TOPIC:

DATE:

FILE UNDER:

PAGE:

Secant method

If f' is difficult to compute (almost always) ~~are~~ or we are lazy (always) the the secant method provides an approximation.

Newton step / secant method

$$\cancel{x_{n+1} = x_n - \frac{f(x_n)}{}}$$

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$$

$$\Rightarrow f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

~~Pro~~ Note: Need 2 starting values.

Pro: easy as pie

Con: Convergence only super linear.

Statistics application

$l(\theta): \mathbb{R}^K \to \mathbb{R}$, log-likelihood, $\theta = (\theta_1, \ldots, \theta_K)$

$l'(\theta): \mathbb{R}^K \to \mathbb{R}^K$, ~~Score~~ gradient

$l''(\theta): \mathbb{R}^K \to \mathbb{R}^{K \times K}$, hessian

$x_n - x_{n-1} =$

$x_n - x_\infty + x_\infty - x_{n-1}$

$\varepsilon_n \quad \varepsilon_{n-1}$

$$\varepsilon_{n+1} = \varepsilon_n - \frac{f(x_\infty + \varepsilon_n)(x_n - x_{n-1})}{f(x_\infty + \varepsilon_n) - f(x_\infty + \varepsilon_{n-1})}$$

$$\varepsilon_n - \frac{\left(\varepsilon_n f' + \frac{\varepsilon_n^2}{2} f''\right)(\quad)}{\varepsilon_n f' + \frac{\varepsilon_n^2}{2} f'' - \varepsilon_{n-1} f' - \frac{\varepsilon_{n-1}^2}{2} f''}$$

$$= \varepsilon_n - \frac{\left(\varepsilon_n f' + \frac{\varepsilon_n^2}{2} f''\right)(\varepsilon_n - \varepsilon_{n-1})}{(\varepsilon_n - \varepsilon_{n-1}) f' - \left(\frac{\varepsilon_n^2}{2} - \frac{\varepsilon_{n-1}^2}{2}\right) f''}$$

TOPIC: DATE:

FILE UNDER: PAGE:

Fisher Scoring for GLMs

$Y \sim$ ~~[illegible]~~ , $p(y|\mu)$ $\quad E[y] = \mu$

$g(\mu) = X\beta$ $\quad Var(y) = V(\mu)$

$$g(y) \doteq g(\mu) + (y-\mu)g'(\mu)$$

$g(y) \downarrow$ z (working response)

$$\eta = g(\mu)$$

1. Start with $\hat{\mu}_i$
2. Set $z_i = \eta_i + (y_i - \hat{\mu}_i)\left(\frac{\partial \hat{\eta}_i}{\partial \mu_i}\right)$
3. Weighted regression of z on X, i.e. solve

$$X^T W X \beta_n = X^T W z$$

$$\beta_n = (X^T W X)^{-1} X^T W z$$

where $W = \left[\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 V(\mu_i)\right]^{-1}$

ex: $y_i \sim \text{Poisson}(\mu_i)$

$$g(\mu_i) = \log \mu_i = \eta_i$$

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i} \qquad V(\mu_i) = \mu_i$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

1. Set $\hat{\mu}_i$

2. $z_i = \log \mu_i + (y_i - \mu_i)/\mu_i$

3. Regress z on X where

$$W = \left[\frac{1}{\hat{\mu}_i^2} \hat{\mu}_i\right]^{-1} = \hat{\mu}_i$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Poisson regression via Newton

$$L(\beta) \propto e^{-\mu} \mu^{Y}$$

$$\ell(\beta) = Y \log \mu - \mu$$
$$= Y X\beta - \exp(X\beta)$$
$$\ell'(\beta) = XY - X\exp(X\beta)$$
$$\ell''(\beta) = -X \underbrace{\exp(X\beta)}_{\mu} X$$

$$Y^T X\beta - \exp(X\beta)$$
$$X^T Y - X^T \mu$$
$$X^T(Y-\mu)$$
$$-X^T W X$$

$$z = \eta + (Y-\mu)/\mu$$
$$\mu z = \mu\eta + Y - \mu$$
$$Y-\mu = \mu z - \mu\eta$$

