STAT 598G Spring 2011
Sergey Kirshner
Homework 3 (Written Exercises) Solution

March 22, 2011

1. (2 points) **Maximum likelihood for Laplace distribution.** Suppose $x_1, \ldots, x_n \overset{iid}{\sim} \mathsf{Laplace}(\lambda, \mu)$ with Laplace pdf $f_X(x|\lambda, \mu) = \frac{1}{2}\lambda e^{-\lambda|x-\mu|}, x \in \mathbb{R}$. Find MLE $\left(\hat{\lambda}, \hat{\mu}\right)$. Proof that your estimate is truly MLE.

Solution: First, compute the log-likelihood:

$$l(\lambda, \mu) = \sum_{i=1}^{n} \ln f_X(x_i|\lambda, \mu) = -n \ln 2 + n \ln \lambda - \lambda \sum_{i=1}^{n} |x_i - \mu|.$$

The, find critical points by setting the gradient to zero:

$$\frac{\partial l(\lambda, \mu)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} |x_i - \mu| = 0, \tag{1}$$

$$\frac{\partial l(\lambda, \mu)}{\partial \mu} = \lambda \sum_{i=1}^{n} \operatorname{sgn}(x_i - \mu) = \lambda\left(\# \text{ of } x_i > \mu \ - \ \# \text{ of } x_i < \mu\right) = 0. \tag{2}$$

So, from the equation (2) since $\lambda > 0$, $\hat{\mu} = \mathit{median}(x_1, \ldots, x_n)$. Plugging in $\hat{\mu}$ into equation (1), $\hat{\lambda} = n\left[\sum_{i=1}^{n} |x_i - \hat{\mu}|\right]^{-1}$.

To show that $\left(\hat{\lambda}, \hat{\mu}\right)$ is actually MLE and not just some other critical point, we need to show that for all $\lambda > 0$, $\mu \in \mathbb{R}$,

$$\begin{aligned} l(\lambda, \mu) \leq l\left(\hat{\lambda}, \hat{\mu}\right) &= -n \ln 2 + n \ln \frac{n}{\sum_{i=1}^{n} |x_i - \hat{\mu}|} - \frac{n}{\sum_{i=1}^{n} |x_i - \hat{\mu}|} \times \sum_{i=1}^{n} |x_i - \hat{\mu}| \\ &= -n \ln 2 + n \ln n - n \ln \sum_{i=1}^{n} |x_i - \hat{\mu}| - n. \end{aligned}$$

To see that $\left(\hat{\mu}, \hat{b}\right) = \operatorname{argmax}_{\mu, b} l(\mu, b)$ and not just a critical point of $l(\mu, b)$, note that $\forall \mu \in \mathbb{R}$, $\sum_{i=1}^{n} |x_i - \mu| \geq \sum_{i=1}^{n} |x_i - \hat{\mu}| > 0$. Also, for any $a > 0$, $g(\lambda) = n \ln \lambda - \lambda a$ takes the maximum value at $\lambda = \frac{n}{a}$ (as $\frac{\partial g(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - a$ and $\frac{\partial^2 g(\lambda)}{\partial \lambda^2} = -\frac{n}{\lambda^2} < 0$), so

$$n \ln \lambda - \lambda a = g(\lambda) \leq g\left(\frac{n}{a}\right) = n \ln \frac{n}{a} - \frac{n}{a} a = n \ln n - n \ln a - n.$$

so for $\mu \in \mathbb{R},\ \lambda > 0$ (setting $a = \sum_{i=1}^{n} |x_i - \hat{\mu}|$),

$$
\begin{aligned}
l(\lambda, \mu) &= -n \ln 2 + n \ln \lambda - \lambda \sum_{i=1}^{n} |x_i - \mu| \le -n \ln 2 + n \ln \lambda - \lambda \sum_{i=1}^{n} |x_i - \hat{\mu}| \\
&\le -n \ln 2 + n \ln n - n \ln \sum_{i=1}^{n} |x_i - \hat{\mu}| - n = l\left(\hat{\lambda}, \hat{\mu}\right).
\end{aligned}
$$

2. (5 points) Decide whether the distributions below fall within the exponential family or not. If they do, describe their set of (natural) parameters $\boldsymbol{\theta}$, features $\boldsymbol{\phi}(\boldsymbol{x})$, reference distribution $p_0(\boldsymbol{x})$, and the log-partition function $g(\boldsymbol{\theta})$.

