Multivariate Normal Distribution as an Exponential Family distribution

In exponential family, the underlying pdf or pmf is

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c(\boldsymbol{\theta}) \cdot e^{\sum_{j=1}^{k} \theta_j t_j(\mathbf{x})} \cdot h(\mathbf{x})$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$ is the *canonical parameter*. Then, based on the i.i.d. sample $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, the *canonical sufficient statistic* is

$$\mathbf{t}(\mathbf{X}) = (\sum_{i=1}^{n} t_1(\mathbf{X}_i), \ldots, \sum_{i=1}^{n} t_k(\mathbf{X}_i)) := (t_1(\mathbf{X}), \ldots, t_k(\mathbf{X})),$$

which is also complete (if Θ contains k-dimensional parallelepiped), and therefore it is a minimal sufficient statistic. Then we say that the exponential family is *minimally represented* (which also means that it is not overparametrized). Usually, Θ is an open set, in which case, the exponential family is called *regular*.

Proposition 1 *Under the usual regularity conditions (see the Cramér–Rao inequality and the Cramér–Dugué theorem), in regular exponential families the likelihood equation boils down to solving*

$$\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t}(\mathbf{X})) = \mathbf{t}(\mathbf{X}).$$

Proof. The likelihood function has the following form:

$$L_\theta(\mathbf{X}) = c^n(\boldsymbol{\theta}) \cdot e^{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} t_j(\mathbf{X}_i)} \cdot \prod_{i=1}^{n} h(\mathbf{X}_i) = \frac{1}{a(\boldsymbol{\theta})} \cdot e^{\langle \boldsymbol{\theta}, t(\mathbf{X}) \rangle} \cdot b(\mathbf{X}),$$

where

$$a(\boldsymbol{\theta}) = \int_{\mathcal{X}} e^{\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) \rangle} \cdot b(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

is the normalizing constant, while $\mathcal{X} \subset \mathbb{R}^n$ is the sample space. This formula will play a crucial role in our subsequent calculations.

The likelihood equation is

$$\nabla_{\boldsymbol{\theta}} \ln L_{\boldsymbol{\theta}}(\mathbf{X}) = \mathbf{0},$$

that is

$$-\nabla_{\boldsymbol{\theta}} \ln a(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \langle \mathbf{t}(\mathbf{X}), \boldsymbol{\theta} \rangle \;=\; \mathbf{0}. \tag{2}$$

Under the regularity conditions, by (1) we get that

$$\nabla_{\boldsymbol{\theta}} \ln a(\boldsymbol{\theta}) = \frac{1}{a(\boldsymbol{\theta})} \int_{\mathcal{X}} \mathbf{t}(\mathbf{x}) e^{\langle \mathbf{t}(\mathbf{x}), \boldsymbol{\theta} \rangle} \cdot b(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t}(\mathbf{X})).$$

Therefore, (2) is equivalent to

$$-\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{t}(\mathbf{X})) + \mathbf{t}(\mathbf{X}) = \mathbf{0}$$

that finishes the proof. □

Note that this resembles the idea of the moment estimation. Indeed, if $t_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$, …, $t_k(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, then the ML estimator of the

canonical parameter is the same as the moment estimator. This is the case, e.g., when our underlying distribution is Poisson, exponential, or Gaussian.

Observe that with introducing the so-called *log-partition (in other words, cumulant) function* $Z(\boldsymbol{\theta}) := \ln a(\boldsymbol{\theta})$, the likelihood function has the form

$$L_{\boldsymbol{\theta}}(\mathbf{X}) = e^{\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{X})\rangle - Z(\boldsymbol{\theta})} \cdot h(\mathbf{X}). \tag{3}$$

Based on Proposition 1, in *regular exponential families*, the ML equation $\nabla_{\boldsymbol{\theta}} \ln L_{\boldsymbol{\theta}}(\mathbf{X}) = \mathbf{0}$ is also equivalent to

$$\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) = \mathbf{t}. \tag{4}$$

Since $\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \mathbf{t}$, the ML equation (4) means that the canonical sufficient statistic is made equal to its expectation. But when is it possible? Now we briefly summarize existing theoretical results on this issue.

Let $\mathcal{M} = \{\mathbb{E}_{\boldsymbol{\theta}} \mathbf{t} \,:\, \boldsymbol{\theta} \in \Theta\}$ denote the so-called *mean parameter space*; it is necessarily convex. Let $\mathcal{M}^0$ denote its interior.

Proposition 2 *In exponential family, the gradient mapping $\nabla Z : \Theta \to \mathcal{M}$ is one-to-one if and only if the exponential family representation is minimal.*

Proposition 3 *In a minimally represented exponential family, the gradient mapping ∇Z is onto $\mathcal{M}^0$.*

By Propositions 2 and 3, any parameter in $\mathcal{M}^0$ is uniquely realized by the $\mathbb{P}_\theta$ distribution for some $\theta \in \Theta$. Also, in a regular and minimal exponential family, $\mathcal{M}$ is an open set and is identical to $\mathcal{M}^0$.

As the ML estimate of $\boldsymbol{\theta}$ is the solution of (4), we have the following.

