Statistics 860 Lecture 10
Exponential Families: Including Gaussian and Non-Gaussian data:

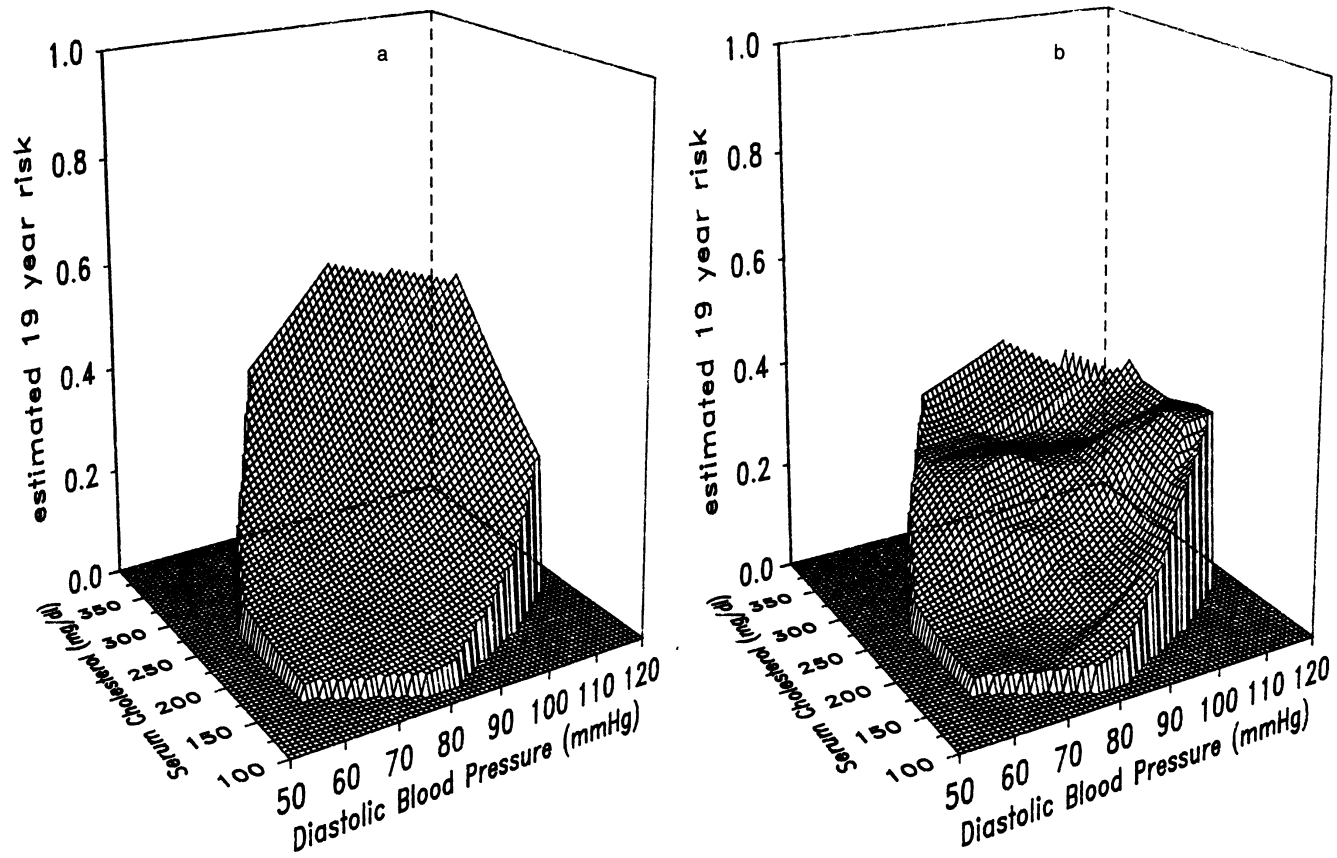


Figure 3. Risk Estimates: (a) from linear generalized linear model; (b) smoothed.