$$\beta_{n+1} = \beta_n + \left(-X^T W X\right)^{-1}\left(-X^T(Y-\mu)\right)$$
$$= \beta_n + (X^T W X)^{-1} X^T(\mu z - \mu\eta)$$
$$= \beta_n + (X^T W_n X)^{-1} X^T W_n (z_n - \eta_n)$$
$$= (X^T W X)^{-1} X^T W z + \underbrace{\left(\beta_n - (X^T W X)^{-1} X^T W \eta\right)}_{=0}$$
$$= (X^T W X) X^T W z$$

$$\eta = \log \mu$$
$$= X\beta$$

$$\eta = g(\mu) = \log \mu$$

$$\eta = z - \frac{Y-\mu}{\mu}$$

$X\beta$

$\mu = \exp$

$\log =$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Summary of Minimization

$$\min_{x} f(x) \quad \text{for} \quad f : \mathbb{R}^K \to \mathbb{R}, \; x \in \mathbb{R}^K$$

Line Search: ~~Step~~

① Steepest descent

$$X_{n+1} = \underbrace{X_n + \alpha}_{\text{scalar step length}} \underbrace{\left[-f'(X_n)\right]}_{\text{direction of steepest descent}}$$

$\Rightarrow$ linear convergence

② Newton

$$X_{n+1} = X_n + 1 \underbrace{\left[f''(X_n)\right]^{-1}\left[-f'(X_n)\right]}_{\text{Newton direction}}$$

Step length $= 1$

$\Rightarrow$ quadratic convergence

$\rightarrow$ in stat, estimate of asymptotic covariance

③ Quasi-Newton

$$X_{n+1} = X_n + \alpha B_n \left[-f'(X_n)\right]$$

ⓐ $B_n = \arg\min_{B} \| B - B_{n-1} \|$

$B = B^T$ and $f'(X_n) - f'(X_{n-1}) = B_n (X_n - X_{n-1})$

TOPIC:

DATE:

FILE UNDER:

PAGE:

4) Conjugate Gradient ("Modified steepest descent")

$$x_{n+1} = x_n + \alpha p_n$$

$$p_{n+1} = -f_n'(x_{n+1}) + \frac{\|f'(x_{n+1})\|^2}{\|f'(x_n)\|^2} p_n$$

$$p_0 = -f'(x_0)$$

$$\cancel{p_0^T p_1 = f_0'^T f_1' + \frac{\|f_1'\|^2}{\|f_0\|^2} f_0'^T f_0}$$

Coordinate Descent:

$$x_{n+1}^{(k)} = \arg\min_{x^{(k)}} f(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)}, \ldots, x_n^{(p)})$$

$\Rightarrow$ has "descent property", non-increasing

TOPIC:

DATE:

FILE UNDER:

PAGE:

Do this $\Rightarrow$

~~Method of steepest descent/ascent uses~~

~~$\theta_{n+1} = \theta_n - \alpha\, l'(\theta_n)$~~

~~Where α is chosen so that θ_{n+1} has a larger likelihood value than θ_n.~~

Summary:

Knowledge of l''

✓ (1) Newton: Fastest (quadratic convergence), requires calculating l'', gives asymptotic $Var(\hat{\theta})$, unstable if starting value too far.

~ (2) Scoring: Super linear convergence, but equiv. to Newton if l'' does not depend on Y (true in many common cases), requires calculating $E[l'']$, can be unstable, but often quite stable in typical statistics apps.

~ (3) Quasi-Newton: Superlinear convergence, does not require l'' or $E[l'']$, more stable than Newton, No estimate of $Var(\hat{\theta})$

~~(4) Steepest descent: linear convergence, Very stable~~

X (4) Steepest descent: linear convergence, stable

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~**The EM algorithm**~~

~~EM stands for Expectation-Maximization.~~

~~Originally outlined in DLR (1977) but ideas go much farther back. DLR unified many different ideas and gave examples of the broad applicability of EM.~~

~~EM is not an "algorithm". It is an algorithm for creating other algorithms.~~

Generalized Additive Models (Hastie + Tibshirani)

Usual linear model has:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

GAMS say:

$$Y_i = \alpha + S_1(X_{1i}) + S_2(X_{2i}) + \cdots + S_p(X_{pi}) + \varepsilon_i$$

$$\Rightarrow Y = \alpha + \sum_{j=1}^{p} S_j(X_j) + \varepsilon \qquad (*)$$

where $S_j(\,)$ are smooth.