(a) **Multivariate Normal.** $p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\ \boldsymbol{x} \in \mathbb{R}^d$ with mean parameter $\boldsymbol{\mu} \in \mathbb{R}^d$, and a symmetric positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}_+^d$ (a set of symmetric positive semidefinite matrices with entries in $\mathbb{R}^{d\times d}$).

Solution:

$$
p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}|\right),
$$

so setting $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_1 = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\theta}_2 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $\phi_1(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}\boldsymbol{x}^T$ ($\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\boldsymbol{x}^T\right)$), $\phi_2(\boldsymbol{x}) = \boldsymbol{x}$, $g(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| = \frac{1}{2}\left(\boldsymbol{\theta}_1^{-1}\boldsymbol{\theta}_2\boldsymbol{\theta}_1 + d\ln(2\pi) - \ln|\boldsymbol{\theta}_1|\right)$, and $p_0(\boldsymbol{x}) = 1$, and multivariate normal is a member of the exponential family.

(b) **Bounded Pareto.** $p(x|\alpha) = \frac{\alpha l^\alpha}{x^{\alpha+1}\left(1-\left(\frac{l}{m}\right)\right)^\alpha}$, $x \in (l, m)$ ($0 < l < m$ are both assumed given) with parameter $\alpha > 0$.

Solution:

$$
p(x|\alpha) = \frac{1}{x}\exp\left(\alpha(-\ln x) - \left(-\ln\alpha + \alpha\ln l + \ln\left(1 - \left(\frac{l}{m}\right)^\alpha\right)\right)\right)
$$

with $\theta = \alpha$, $p_0(x) = \frac{1}{x}$, $\phi(x) = -\ln x$, $g(\theta) = -\ln\theta + \theta\ln l + \ln\left(1 - \left(\frac{l}{m}\right)^\theta\right)$, so bounded Pareto falls within the exponential family.

(c) **Laplace.** $p(\boldsymbol{x}|\mu, \lambda) = \frac{1}{2}\lambda e^{-\lambda|x-\mu|}$, $x \in \mathbb{R}$ with parameters $\mu \in \mathbb{R}$, $\lambda > 0$.

Solution:

$$
p(x|\lambda, \mu) = \exp(-\ln 2 + \ln\lambda - \lambda|x-\mu|).
$$

However, $|x-\mu|$ does not decompose into a product of two function, one is only of x, and the other one is only of μ, so unless $\mu = 0$, Laplace distribution is not a member of the exponential family.

(d) **Von Mises distribution.** $f(x|\kappa, \mu) \propto e^{\kappa\cos(x-\mu)}$, $x \in (-\pi, \pi]$ with parameters $\mu \in (-\pi, \pi]$, $\kappa > 0$.

Solution:

$$
p(x|\mu, \kappa) = \exp(\kappa\cos\mu\cos x + \kappa\sin\mu\sin x - g(\kappa)),
$$

so setting $\boldsymbol{\theta} = (\kappa\cos\mu, \kappa\sin\mu)$, $\boldsymbol{\phi}(\boldsymbol{x}) = (\cos x, \sin x)$, $g(\boldsymbol{\theta})$ is a log-partition function, and $p_0(x) = 1$, so von Mises distribution falls within the exponential family.

(e) **Gumbel distribution.** $f(x|\mu, \sigma) = e^{-\frac{x-\mu}{\sigma}} e^{-e^{-\frac{x-\mu}{\sigma}}}$, $x \in \mathbb{R}$ with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$.

Solution:

$$p(x|\mu,\sigma) = \exp\left(-\frac{x-\mu}{\sigma} - e^{-\frac{x-\mu}{\sigma}}\right).$$

The first term in the exponential is in the linear form, but the second one cannot separate x and the rest of parameters, so Gumbel's distribution is not a member of the exponential family.