From O'Sullivan,Yandell and Raynor, JASA, (1986): 19-year risk of a heart attack given cholesterol and diastolic blood pressure at the start of the study. (Copy of the paper in pdf1/osullivan.yandell.raynor.pdf)

$$y = \begin{cases} 1 & \text{have heart attack(in 19 years)} \\ 0 & \text{do not have heart attack} \end{cases}$$

$$t_i = (x_{i_1}, x_{i_2})$$

x_1=cholesterol, x_2=diastolic blood pressure

$p(t) = p(1|t)$ =Probablity of have heart attack given t at start of study

$f(t) = \log \frac{p(t)}{1-p(t)}, \quad p(t) = \frac{\exp f(t)}{1+\exp f(t)}$

Find $f \in \mathcal{H}$ to minimize $\quad (f_i = f(t_i))$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} -y_i f_i + \log\left(1 + e^{f_i}\right)}_{\text{negative log likelihood}} + \lambda \mathcal{J}(f)$$

$$\mathcal{J}(f) = \iint f_{x_1 x_1}^2 + 2 f_{x_1 x_2}^2 + f_{x_2 x_2}^2 dx_1 dx_2$$

Thin plate penalty.

$y_i \in \{0, 1\}$: “Bernoulli data”

The Bernoulli distribution is an important member of the exponential family .

Gaussian case:

$$y_i = f(t_i) + \epsilon_i \quad , \quad \epsilon_i \sim N(0, \sigma^2)$$

$$y_i \sim N(f(t_i), \sigma^2)$$

$$F_{y,f} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - f(t_i))^2} \quad f_i = f(t_i)$$

$$\begin{aligned} -\text{log likelihood} &= \frac{1}{2\sigma^2}(y_i - f_i)^2 + \frac{1}{2}\log(2\pi\sigma) \\ &= \frac{1}{\sigma^2}[-y_i f_i + \frac{f_i^2}{2}] + \frac{y_i^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma \\ &= \frac{1}{\sigma^2}[-y_i f_i + b(f_i)] + c(y_i, \sigma) \\ b(f_i) &= \frac{f_i^2}{2} \end{aligned}$$

General case with parameter f_i:

$$
\begin{aligned}
-\text{log likelihood} &= \frac{1}{a(\phi_i)}[-y_i f_i + b(f_i)] + c(y_i, \phi_i) \\
b'(f_i) &= Ey_i \\
a(\phi_i)b''(f_i) &= Var(y_i) = \sigma^2
\end{aligned}
$$

Gaussian :

$$
\begin{aligned}
b'(f_i) &= f_i \\
b'' &= 1
\end{aligned}
$$

The "Canonical Link" $\mathcal{L}$ relates f_i to the parameter of interest. Since f_i is the parameter of interest $\mathcal{L}(f_i) = f_i$

Reference: McCullagh and Nelder, Generalized Linear Models, Chapman and Hall, 1983.

Bernoulli data:

$$y_i = \begin{cases} 1 & \text{with probablity } p_i \\ 0 & \text{with probablity } 1 - p_i \end{cases}$$

$$f_i = \log \frac{p_i}{(1 - p_i)}$$

$$F_p = p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$\text{-Log liklihood} = -y_i \log p_i - (1 - y_i) \log (1 - p_i)$$

$$= -y_i f_i + b(f_i)$$

$$b(f_i) = \log (1 + e^{f_i})$$

$$b'(f_i) = \frac{e^{f_i}}{1 + e^{f_i}} = Ey_i = p_i$$

$$b'' = \frac{e^{f_i}}{(1 + e^{f_i})^2} = p_i(1 - p_i) = Var(y_i)$$

Canonial link $\mathcal{L}(p_i) = f_i = \log \frac{p_i}{1-p_i}$

Before we leave the Bernoulli data optimization problem we remark that the result is an estimate of the probability that the subject with given attributes in class ”1”, and there will be an interesting connection to the optimization problem that implements the support vector machine (SVM) - the SVM makes a ’hard’ classification, whereas the Bernoulli likelihood estimate makes a ’soft’ or probabilistic classification. The SVM will be discussed later. In the meantime, See `http://www.pnas.org/content/99/26/16524.full`.

Poisson:

$y_i = k$ with probablity $\frac{\lambda_i^k e^{-\lambda_i}}{k!}, \quad k = 1, 2, \cdots$

-Log likelihood $= -y_i f_i + e^{f_i} + \log(y_i!)$

Canonical Link: $\mathcal{L}(f_i) = \log \lambda_i$

$$b(f_i) = e^{f_i} = \lambda_i$$
$$b'(f_i) = e^{f_i} = \lambda_i = Ey_i$$
$$b''(f_i) = e^{f_i} = \lambda_i = Vary_i$$

Risk Factor estimation:

$$y_i = \begin{cases} 1 & \text{with probablity} p_i = \frac{e^{f_i}}{1+e^{f_i}} \\ 0 & \text{with probablity} 1 - p_i = \frac{1}{1+e^{f_i}} \end{cases}$$

find $f \in \mathcal{H}$ to minimize

$$\underbrace{\mathcal{L}(y, f)}_{\text{negative log likelihood(+constant)}} + \lambda \|P_1 f\|^2$$

$$\mathcal{L}(y, f) = -\sum y_i f_i + \log\left(1 + e^{f_i}\right)$$

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1, \mathcal{J}(f) = \|P_1 f\|^2$$

$$f_\lambda = \sum d_\nu \phi_\nu + \sum c_i \xi_i$$

If $\mathcal{L}(y, f)$ is log likelihood from an exponential family , then it is strictly convex.