$S_j(\,)$ can be any kind of "smoother", even a mixture of different kinds

ex Smoothers: parametric splines, smoothing splines, penalized splines, loess, running lines, running median

TOPIC:

DATE:

FILE UNDER:

PAGE:

GAM algorithm:

$Y = (Y_1, \ldots, Y_n)$

Given model $Y = \alpha + \sum_{j=1}^{p} S_j(X_j) + \varepsilon$

(1) Initialize $\alpha = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $s_j = S_j^0 \overset{?}{=} 0$

(2) For $j = 1, 2, \ldots, p$

Let $r_j = Y - \alpha - \sum_{k \neq j} s_k(X_k)$

~~where~~ $\cancel{s_k = (s_k(x_{k1}), s_k(x_{k2}), s_k(x_{k3}), \ldots, s_j(x_{jn}))}$

where $S_k = (S_k(X_{k1}), S_k(X_{k2}), \ldots, S_k(X_{kn}))$

i.e. $s_k(\,)$ evaluated at data points for X_k

$S_j = \text{Smooth}(r_j \mid X_j)$

$\hat{S}_j = S_j - \sum_{i=1}^{n} s_j(x_{ij})$

(3) Evaluate $\Delta = \sum_{j=1}^{p} \| s_j - s_j^0 \|$

or $\Delta = \dfrac{\sum_{j=1}^{p} \| S_j - S_j^0 \|}{\sum \| S_j^0 \|}$

If $\Delta < \varepsilon$, stop.

Else set $S_j^0 = S_j$ for $j = 1, \ldots, p$.

~~Got~~ Goto (2)

TOPIC:

DATE:

FILE UNDER:

PAGE:

Local Scoring

Set adjusted response

$$z_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

Fit an additive model for

$z \sim x_1, \ldots, x_p$ with observation weights

$$w_i = \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-2} V(\mu_i)^{-1}$$

GAM algorithm = "backfitting"

~ Alternating conditional expectation
~ Cyclic coordinate descent

⇒ linear convergence algorithm

⇒ sidesteps curse of dimensionality by additivity constraint (bz!)

TOPIC:

DATE:

FILE UNDER:

PAGE:

EM Algorithm

EM stands for Expectation-~~Maximization~~ Maximization.

Originally by DLR, 1977 but ideas go much farther back. DLR united many different ideas and put them in a statistical framework.

meta-algorithm?

EM is not strictly an "algorithm". It is an (abstract) algorithm for creating other algorithms

The basic principle of EM is straightforward.

We observe some data Y but there are some data that are unavailable or "missing". Call these data Z.

The observed data ~~X~~ Y with the missing data Z are the complete data $X = (Y, Z)$

(1) The complete data have a joint density $g(\overbrace{Y,Z}^{\text{complete data}} \mid \theta)$

(2) Because of missing Z, we cannot evaluate g. We observe Y with a joint density

$$f(y \mid \theta) = \int g(y, z \mid \theta)\, dz$$

$$\ell(\theta \mid y) = \log f(y \mid \theta)$$

(3) $\ell(\theta \mid y)$ is hard to evaluate! (because of $\int$)

(although, maybe not!)

TOPIC:

DATE:

FILE UNDER:

PAGE:

$\mathbb{E}\log X$

(if $f(y) \le f(\mathbb{E}X)$)

Direct ML maximizes $\ell(\theta|y)$. This may be possible!