3. (1 point) (BV, Exercise 3.3) **Inverse of an increasing convex function.** Suppose $f:\mathbb{R}\to\mathbb{R}$ is increasing and convex on its domain (a,b). Let g denote its inverse, i.e., the function with domain $(f(a), f(b))$ and $g(f(x)) = x$ for $a < x < b$. What can you say about convexity or concavity of g?

   Solution:

   $$g'(x) = 1/f'(g(x)), \qquad g''(x) = -\frac{1}{(f'(g(x)))^2} f''(g(x))\, g'(x).$$

   Since $f(x)$ is convex, $f''(g(x)) \ge 0$. Since $f(x)$ is increasing, so is $g(x)$, so $f'(x) > 0$, so $g''(x) \le 0$, and therefore, g is concave.

4. (1 point) Prove (rigorously) that $f(x,y) = x^2/y$ is convex for $y > 0$.

   Solution: Consider the Hessian $\nabla_f^2(x,y)$ (matrix of second partial derivatives of f) – we will show that it is positive semi-definite (i.e., $(w,z)\nabla_f^2(w,z)^T \ge 0 \forall w,z\in\mathbb{R}$:

   $$\begin{aligned}
   \frac{\partial f(x,y)}{\partial x} &= \frac{2x}{y}, \\
   \frac{\partial f(x,y)}{\partial y} &= -\frac{x^2}{y^2}, \\
   \frac{\partial^2 f(x,y)}{\partial x^2} &= \frac{2}{y}, \\
   \frac{\partial^2 f(x,y)}{\partial x \partial y} &= \frac{\partial^2 f(x,y)}{\partial y \partial x} = -\frac{2x}{y^2}, \\
   \frac{\partial^2 f(x,y)}{\partial y^2} &= \frac{2x^2}{y^3}, \\
   (w,z)\nabla_f^2(x,y)(w,z)^T &= \frac{2}{y}w^2 - 2\frac{2x}{y^2}wz + \frac{2x^2}{y^3}z^2 = \frac{2}{y^3}\left(y^2w^2 - 2xywz + x^2y^2z^2\right) = \frac{2}{y^3}(yw - xz)^2 \ge 0 \text{ if } y > 0.
   \end{aligned}$$

5. (1 point) Given $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Consider the Lasso logistic regression problem

   $$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\lambda) \triangleq \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right) + \lambda \sum_{j=1}^{d} |\beta_j| \right\}. \qquad (3)$$

   Prove that the objective function $L(\lambda)$ is convex with respect to $\boldsymbol{\beta}$ for $\lambda \ge 0$.

   Solution: The absolute value is a convex function, as is a linear combination of convex functions, so $\lambda\sum_{j=1}^d |\beta_j|$ is convex. We will show that $\ln\left(1 + \exp\left(-y_i\boldsymbol{\beta}^T\boldsymbol{x}_i\right)\right)$ is convex by showing that its Hessian $\nabla_f^2(\boldsymbol{\beta})$ is positive semi-definite, and again, as the linear combination of convex functions is convex, so would be $L(\lambda)$ in (3). Assume that $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$.

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} \ln\left(1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right) &= \frac{-\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right) y_i x_{ij}}{1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)} = \left(\frac{1}{1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)} - 1\right) y_i x_{ij}, \\
\frac{\partial^2}{\partial \beta_j \beta_k} \ln\left(1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right) &= -\frac{y_i x_{ij}}{\left(1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right)^2} \exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\left(-y_i x_{ik}\right) \\
&= \frac{y_i^2 \exp\left(-y_y \boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\left(1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right)^2} \times x_{ij} x_{ik},
\end{aligned}$$

which is a positive semi-definite matrix as

$$\nabla_f^2(\boldsymbol{\beta}) = \frac{y_i^2 \exp\left(-y_y \boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\left(1+\exp\left(-y_i \boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right)^2} \boldsymbol{x}_i^T \boldsymbol{x}_i \succeq 0.$$