Bernoulli case:

Find c,d to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(-y_i f_i + \log\left(1 + e^{fi}\right)) + \lambda c'\Sigma c$$

Where $\Sigma_{ij} = \langle \xi_i, \xi_j \rangle$,

$$\begin{pmatrix} f_1 \\ \cdot \\ \cdot \\ \cdot \\ f_n \end{pmatrix} = T'd + \Sigma c, \quad T'c = 0,$$

using the Newton-Raphson method.

Newton-Raphson:

find $\theta = (\theta_1, \cdots, \theta_k)$ to minimize $I(\theta)$ where $I(\theta)$ is a strictly convex function of θ.
The second order Taylor expansion of $I(\theta)$ about the l^{th} iterate $\theta^{(l)}$ is

$$I(\theta) \simeq I(\theta^{(l)}) + \nabla I(\theta - \theta^{(l)}) + \frac{1}{2}(\theta - \theta^{(l)})'\nabla^2 I(\theta - \theta^{(l)}) + \cdots \quad (*)$$

where

$$\nabla I = (\frac{\partial I}{\partial \theta_1}, \cdots, \frac{\partial I}{\partial \theta_k})'_{\theta=\theta^{(l)}} \qquad \text{Gradient}$$

$$\{\nabla^2 I\}_{jk} = \frac{\partial^2 I(\theta)}{\partial \theta_j \partial \theta_k}|_{\theta=\theta^{(l)}} \qquad \text{Hessian}$$

Then $\theta = \theta^{(l+1)}$ is the minimizer of $(*)$

$$\theta^{(l+1)} = \theta^{(l)} - (\nabla^2 I)^{-1}(\nabla I)'$$

What is a good criteria for choosing λ?

Gaussian case:(know σ^2)

$$y_i = f_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$
$$f_i : \text{true} \qquad f = (f_i, \cdots, f_n)$$
$$f_{i\lambda} : \text{estimation} \qquad f = (f_{i\lambda}, \cdots, f_{n\lambda})$$

$$KL(g_\lambda, g) = E_g(\log \frac{g}{g_\lambda}) \qquad \text{(Kullback-Leibler distance)}$$
$$g_i \sim N(f_i, \sigma^2)$$
$$g_{i\lambda} \sim N(f_{i\lambda}, \sigma^2)$$

$$E_{f_i}\{-\frac{1}{2\sigma^2}[(f_i - y_i)^2 - (f_{i\lambda} - y_i)^2]\}$$
$$= E_{f_i}\{-\frac{1}{2\sigma^2}[f_i^2 - 2f_i y_i + y_i^2 - f_{i\lambda}^2 + 2f_{i\lambda} y_i - y_i^2]\}$$
$$= -\frac{1}{2\sigma^2}[f_i^2 - 2f_i^2 - f_{i\lambda}^2 + 2f_{i\lambda} f_i]$$
$$= -\frac{1}{2\sigma^2}[-f_i^2 + 2f_{i\lambda} f_i - f_{i\lambda}^2]$$
$$= \frac{1}{2\sigma^2}[(f_i - f_{i\lambda})^2] = \frac{\text{Predictive MSE}}{2\sigma^2}$$

If know σ^2-use unbiased risk estimator for **this target**, otherwise GCV.

Comparative KL(for λ)-remove anything that does not depend on λ

$$\begin{aligned} CKL(\lambda) =& \frac{1}{2\sigma^2}[f_i^2 - 2f_{i\lambda}f_i + f_{i\lambda}^2] - \frac{1}{2\sigma^2}f_i^2 \\ =& \frac{1}{2\sigma^2}[-f_{i\lambda}f_i + \frac{f_{i\lambda}^2}{2}] \\ =& \frac{1}{2\sigma^2}[-\mu_i f_{i\lambda} + b(f_{i\lambda})] \end{aligned}$$

$\mu_i = E_{f_i} y_i = f_i, \qquad \frac{f_{i\lambda}^2}{2} = b(f_{i\lambda})$ for Gaussian case

.

