

CS/CNS/EE/IDS 165: Foundations of Machine Learning

Sufficient Statistics

<http://tensorlab.cms.caltech.edu/users/anima/cms165-2020.html>

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena, CA 91125
anima@caltech.edu
Copyright ©2013

Outline

Concepts

- Parametric statistical model.
- Statistics, sufficient statistics, and minimal sufficient statistics.
- Exponential families.

References

1. H.V. Poor, [An Introduction to Signal Detection and Estimation](#), 2nd Ed., Springer-Verlag, 1994, Chapter IV.C.
2. L. L. Scharf, [Statistical Signal Processing: Detection, Estimation and Time Series Analysis](#), Addison-Wesley, Publishing Company, Inc., 1991, Chapter 3.
3. P.J. Bickel and K.A. Doksum, [Mathematical Statistics: Basic Ideas and Selected Topics](#), Prentice Hall, 1977, Chapter 2.
4. T. S. Ferguson, [Mathematical Statistics: A Decision Theoretic Approach](#), Academic Press, 1967, Chapter 3.3.
5. J. Shao, [Mathematical Statistics](#), Springer-Verlag, 1999, Chap. 2.

Motivating Examples

Coin Flip

The experiment of flipping a coin with probability of showing head θ can be modeled by pmfs indexed by θ

$$f(y|\theta) \triangleq \begin{cases} \theta & y = 1 \\ 1 - \theta & y = 0 \end{cases}, \quad \theta \in \Theta \triangleq [0, 1]$$

Binary signaling in Gaussian noise

The transmission of $\theta \in \{1, -1\}$ over an AWGN channel

$$Y = \theta + N, \quad N \sim \mathcal{N}(0, \sigma^2)$$

with known σ^2 can be modeled by pdfs indexed by $\theta \in \{\pm 1\}$

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \theta)^2}{2\sigma^2}\right\}, \quad \theta \in \Theta \triangleq \{\pm 1\}$$

Channel Estimation

An unknown linear fading channel in Gaussian noise

$$Y_1 = \theta s_1 + N_1, \quad Y_2 = \theta s_2 + N_2, \quad N_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

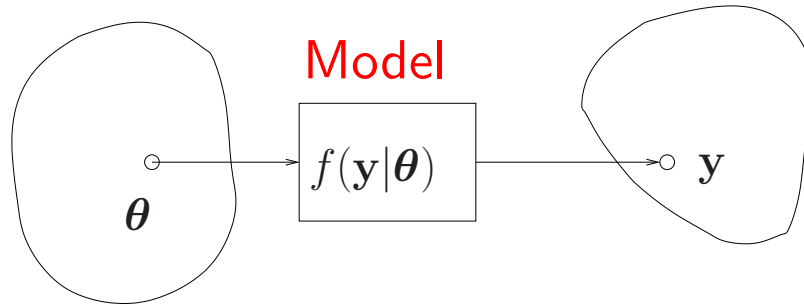
with known input s_1, s_2 and σ^2 can be modeled by

$$f(y_1, y_2|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_1 - s_1\theta)^2 + (y_2 - s_2\theta)^2}{2\sigma^2}\right\}, \quad \theta \in \Theta \triangleq \mathcal{R}$$

Parametric Model

Parameter Space Θ

Observation Space Γ



Frequentist Model

The statistic model is defined by the probability density (or pmf) function $f(\mathbf{y}|\theta)$ on the observation space Γ indexed by deterministic parameter $\theta \in \Theta$. Note that $f(\mathbf{y}|\theta)$ is not the conditional PDF (θ is deterministic); it is merely for notational convenience.

