Sufficiency https://www.cs.princeton.edu/~bee/courses/scribe/lec_09_02_2013.pdf

[5] Alex Smola's Lecture https://www.youtube.com/watch?v=Yq_CLto7IWY