Note on Generalized Linear Models

Consider the following model for data $\boldsymbol{y} \in \mathbb{R}^n$:

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} \;+\; \boldsymbol{w}\,,$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a known matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. In other words, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I})$. In signal processing terms, the data $\boldsymbol{y}$ are generated by an unknown signal $\boldsymbol{X}\boldsymbol{\beta}$ (i.e., belonging to the subspace spanned by the columns of $\boldsymbol{X}$) plus Gaussian noise $\boldsymbol{w}$. Assuming that $\boldsymbol{X}$ has full rank p, the maximum likelihood estimate (also MVUB estimator) is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\,.$$

It is possible to consider other noise models. For example, suppose that $\boldsymbol{w}$ is a vector with each entry i.i.d. $p(w) = \frac{1}{2}e^{-|w|}$, for $w \in \mathbb{R}$. This is a double exponential distribution. The resulting log likelihood function of $\boldsymbol{y}$ is proportional to $\sum_{i=1}^n |y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|$, where $\boldsymbol{x}_i^T$ is the i-th row of $\boldsymbol{X}$. Note that $\sum_{i=1}^n |y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}| = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_1$, so using the double-exponential distribution to model the noise leads to the estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_1\,.$$

There is no closed-form linear algebraic solution to this optimization, but it is a convex optimization that is easy to solve numerically. The ℓ_1 minimization is less sensitive to large differences between $\boldsymbol{y}$ and $\boldsymbol{X}\boldsymbol{\beta}$ compared to the least squares solution. This is not surprising, since the double exponential noise model has heavier-than-Gaussian tails and thus probably generates more extremely large errors. The ℓ_2 minimization of least squares could be dominated by large errors, whereas the ℓ_1 minimization is less influenced by these errors (i.e., doesn't overfit to these large errors).

The Gaussian and double exponential cases above are both examples of an *additive* noise model. This sort of model isn't always the most appropriate way to model randomness present in our data. For example, if the data are counts or binary-valued, then Poisson and Binomial models are more natural than additive noise models. Generalized linear models are a well developed framework that extend this linear modeling approach to other probability distributions and noise/error models. The basic idea is to consider other probability models (e.g., Poisson, Exponential, Binomial, etc.) and parameterize the mean in terms of a linear model like $\boldsymbol{X}\boldsymbol{\beta}$. Specifically, we will consider models of the form $\boldsymbol{y} \sim \prod_{i=1}^n p(y_i|\theta_i)$ and each parameter $\theta_i = g(\boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\boldsymbol{x}_i^T$ is the i-th row of $\boldsymbol{X}$ and g is a known scalar function that is suited to the particular form of the distribution (more on this later in the note). The Gaussian model above fits this framework with $g(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, the identity.

1 The Exponential Family of Distributions

The **Exponential Family** is a class of distributions with the following form:

$$p(y|\theta) \;=\; b(y)\exp(\theta^T T(y) - a(\theta))\,.$$

The parameter θ is called the **natural parameter** of the distribution and $T(y)$ is the **sufficient statistic**. In many cases, $T(y) = y$ and then the distribution is said to be in **canonical form** and θ is called the **canonical parameter.** The quantity $e^{-a(\theta)}$ is a normalization constant, ensuring that $p(y|\theta)$ sums or integrates to

1. The factor $b(y)$ is the non-negative **base measure**, and in many cases it is equal to 1. Many familiar distributions belong to the exponential family (e.g., Gaussian, exponential, log-normal, gamma, chi-squared, beta, Dirichlet, Bernoulli, Poisson, geometric).

In general, the parameter θ is not the mean of the distribution. We can view θ as a function of the mean $\mu = \mathbb{E}y$, and write $\theta(\mu)$. To illustrate this idea, let us consider the following examples.

Example 1. *The Bernoulli distribution is written in terms of its mean* $0 \leq \mu \leq 1$ *as*

$$\begin{aligned} p(y|\mu) &= \mu^y (1-\mu)^{1-y} \\ &= \exp(y \log \mu + (1-y)\log(1-\mu)) \\ &= \exp\left(\log\left(\frac{\mu}{1-\mu}\right) y + \log(1-\mu)\right) . \end{aligned}$$

Thus, the natural parameter is $\theta = \log\left(\frac{\mu}{1-\mu}\right)$. *Conversely, we can write* μ *in terms of* θ *as* $\mu = \frac{1}{1+e^{-\theta}}$.

Example 2. *The Gaussian distribution is*

$$\begin{aligned} p(y|\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{(y-\mu)^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{y^2}{2}\right) \exp\left(\mu y - \frac{\mu^2}{2}\right) . \end{aligned}$$

Thus, the natural parameter is $\theta = \mu$.

The function that maps the mean μ to θ is denoted by g and is called the **link function.** Its inverse is called the **response function**. In other words, $\theta = g(\mu)$ and $\mu = g^{-1}(\theta)$.