Bayesian Model

If the parameter can be modeled as random with **prior pdf** $\pi(\theta)$, we then have a Bayesian model

$$f(\mathbf{y}, \theta) = \pi(\theta)f(\mathbf{y}|\theta).$$

The **posterior** distribution of Θ given observation \mathbf{y} is

$$f(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y}|\theta)}{\int \pi(\mathbf{t})f(\mathbf{y}|\mathbf{t})d\mathbf{t}}$$

Statistics vs. Probability

In statistics, we are interested in inferring θ **after** observing $Y = \mathbf{y}$. In probability, we are interested in deducing the chance of various outcomes **without** experiments.

Frequentist vs. Bayesian

Frequentist Viewpoint

- Probability is objective; it is connected to the physical world through the relative frequency of event occurrence.
- Parameters are deterministic and unknown; it does not make sense to calculate $\Pr(\theta \in \mathcal{X} | \mathbf{Y} = \mathbf{y})$.
- Statistical procedures should have well-behaved long-run properties.

Bayesian Viewpoint

- Probability is subjective; it merely describes the degree of a belief. (“tomorrow, 30% chance of snow”).
- Even if θ is deterministic, we can assign certain distribution of prior belief.
- The inference of a parameter is made based on the posterior distribution $f(\theta | \mathbf{y})$.

Likelihood

Likelihood Function

Given the observation data $\mathbf{Y} = \mathbf{y}$, then the **likelihood** function of θ is a function of the form

$$l(\theta; \mathbf{y}) \triangleq \gamma(\mathbf{y})f(\mathbf{y}|\theta)$$

where $\gamma(\mathbf{y})$ does not depend on θ . A standard choice is when $\gamma(\mathbf{y}) = 1$.

- A likelihood function should be viewed as a function of parameter θ , and it is not uniquely defined.
- Sometimes, it is more convenient to work with log-likelihood function

$$L(\theta; \mathbf{y}) = \log f(\mathbf{y}|\theta).$$

- The average log-likelihood function happens to be the **entropy**:

$$H_{\theta}(\mathbf{Y}) \triangleq \mathbb{E}_{\theta}(-L(\theta; \mathbf{Y})) = - \int f(\mathbf{y}|\theta) \log f(\mathbf{y}|\theta) d\mathbf{y}$$

Note that the connection between entropy and likelihood function is only valid when the expectation is taken using the same probability model that the observations are generated.

Example: Uniform Distribution

Consider N independent random samples $Y_i \stackrel{i.i.d.}{\sim} \mathcal{U}(0, \theta)$. The parametric model is then given by the PDF

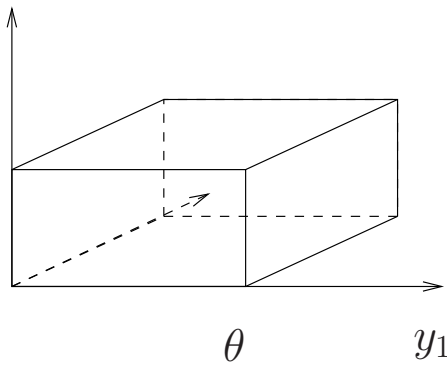
$$f(\mathbf{y}|\theta) = \begin{cases} \frac{1}{\theta^n} & \theta \geq \max\{y_i\} \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function $l(\theta; \mathbf{y})$ defined

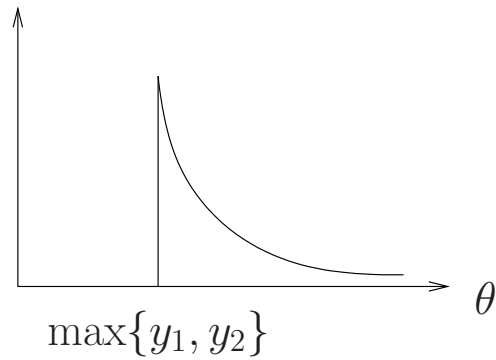
$$l(\theta; \mathbf{y}) \triangleq f(\mathbf{y}|\theta)$$

has a very different look from the PDF.

