18.650. Statistics for Applications
Fall 2016. Problem Set 11

Due Friday, Dec. 9 at 12 noon

NOTE: there was a typo in the definition of the logistic function in Problem 3, Question 4.

Problem 1 Exponential families

For each of the following families of distributions, tell whether it is an exponential family:

- $\mathsf{Ber}(p), p \in (0,1)$;
- $\mathcal{N}(\mu, 1), \mu \in \mathbb{R}$;
- $\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$;
- $\mathsf{Exp}(\lambda), \lambda > 0$;
- $\mathcal{U}([0, \vartheta]), \vartheta > 0$;
- $\Gamma(\alpha, \beta), \alpha > 0, \beta > 0$;
- $\mathsf{Poiss}(\lambda), \lambda > 0$.

Recall that the Gamma distribution with parameters $\alpha, \beta > 0$ has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

where Γ is the Gamma function.

Problem 2 One parameter canonical Exponential families

Let $\Theta \subseteq \mathbb{R}$ and consider the family of densities $f_\theta, \theta \in \Theta$ defined on a subset $\mathcal{X}$ of $\mathbb{R}$ (called *sample space*) by

$$f_\theta(x) = a(x) \exp(-\theta x + b(\theta)), \quad x \in \mathcal{X},$$

where a is a positive function defined on $\mathcal{X}$ and b is a function defined on the space Θ (called *parameter space*).

Let $\ell(\theta) = \ln f_\theta(X), \theta \in \Theta$, be the log-likelihood function, where X is a random variable on $\mathcal{X}$.

1. In the rest of the problem, if $\theta \in \Theta$ and g is a function defined on $\mathbb{R}$, denote by $\mathbb{E}_\theta[g(X)]$ (resp. $\mathsf{Var}_\theta[g(X)]$) the expectation (resp. variance) of $g(X)$ under the assumption that X has density f_θ.
   a) Why is it true that
   $$\int_{\mathbb{R}} f_\theta(x)\,\mathrm{d}x = 1, \quad \forall \theta \in \Theta \quad ?$$
   b) Assuming that you can switch expectations and derivatives with respect to θ, prove the following identities:
      1. $\mathbb{E}_\theta\left[\ell'(\theta)\right] = 0$;
      2. $\mathsf{Var}_\theta\left[\ell'(\theta)\right] = -\mathbb{E}_\theta\left[\ell''(\theta)\right]$.

   c) What is the name of the last quantity in the previous question ?
2. Compute $\ell'(\theta)$ and $\ell''(\theta)$ in terms of the functions a and b.
3. Using the previous functions, compute $\mathbb{E}_\theta[X]$ and $\mathbf{Var}_\theta[X]$, for all $\theta \in \Theta$.
4. Example: Assume that $\Theta = (0, \infty)$ and for $\theta > 0$, f_θ is the density of the Gamma distribution with parameters α and θ, where α is a fixed number.
   a) What is the sample space $\mathcal{X}$?
   b) What are the functions a and b ?
   c) Using the previous questions, compute the expectation and the variance of the Gamma distribution with parameters $\alpha, \beta > 0$.

Problem 3 Linear model with latent variables

Consider the linear regression

$$Y = X'\beta + \varepsilon,$$

where $\beta \in \mathbb{R}^p$ is the unknown parameter, $X \in \mathbb{R}^p$ is the vector or explanatory variables and, $Y \in \mathbb{R}$ is the response variable and $\varepsilon \in \mathbb{R}$ is the error term. Assume that ε and X are independent and let F be the cdf of ε.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of i.i.d. copies of (X, Y). Assume that for each observation, X_i is observed but Y_i is not observed. Instead, what is observed is

$$Z_i = \mathbb{1}_{Y_i \geq 0}.$$

The random variables Y_i are called *latent variables* because they are not observed and the observed sample is $(X_1, Z_1), \ldots, (X_n, Z_n)$.

1. Conditional on X_1, what is the distribution of Z_1 ?
2. Write the link function in terms of F.
3. If ϵ is standard Gaussian, prove that the link function is Φ^{-1}, where Φ is the cdf of $\mathcal{N}(0, 1)$. What is the name of the model in that case ?

4. Assume that the density of ϵ is the logistic function:

$$f(t) = \frac{e^{-t}}{(1+e^{-t})^2}, \quad t \in \mathbb{R}.$$

a) Compute $F(t)$, for $t \in \mathbb{R}$.
b) Compute the link function.
c) What is the name of the model in that case ?