Proposition 4 *Assume, the (canonical) parameter space Θ is open (i.e., we are in a regular exponential family). Then there exists a solution $\hat{\boldsymbol{\theta}} \in \Theta$ to the ML equation $\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) = \mathbf{t}$ if and only if $\mathbf{t} \in \mathcal{M}^0$; further, if such a solution exists, it is also unique.*

Note that in regular and minimal exponential families, $\mathcal{M}^0$ is also the interior of $\mathcal{T}$, which is the convex hull of all possible values of $\mathbf{t}$. $\mathcal{T}$ is usually not open, and its boundary may have positive probability (in particular, when the underlying distribution is discrete). If we unfortunately start with a sampling statistic $\mathbf{t}$ on this boundary, then we have no solution to the ML equation.

Example 1 (discrete distribution)
Let us apply the above theory for the $\mathcal{P}(\lambda)$ Poisson distribution. Its pmf is

$$p(x) = e^{-\lambda} e^{x \ln \lambda} \frac{1}{x!} = e^{\theta x - e^\theta} \frac{1}{x!}, \quad x \in \mathcal{X} = \{0.1.2, \dots\}.$$

From here we can see that the canonical parameter is $\theta = \ln \lambda$, and the canonical sufficient statistic based on an n-element sample $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$ is $\sum_{i=1}^n X_i$. Based on it, the likelihood equation boils down to

$$\mathbb{E}(\sum_{i=1}^n X_i) = n\lambda = \sum_{i=1}^n X_i.$$

The mean parameter is $\mathbb{E}(X) = \lambda$.

There is indeed a one-to-one correspondence between the canonical parameter-space $\mathbb{R}$ and the mean parameter-space $(0, \infty)$. They are open sets, so $\mathcal{M}^0 = (0, \infty)$. By proposition 4, the likelihood equation has a (unique) solution if and only if $\sum_{i=1}^n X_i \in (0, \infty)$. But here $\mathcal{T} = [0, \infty)$, the interior of which is indeed $\mathcal{M}^0$. Therefore, if $\sum_{i=1}^n X_i$ is on the boundarry of $\mathcal{T}$, i.e., if $\sum_{i=1}^n X_i = 0$, then the likelihood equation has no solution. This happens with positive probability $e^{-n\lambda}$, albeit this probability exponentially decreases with increasing n. Otherwise, the unique solution of the ML equation is $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$ by the above theory, no further considerations are needed.

Example 2 (absolutely continuous distribution)
Let us apply the above theory for the $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{C})$ distribution, where the dimension (the positive integer p) is fixed. The usual parameters $\boldsymbol{\mu}, \mathbf{C}$ are, in fact, equivalent to the mean value parameters here. Let us find the canonical parameters, and based on them, the ML estimators of $\boldsymbol{\mu}$ and $\mathbf{C}$. Assume that $\mathbf{C} > 0$. Introduce the following notation:

$$\mathbf{K} = \mathbf{C}^{-1}, \quad \mathbf{h} = \mathbf{K}\boldsymbol{\mu}.$$

$\mathbf{K}$ is called *concentration matrix*.

Proposition 5 $\mathbf{K}$ *and* $\mathbf{h}$ *are canonical parameters of the p-variate normal distribution.*

Proof. For $\mathbf{x} \in \mathbb{R}^p$ we have

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})} = \frac{|\mathbf{K}|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}(\mathbf{x}-\boldsymbol{\mu})}$$
$$= \frac{|\mathbf{K}|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{K}\boldsymbol{\mu}} e^{-\frac{1}{2}\mathrm{tr}[(\mathbf{x}\mathbf{x}^T)\mathbf{K}]+\mathbf{x}^T\mathbf{h}} = \frac{|\mathbf{K}|^{1/2}}{(2\pi)^{p/2}} e^{-\frac{1}{2}\mathbf{h}^T\mathbf{K}^{-1}\mathbf{h}} e^{-\frac{1}{2}\mathrm{tr}[(\mathbf{x}\mathbf{x}^T)\mathbf{K}]+\mathbf{x}^T\mathbf{h}}$$

that finishes the proof. □

It also turns out from the proof that, based on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the canonical sufficient statistics are

$$\sum_{i=1}^n \mathbf{X}_i \quad \text{and} \quad \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T.$$

Therefore, by Proposition 1, the ML equations boil down to

$$\mathbb{E}\left[\sum_{i=1}^n \mathbf{X}_i\right] = \sum_{i=1}^n \mathbf{X}_i \quad \text{and} \quad \mathbb{E}\left[\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\right] = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T.$$

Equivalently,

$$n\boldsymbol{\mu} = \sum_{i=1}^n \mathbf{X}_i \quad \text{and} \quad n\mathbf{C} + n\boldsymbol{\mu}\boldsymbol{\mu}^T = \mathbf{S} + n\bar{\mathbf{X}}\bar{X}^T,$$

from where

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} \quad \text{and} \quad \hat{\mathbf{C}} = \frac{1}{n}\mathbf{S}$$

easily follows.

It can be proven that the sufficient statistic (only the second moment ones are critical) is on the boundary of $\mathcal{T}$ if and only if $\mathbf{S}$ is singular. It is always true if $n \leq p$, but in the $n > p$ case we learned that $\mathbf{S} > 0$ with probability 1, so it is on the boundary of $\mathcal{T}$ with 0 probability. Consequently, provided $n > p$ holds, the ML equation has a solution with probability 1.