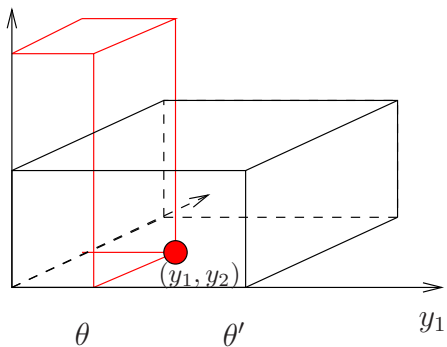
$f(y_1, y_2|\theta)$



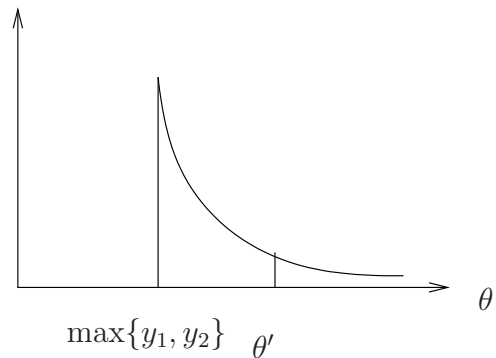
$l(\theta; y_1, y_2)$



$f(y_1, y_2|\theta)$



$l(\theta; y_1, y_2)$



Examples: The Gaussian Population

Independent Sampling

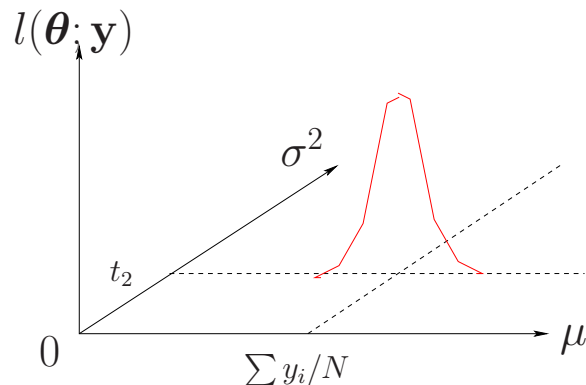
Consider N independent random samples $Y_i \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. With $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathcal{R} \times \mathcal{R}^+$, the parametric model is then given by

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp\left\{-\frac{\sum_{i=1}^N y_i^2 - 2\mu \sum_{i=1}^N y_i + N\mu^2}{2\sigma^2}\right\}$$

The likelihood function can be defined as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \exp\left\{-N \frac{\frac{1}{N} \sum_{i=1}^N y_i^2 - 2\mu \frac{1}{N} \sum_{i=1}^N y_i + \mu^2 + 2\sigma^2 \ln \sigma}{2\sigma^2}\right\}$$

$$L(\boldsymbol{\theta}; \mathbf{y}) = -N \frac{\frac{1}{N} \sum_{i=1}^N y_i^2 - 2\mu \frac{1}{N} \sum_{i=1}^N y_i + \mu^2 + 2\sigma^2 \ln \sigma}{2\sigma^2}$$



Remark

- The likelihood function depends only on data summary $(\sum_i y_i, \sum_i y_i^2)$.
- What happens when $N \rightarrow \infty$? By the law of large numbers, we have roughly

$$\frac{1}{N} L(\boldsymbol{\theta}; \mathbf{y}) \rightarrow -\frac{1 + 2 \ln \sigma^2}{2}$$