Heuristically, the EM algorithm is as follows:

① E-step. Given estimate $\theta = \theta_0$

Define $Q(\theta|\theta_0) = \mathbb{E}\left[\log g(y,z|\theta) \,\middle|\, y, \theta_0\right]$

$$= \int p(z|y,\theta_0) \log g(y,z|\theta)\, dz$$

~~[illegible]~~ $\Rightarrow$ switch $\int$ and log

$\Rightarrow$ add $p(z|y,\theta_0)$

② Maximize $Q(\theta|\theta_0)$ wrt θ, ~~and set~~

~~Call set~~ $\cancel{\hat{\theta}_1 = \arg\max_{\theta} Q(\theta|\theta_0)}$

~~Set~~ $\cancel{\theta_0 = \hat{\theta}_1}$ ~~and goto ①~~

~~Create a Sequence~~

Set $\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n)$

Under broad assumptions, $\theta_n \xrightarrow[n\to\infty]{} \hat{\theta}$, MLE

(More later)

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex: $Y_1, \ldots, Y_m, Z_{m+1}, \ldots, Z_n \sim \text{Poisson}(\mu)$

$$g(x|\mu) \propto \prod_{i=1}^{n} \mu^{x_i} e^{-\mu}$$

$$\log g(x|\mu) = \sum_{i=1}^{n} x_i \log \mu - \mu$$

$$= \log \mu \sum x_i - n\mu$$

$$\mathbb{E}\left[\sum_{i=1}^{n} x_i \,\middle|\, Y\right] = \sum_{i=1}^{m} y_i + \sum_{i=m+1}^{n} \tilde{\mu}$$

$$\mathbb{E}_{\mu}\left[\sum_{i=1}^{n} x_i\right] = \mu n$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex: Censored exponential data

$Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$ but some cases are censored on right

Let ~~complete~~ observed data be

$(\min(Y_1, C_1), \delta_1), \ldots, (\min(Y_n, C_n), \delta_n)$

where $\delta_i = 1$ if $Y_i \le C_i$ and

$= 0$ if Y_i censored

$$g(x|\lambda) \propto \prod_{i=1}^{n} \frac{1}{\lambda} \exp(-x_i/\lambda)$$

$$\log g(x|\lambda) = -n \log \lambda - \frac{1}{\lambda} \sum x_i$$

$$= -n \log \lambda - \frac{1}{\lambda} \left\{ \sum_{obs} Y_i + \sum_{censored} Z_i \right\}$$

$$\mathbb{E}\left[\log g(x|\lambda) \,\middle|\, Y, \tilde{\lambda}\right]$$

$$= -n \log \lambda - \frac{1}{\lambda} \left[\sum_{obs} Y_i + \sum_{mis} C_i + 1/\tilde{\lambda} \right]$$

$$\mathbb{E}[Z_i \mid Y_i, C_i]$$

$$\Rightarrow \hat{\lambda} = \frac{1}{n} \left[\sum_{obs} Y_i + \sum_{mis} C_i + 1/\tilde{\lambda} \right]$$

TOPIC: EM

DATE:

FILE UNDER:

PAGE:

One sample

$Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$

$Y_1, Y_2^*, Y_3^*, Y_4, \ldots, Y_n \sim N(\mu, \sigma^2)$

$$L(\mu, \sigma^2) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_i - \mu)^2\right)$$

$$\ell = \sum_{i=1}^{n} -\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum(Y_i - \mu)^2$$

$$= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum(Y_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2}\sum(Y_i - \mu)$$

$Y_i^2 - 2\mu Y_i - \mu^2$

$\sigma_0^2 + \mu_0^2 - 2\mu\mu_0 - \mu^2$

$\sum 2\mu_0 - 2\mu$

$$\mathbb{E}[Y_i \mid \mu_0, \sigma_0^2] = \mu_0$$

$\mathbb{E} Y_i^2 = \sigma^2 + \mu^2$

Bivariate

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ & \sigma_2^2 \end{pmatrix}\right)$$

$\mu \qquad V$

$$L \propto \prod_{i=1}^{n} |V|^{-1/2} \exp\left(-\frac{1}{2}(Y_i - \mu)^T V^{-1}(Y_i - \mu)\right)$$

$$\ell = \sum_{i=1}^{n} -\frac{1}{2}\log|V| - \frac{1}{2}(Y_i - \mu)^T V^{-1}(Y_i - \mu)$$

$$= -\frac{n}{2}\log|V| - \frac{1}{2}\operatorname{tr}\left((Y - \mu) V^{-1}(Y_i - \mu)^T\right)$$

$$-\frac{1}{2}\operatorname{tr}\left((Y - \mu)(Y - \mu)^T V^{-1}\right)$$

$$-\frac{1}{2}\operatorname{tr}\left(\left(YY^T - Y\mu^T - \mu Y^T - \mu\mu^T\right)V^{-1}\right)$$