2 Generalized Linear Modeling

Assume that $\boldsymbol{y} \sim \prod_{i=1}^n p(y_i|\theta_i)$, where $p(y_i|\theta_i)$ is in the Exponential Family and θ_i is the natural parameter of the distribution. Let $\boldsymbol{\theta} = [\theta_1\, \theta_2\, \ldots\, \theta_n]^T$. The key idea of the **Generalized Linear Model** (GLM) is to assume that the canonical parameters are described by the linear model $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{X}$ is a known $n \times p$ matrix and $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown. In other words, $\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. This model represents linear relationships between the elements of $\boldsymbol{\theta}$.

Now assume that the distribution is in canonical form; i.e., $T(y_i) = y_i$. Then note that the log likelihood is

$$\log \prod_{i=1}^n p(y_i|\theta_i) = \sum_{i=1}^n \left(\boldsymbol{\beta}^T \boldsymbol{x}_i y_i - a(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right) + \log b(y_i) .$$

Thus, just as in the Gaussian linear model we started with at the beginning of the note, the sufficient statistic $\boldsymbol{X}^T \boldsymbol{y}$ summarizes all our information about $\boldsymbol{\beta}$. This is the reason for the name GLM. The mean parameters can be obtained using the response function: $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$.

Example 3. *Consider the GLM for independent Bernoulli observations* $y_i \sim$ *Bernoulli*(μ_i), $i = 1, \ldots, n$. *Recall that the natural parameter is* $\theta = \log\left(\frac{\mu}{1-\mu}\right)$. *Conversely, we can write mean* μ *in terms of* θ *as* $\mu = \frac{1}{1+e^{-\theta}}$. *In other words, the response function* $g^{-1}(\theta) = \frac{1}{1+e^{-\theta}}$, *which is usually called the* **logistic**

function. *Note that this function maps the real line smoothly into the interval* $[0, 1]$. *It has an"S" shaped* sigmoid curve. *The log likelihood is*

$$L(\boldsymbol{\theta}) = \sum_{i=1} \theta_i y_i + \log\left(\frac{e^{-\theta_i}}{1+e^{-\theta_i}}\right) .$$

Now we can substitute the linear model $\theta_i = \boldsymbol{\beta}^t \boldsymbol{x}_i$ *to express the likelihood as a function of* $\boldsymbol{\beta}$*:*

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1} \boldsymbol{\beta}^T \boldsymbol{x}_i y_i + \log\left(\frac{e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i}}\right) \\ &= \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} + \sum_{i=1} \log\left(\frac{e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i}}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i}}\right) . \end{aligned}$$

So we see that the statistic $\boldsymbol{X}^T \boldsymbol{y}$ *is sufficent for* $\boldsymbol{\beta}$*, just as it is for the Gaussian linear model we looked at first in this note. Alternatively, the log likelihood can be written in terms of the mean* $\mu_i = \frac{1}{1+e^{-\theta_i}}$*:*

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{i=1} \theta_i y_i + \log\left(\frac{e^{-\theta_i}}{1+e^{-\theta_i}}\right) \\ &= \sum_{i=1} y_i \log \mu_i + (1-y_i)\log(1-\mu_i) . \end{aligned}$$

Thus, if $y_i = 1$*, then the corresponding term is given by* $\log\left(\frac{1}{1+e^{-\theta_i}}\right)$. *If* $y_i = 0$*, then term becomes* $\log\left(\frac{1}{1+e^{\theta_i}}\right)$. *So, we can write the log likelihood as*

$$L(\boldsymbol{\beta}) = \sum_{i=1} \log\left(\frac{1}{1+e^{-\theta_i z_i}}\right) ,$$

where $z_i = 2y_i - 1$. *Now we can substitute the linear model* $\theta_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ *to express the likelihood as a function of* $\boldsymbol{\beta}$*:*

$$L(\boldsymbol{\theta}) = \sum_{i=1} \log\left(\frac{1}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i z_i}}\right) ,$$

Maximizing this function (or equivalently minimizing its negation) with respect to $\boldsymbol{\beta}$ *is called* **logistic regression**. *The function* $-\log\left(\frac{1}{1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i z_i}}\right) = \log(1+e^{-\boldsymbol{\beta}^T \boldsymbol{x}_i z_i})$ *is called the* **logistic loss**. *The solution to this optimization does not have a simple linear algebraic form, but is easy to compute numerically.*

Example 4. *Consider the GLM for independent Gaussian observations* $y_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, n$. *Recall* *that the natural parameter is $\theta_i = \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. The log likelihood is*

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_{i=1} \left(\boldsymbol{\beta}^T \boldsymbol{x}_i y - \frac{(\boldsymbol{\beta}^T \boldsymbol{x}_i)^2}{2} + \frac{y_i^2}{2} - \frac{1}{2} \log(2\pi) \right) \\
&= \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} - \frac{\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}}{2} + \sum_{i=1}^{n} \frac{y_i^2}{2} - \frac{n}{2} \log(2\pi) \, .
\end{aligned}
$$

Maximizing this function with respect to $\boldsymbol{\beta}$ is called **linear regression***. This optimization is easy to solve by simply setting $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, which yields the equation $\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}$, resulting in the least squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ that opened our discussion at the beginning of the note.*