Example: Independent Bernoulli Trials

The Model

Suppose that we conduct N independent Bernoulli trials with probability of success $\Pr(Y_i = 1) = \theta$, $\Pr(Y_i = 0) = 1 - \theta$, and $\theta \in \{\theta_1, \theta_2\}$, and $\theta_1 \neq \theta_2$. The parametric model is then given by

$$f(\mathbf{y}|\theta) = \theta^{\sum y_i} (1 - \theta)^{N - \sum y_i}$$

Remarks

- Again, the model depends not on the entire \mathbf{y} but only on a single number $t(\mathbf{y}) = \sum_i y_i$ —the total number of successes in N trials, *i.e.*, the model can be written as

$$f(\mathbf{y}|\theta) = g(t(\mathbf{y}); \theta)$$

- A less obvious but more fundamental fact is that the model depends only on the **likelihood ratio**

$$r(\mathbf{y}) = \frac{f(\mathbf{y}|\theta_1)}{f(\mathbf{y}|\theta_2)} = \left(\frac{\theta_1}{\theta_2}\right)^{\sum_i y_i} \left(\frac{1 - \theta_1}{1 - \theta_2}\right)^{n - \sum_i y_i}$$

This follows from

$$r(\mathbf{y}) \rightarrow t(\mathbf{y}) \rightarrow f(\mathbf{y}|\theta) = q(r(\mathbf{y}); \theta)$$

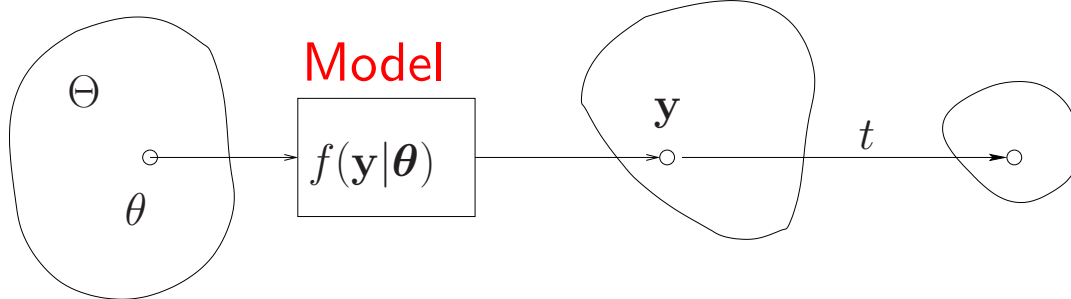
- If we can write $f(\mathbf{y}|\theta) = g(t(\mathbf{y}); \theta)$, can we discard \mathbf{y} using only $t(\mathbf{y})$?

Statistics

Given a parametric model $f(\mathbf{y}|\boldsymbol{\theta})$, a (measurable) function $t(\mathbf{Y})$ of the random observation $\mathbf{Y} \sim f(\mathbf{y}|\boldsymbol{\theta})$ is called a **statistic**.

Parameter Space

Observation Space



- A statistic is a random vector that conveys information about the original parametric model. It often has lower dimension than \mathbf{y} and less complex; it represents a (possibly lossy) data reduction.
- There are many statistics. The original observation \mathbf{Y} is a trivial statistic.
- Statistics are used for inference. It is therefore desirable that (i) they do not lose information about the model—**sufficiency** and (ii) their dimension is as low as possible—**parsimony**.

Sufficiency

A statistic $t(\mathbf{Y})$ is a **sufficient statistic** for model $f(\mathbf{y}|\theta)$ if the conditional density of r.v. \mathbf{Y} given $t(\mathbf{Y}) = \mathbf{u}$ is not a function of θ for all \mathbf{u} . A sufficient statistic $t(\mathbf{Y})$ is a **minimal sufficient** statistic if, for any other sufficient statistic \tilde{t} , there is a (measurable) function $h(\cdot)$ such that $t(\mathbf{y}) = h(\tilde{t}(\mathbf{y}))$.

Example

Consider n Bernoulli trials $Y_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{B}(\theta)$. Denote $\mathbf{Y} = (Y_1, \dots, Y_n)$. We claim that $t(\mathbf{Y}) = \sum Y_i$ is a sufficient statistic.