Target for choosing λ to minimize

$$\sum_i (f_i - f_{i\lambda})^2$$

equivalent to minimizing

$$\sum_i (-\mu_i f_{i\lambda} + b(f_{i\lambda}))$$

General exponential family with no nuisance parameter:

$$g(y_i, f_i) = e^{\{y_i f_i - b(f_i) + c(y_i)\}}$$
$$g(y_i, f_{i\lambda}) = e^{\{y_i f_{i\lambda} - b(f_{i\lambda}) + c(y_i)\}}$$

Bernoulli data:

$$y_i = \begin{cases} 1 & \text{with probablity} p_i \\ 0 & \text{with probablity} 1 - p_i \end{cases}$$

$$f_i = \log \frac{p_i}{1 - p_i}$$
$$b(f_i) = \log (1 + e^{f_i})$$
$$CKL(\lambda) = KL(f_i, f_{i\lambda}) - [y_i f_i - b(f_i)]$$
$$= -\mu_i f_{i\lambda} + b(f_{i\lambda})$$
$$\mu_i = b'(f_i) = Ey_i = p_i$$

GACV estimate for λ:

- D.Xiang and G.Wahba, *A generalized approximate cross validation for smoothing splines with non-Gaussian data*. Statistica Sinica 6, 675-692, 1996. `xiang.wahba.sinica.pdf`

- X.Lin, G.Wahba, D.Xiang, F.Gao, R.Klein, and B.Klein. *Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV*.Ann. Statist., 28:1570–1600, 2000. `lin.wahba.xiang.gao.pdf`

The GACV estimate of λ. Xiang Wahba:1996

$$
\begin{aligned}
OBS(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{i\lambda} + b(f_{i\lambda})] \\
CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{i\lambda}^{[-i]} + b(f_{i\lambda})] \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}[y_i(y_i - \mu_{i\lambda}^{[-i]})]\left[\frac{f_{i\lambda} - f_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}}\right] \\
&= OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\left[y_i(\frac{y_i - \mu_{i\lambda}}{1 - \frac{\mu_{i\lambda} - \mu_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}}})\right]\left[\frac{f_{i\lambda} - f_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}}\right] \\
&\approx OBS(\lambda) + \frac{1}{n}\sum_{i=1}^{n}\left[y_i(\frac{y_i - \mu_{i\lambda}}{1 - \sigma_{i\lambda}^2\left[\frac{f_{i\lambda} - f_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}}\right]})\right]\left[\frac{f_{i\lambda} - f_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}}\right]
\end{aligned}
$$

$\sigma_{i\lambda}^2 = \sigma^2(f_{i\lambda})$

The last approximation comes from recalling that $\mu = e^f/(1+e^f)$ and $\frac{\partial \mu}{\partial f} = \sigma^2$ and setting $\frac{\mu_{i\lambda} - \mu_{i\lambda}^{[-i]}}{f_{i\lambda} - f_{i\lambda}^{[-i]}} \approx \sigma_{i\lambda}^2$.

If $J(f) = \|f\|^2$ then letting $f = (f_1, \cdots, f_n)'$, $f = \Sigma c, \|f\|^2 = c'\Sigma c = f'\Sigma^{-1} f$. in general, let $f = \Sigma c + Td$. Then, let Σ_λ be twice the matrix of the penalty quadratic form in terms of f. It can be shown that Σ_λ is given by

$$\Sigma_\lambda = 2\lambda(\Sigma^{-1} - \Sigma^{-1}T(T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1})$$

$$I_\lambda(f, Y) = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n}[-y_i f_i + b(f_i)] + \frac{1}{2}f'\Sigma_\lambda f. \quad (1)$$

Let $W = W(f)$ be the $n \times n$ diagonal matrix with σ_i^2 in the iith position. For Bernoulli data $\sigma_i^2 \equiv \mu_i(1 - \mu_i)$. Using the fact that σ_i^2 is the second derivative of $b(f_i)$, we have that $H = [W + \Sigma_\lambda]^{-1}$ is the inverse Hessian of the variational problem (1).