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y} | t(\mathbf{Y}) = j) &= \frac{\Pr(\mathbf{Y} = \mathbf{y}, t(\mathbf{Y}) = j)}{\Pr(t(\mathbf{Y}) = j)} \\ &= \begin{cases} \frac{\theta^j (1-\theta)^{n-j}}{\binom{n}{j} \theta^j (1-\theta)^{n-j}} & \text{if } t(\mathbf{y}) = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Remarks:

- If we know $t(\mathbf{y})$, then we can discard \mathbf{y} since, given $t(\mathbf{Y}) = t(\mathbf{y})$, the probability of \mathbf{Y} no longer depends on θ ; the outcome of $\mathbf{Y} = \mathbf{y}$ is no longer informative.
- How to find sufficient statistics?

The Neyman-Fisher Factorization Theorem

Theorem: A statistic $t(\mathbf{Y})$ is sufficient if and only if the pdf $f(\mathbf{y}|\boldsymbol{\theta})$ has the factorization

$$f(\mathbf{y}|\boldsymbol{\theta}) = g(t(\mathbf{y}); \boldsymbol{\theta})h(\mathbf{y})$$

where g and h are non-negative functions.

Proof for the discrete case: If $f(\mathbf{y}|\boldsymbol{\theta}) = g(t(\mathbf{y}); \boldsymbol{\theta})h(\mathbf{y})$, then

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y} | t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta}) &= \frac{\Pr(\mathbf{Y} = \mathbf{y}, t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta})}{\Pr(t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta})} \\ &= \begin{cases} \frac{g(\mathbf{u}, \boldsymbol{\theta})h(\mathbf{y})}{\Pr(t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta})} & \text{if } t(\mathbf{y}) = \mathbf{u} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

But

$$\Pr(t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta}) = \sum_{\mathbf{y}, t(\mathbf{Y}) = \mathbf{u}} f(\mathbf{y}|\boldsymbol{\theta}) = g(\mathbf{u}; \boldsymbol{\theta}) \sum_{\mathbf{y}, t(\mathbf{y}) = \mathbf{u}} h(\mathbf{y})$$

Hence

$$\Pr(\mathbf{Y} = \mathbf{y} | t(\mathbf{Y}) = \mathbf{u}; \boldsymbol{\theta}) = \begin{cases} \frac{h(\mathbf{y})}{\sum_{\mathbf{y}, t(\mathbf{y}) = \mathbf{u}} h(\mathbf{y})} & \text{if } t(\mathbf{y}) = \mathbf{u} \\ 0 & \text{otherwise} \end{cases}$$

If $t(\mathbf{Y})$ is sufficient, let

$$\begin{aligned} g(t(\mathbf{y}); \boldsymbol{\theta}) &\triangleq \Pr(t(\mathbf{Y}) = t(\mathbf{y}); \boldsymbol{\theta}), \\ h(\mathbf{y}) &= \Pr(\mathbf{Y} = \mathbf{y} | t(\mathbf{Y}) = t(\mathbf{y})) \end{aligned}$$

Then

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta}) &= \Pr(\mathbf{Y} = \mathbf{y}; \boldsymbol{\theta}) = \Pr(\mathbf{Y} = \mathbf{y}, t(\mathbf{Y}) = t(\mathbf{y}); \boldsymbol{\theta}) \\ &= g(t(\mathbf{y}); \boldsymbol{\theta})h(\mathbf{y}) \end{aligned}$$

Sufficiency of Likelihood

Corollary Consider a binary hypothesis model given by $\mathbf{y} \sim p(\mathbf{y}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$. Define the statistic by the likelihood ratio

$$r(\mathbf{Y}) \triangleq \frac{f(\mathbf{Y}|\boldsymbol{\theta}_1)}{f(\mathbf{Y}|\boldsymbol{\theta}_0)}.$$

We then have $p(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}_0)g(r(\mathbf{y}); \boldsymbol{\theta})$, where

$$g(r(\mathbf{y}); \boldsymbol{\theta}) = \begin{cases} 1 & \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ r(\mathbf{y}) & \boldsymbol{\theta} = \boldsymbol{\theta}_1 \end{cases}$$

By the Neyman-Fisher factorization, $r(\mathbf{Y})$ is a sufficient statistic.

Remarks:

- For the general discrete model $\Theta = \{\theta_1, \dots, \theta_M\}$, the M -dimensional vector of likelihood functions $l(\mathbf{y}) = [p(\mathbf{y}; \theta_1), \dots, p(\mathbf{y}; \theta_M)]$ or the $M - 1$ dimensional vectors of likelihood ratios

$$r(\mathbf{Y}) = \left[\frac{f(\mathbf{Y}|\theta_2)}{f(\mathbf{Y}|\theta_1)}, \dots, \frac{f(\mathbf{Y}|\theta_M)}{f(\mathbf{Y}|\theta_1)} \right]$$

are also sufficient statistics.

- If we broaden the notion of statistic whose values are functions of θ , the the likelihood function $r(\boldsymbol{\theta}; \mathbf{Y})$ is minimal sufficient (Dynkin,1951)[†].

[†]E.B. Dynkin, "Necessary and sufficient statistics for families of distributions," *Sel. Transl. Math., Stat., and Prob.*, vol. 1, pp. 23–41, 1951.

Examples

I.I.D. Gaussian Model: Consider $Y_i \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

$$pf(\mathbf{y}|\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma^2)^n} e^{\{-\frac{\sum y_i^2}{2\sigma^2} + \frac{\mu \sum y_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\}} \rightarrow \mathbf{t}(\mathbf{y}) = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i^2 \end{pmatrix}.$$

I.I.D. Poisson Model: Consider $Y_i \stackrel{i.i.d.}{\sim} \mathcal{P}(\lambda)$.

$$f(\mathbf{y}|\lambda) = \frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod y_i!} \rightarrow t(\mathbf{y}) = \sum_i y_i.$$

Extreme Statistic. Suppose $Y_i \stackrel{i.i.d.}{\sim} \mathcal{U}(0, \theta)$.

$$f(\mathbf{y}|\theta) = \begin{cases} \frac{1}{\theta^{-n}} & 0 < y_i < \theta, \quad \forall i \\ 0 & \text{otherwise} \end{cases} = h(\mathbf{y})g(\theta, \max_i y_i)$$
$$h(\mathbf{y}) = \begin{cases} 1 & y_i > 0, \quad \forall i \\ 0 & \text{otherwise} \end{cases} \quad g(\theta, t) = \begin{cases} \frac{1}{\theta^{-n}} & t < \theta, \\ 0 & \text{otherwise} \end{cases}$$

Channel Estimation in AWGN. Given

$$y_n = x_0 s_n + x_1 s_{n-1} + w_n, \quad w_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad n = 0, 1, \dots, N-1,$$

To estimate $\boldsymbol{\theta} = [x_0 \ x_1]^T$ with known s_n , let

$$\mathbf{y} = \begin{pmatrix} y_0 \\ \vdots \\ y_{N-1} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} s_0 & s_{-1} \\ s_1 & s_0 \\ \vdots & \vdots \\ s_{N-1} & s_{N-2} \end{pmatrix}.$$

Then

$$f(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{S}\boldsymbol{\theta}\|^2}{2\sigma^2}\right\}$$
$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{\frac{2\mathbf{y}'\mathbf{S}\boldsymbol{\theta} - \|\mathbf{S}\boldsymbol{\theta}\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}\right\}$$

Hence

$$t(\mathbf{y}) = \mathbf{y}'\mathbf{S} = \begin{pmatrix} \sum_i s_i y_i \\ \sum_i s_{i-1} y_i \end{pmatrix}$$