Note: This makes use of the special structure of exponential families. $b'' > 0$ always. The variational problem is strictly convex if T is of full rank, and

$$f_\lambda^{Y+\epsilon} - f_\lambda^Y \approx (W(f_\lambda^Y) + n\Sigma_\lambda)^{-1}\epsilon \equiv H\epsilon, \qquad (2)$$

where $H = H(\lambda)$. Equation (2) can also be invoked (roughly) to justify the approximation

$$\frac{f_{i\lambda} - f_{i\lambda}^{[-i]}}{y_i - \mu_{i\lambda}^{[-i]}} \approx h_{ii},$$

where h_{ii} is the iith entry of H. (h_{ii} plays the same role as $a_{ii}(\lambda)$).

Leaving Out One Lemma will also give the same approximation involving h_{ii}:

Let $Y^{[-i]} = (y_1, \ldots, y_{i-1}, \mu_\lambda^{[-i]}(x_i), y_{i+1}, \ldots, y_n)'$

$$
\begin{aligned}
f_\lambda^Y - f_\lambda^{Y^{[-i]}} &\approx \left(W(f_\lambda^Y) + n\Sigma_\lambda\right)^{-1} \begin{pmatrix} \\ Y - Y^{[-i]} \\ \\ \end{pmatrix} \\
&= \left(W(f_\lambda^Y) + n\Sigma_\lambda\right)^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ y_i - \mu_\lambda^{[-i]}(x_i) \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\end{aligned}
$$

$$\frac{f_\lambda(x_i) - f_\lambda^{[-i]}(x_i)}{y_i - \mu_\lambda^{[-i]}(x_i)} \approx h_{ii}$$

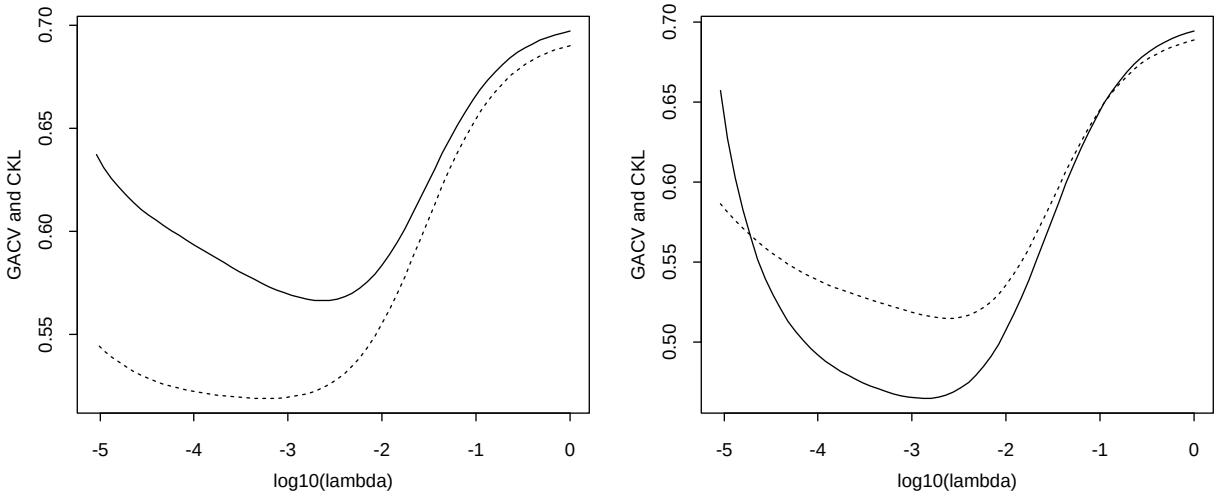
Then $CV(\lambda) \approx ACV(\lambda)$:

$$ACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_{i\lambda} + b(f_{i\lambda})] + \frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\frac{y_i - \mu_{i\lambda}}{1-\sigma_{i\lambda}^2 h_{ii}}\right)\right] h_{ii}\,.$$

The $GACV$ is obtained from the ACV by replacing h_{ii} by $\frac{1}{n}\sum_{i=1}^{n} h_{ii} \equiv \frac{1}{n}tr(H)$ and replacing $1-\sigma_{i\lambda}^2 h_{ii}$ by $\frac{1}{n}tr[I-(W^{1/2}HW^{1/2})]$, giving $GACV(\lambda) =$

$$\frac{1}{n}\sum_{i=1}^{n}[-y_i f_{i\lambda} + b(f_{i\lambda})] + \frac{1}{n}trH\frac{\sum_{i=1}^{n} y_i(y_i - \mu_{i\lambda})}{tr[I-(W^{1/2}HW^{1/2})]}\,,$$

where W is evaluated at f_λ.



$GACV(\lambda)$ solid lines, $CKL(\lambda)$ dotted lines.