The K-Parameter Exponential Family

Definition: A family of distributions is said to be a K -parameter exponential family if there exist $c_1(\boldsymbol{\theta}), \dots, c_K(\boldsymbol{\theta}), d(\boldsymbol{\theta}), t_1(\mathbf{y}), \dots, t_K(\mathbf{y}), s(\mathbf{y})$ and a set \mathcal{A} such that

$$f(\mathbf{y}|\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^K c_i(\boldsymbol{\theta})t_i(\mathbf{y}) + d(\boldsymbol{\theta}) + s(\mathbf{y})\right\}1_{\mathcal{A}}(\mathbf{y})$$

where $1_{\mathcal{A}}(\mathbf{y})$ is the indicator function not related to $\boldsymbol{\theta}$. It is often more convenient to use the **canonical form** (or the natural representation) of the exponential distribution

$$f(\mathbf{y}|\boldsymbol{\eta}) = \exp\left\{\sum_{i=1}^K \eta_i t_i(\mathbf{y}) + d(\boldsymbol{\eta}) + s(\mathbf{y})\right\}1_{\mathcal{A}}(\mathbf{y}).$$

Theorem: Let $\{f(\mathbf{y}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Lambda\}$ be a K -parameter exponential family, *i.e.*,

$$f(\mathbf{y}|\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^K c_i(\boldsymbol{\theta})t_i(\mathbf{y}) + d(\boldsymbol{\theta}) + s(\mathbf{y})\right\}I_{\mathcal{A}}(\mathbf{y})$$

If $\{\mathbf{c}(\boldsymbol{\theta}) = [c_1(\boldsymbol{\theta}), \dots, c_K(\boldsymbol{\theta})], \boldsymbol{\theta} \in \Theta\}$ has an interior point, then $\mathbf{t}(\mathbf{y}) = [t_1(\mathbf{y}), \dots, t_K(\mathbf{y})]^T$ is minimal sufficient.

Proof: The sufficiency of $\mathbf{t}(\mathbf{y})$ follows the Neyman-Fisher factorization. The minimality is implied by the completeness of $\mathbf{t}(\mathbf{y})$, which will be discussed later. The reason for the existence of “interior point” is to prevent the trivial cases such as by splitting $c_1(\boldsymbol{\theta}) = c_{11}(\boldsymbol{\theta}) + c_{12}(\boldsymbol{\theta})$ thus increasing the dimension of the statistic.

Examples of Exponential Family

These belong to the exponential family

1. Gaussian. $Y_i \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

2. Binomial: $Y \sim \mathcal{B}(\theta, n)$

$$f(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{(n-k)} = \binom{n}{k} e^{k \ln \frac{\theta}{1-\theta} + n \ln(1-\theta)}$$

3. Multinomial: In n independent trials with s outcomes in each trial. Let p_i be the probability for the i th outcome. Let y_i be the number of trials that have the i th outcome.

$$\begin{aligned} f(y_1, \dots, y_s | p_1, \dots, p_s) &= \frac{n!}{y_1! \dots y_s!} p_1^{y_1} \dots p_s^{y_s} \\ &= \exp(k_1 \ln p_1 + \dots + k_s \ln p_s) h(\mathbf{y}) I_{\mathcal{A}}(\mathbf{y}) \end{aligned}$$

4. Poisson. $Y_i \stackrel{i.i.d.}{\sim} \mathcal{P}(\theta)$

$$f(y_1, \dots, y_n | \theta) = \frac{\theta^{\sum y_i}}{\prod y_i!} e^{-n\theta} = \exp\left\{ \sum y_i \ln \theta - n\theta \right\} h(\mathbf{y})$$

These do not belong to the exponential family

1. Uniform. $Y \sim \mathcal{U}(0, \theta)$.

$$f(y|\theta) = \frac{1}{\theta} I_{(0,\theta)}(y)$$

2.

$$f(y|\theta) = 2 \frac{y + \theta}{1 + 2\theta} = \exp\{\ln 2(y + \theta) - \ln(1 + 2\theta)\}, \quad 0 < y < 1, \quad \theta > 0$$