$n \times 2 \quad 2 \times 2 \quad 2 \times n$

$Y_i \quad n \times 2 \quad 2 \times 2$

$n \times 2 \quad 2 \times n$

$n \times n$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$$\ell = -\frac{n}{2}\log|V| - \frac{1}{2}\mathrm{tr}\left((Y_i-\mu)^T(Y_i-\mu)\,V^{-1}\right)$$

$$-\frac{1}{2}\mathrm{tr}\left(\left(Y_i^T Y - Y^T\mu - \mu^T Y + \mu^T\mu\right)V^{-1}\right)$$

$$\begin{pmatrix}\sum Y_{i1}^2 & \sum Y_{i1}Y_{i2}\\ \sum Y_{i2}Y_{i1} & \sum Y_{i2}^2\end{pmatrix}$$

$$Y_{ij}^{*2} = \left[\hat\mu_1 + \hat\rho\,\sigma_1\left(Y_{i2}-\hat\mu_1\right)\right]^2 + (1-\hat\rho)\hat\sigma_2^2$$

$$= \left(\mathbb{E}\,Y_{ij}\right)^2 + \mathrm{Var}\left(Y_{i1}\right) \Big| \hat\mu_1, \hat\mu_2, \hat\sigma_1^2, \hat\sigma_2^2, \hat\rho$$

$$\mathbb{E}\left[Y_{i1}^* Y_{i2}^* \mid v\right] = \mathbb{E}\left[Y_{i1}^*\right]\mathbb{E}\left[Y_{i2}^*\right] + \hat\rho\,\hat\sigma_1\hat\sigma_2$$

TOPIC: Minorizing function

DATE:

FILE UNDER:

PAGE:

Mbss class
11/26

$$\log f(y|\theta) = \log \int g(y,z|\theta)\,dz$$

$$\log f(y|\theta) - \log f(y|\theta_0)$$

$$= \log \int g(y,z|\theta)\,dz - \log \int g(y,z|\theta_0)\,dz$$

$$= \log \frac{\int g(y,z|\theta)\,dz}{\int g(y,z|\theta_0)\,dz}$$

$$= \log \frac{\int g(y,z|\theta_0) \frac{g(y,z|\theta)}{g(y,z|\theta_0)}\,dz}{\int g(y,z|\theta_0)\,dz}$$

$$= \log \int h(z|y,\theta_0) \frac{g(y,z|\theta)}{g(y,z|\theta_0)}\,dz$$

$$\geq \int h(z|y,\theta_0) \log \frac{g(y,z|\theta)}{g(y,z|\theta_0)}\,dz$$

$$= \int h(z|y,\theta_0) \log g(y,z|\theta)\,dz$$

$$- \int h(z|y,\theta_0) \log g(y,z|\theta_0)\,dz$$

$$\log f(y|\theta) \geq \log f(y|\theta_0) + \mathbb{E}_h\left[\log g(\cancel{y},\cancel{z}|\theta)\right] - \mathbb{E}_h\left[\log g(y,z|\theta_0)\right]$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$f(y|\theta)$ obs.

$g(y, z|\theta)$ complete

Ascent property of EM algorithm

The sequence $\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n)$

has the property that for each n,

$$\ell(\theta_{n+1}|y) \geq \ell(\theta_n|y)$$

$\Rightarrow \log f(y|\theta_{n+1}) \geq \log f(y|\theta_n)$

w/ strict inequality when $Q(\theta_{n+1}|\theta_n) > Q(\theta_n|\theta_n)$

~~Define~~ $\cancel{D(f\|g) = -\log \mathbb{E}}$

Define $D(f\|g) = \mathbb{E}_f[\log f/g]$

$\cancel{D(f\|g)}$ ~~is zero~~ $\cancel{\Leftrightarrow f = g}$

$D(f\|g) \geq 0$ (information inequality)

Pf:

$$D(f\|g) = \mathbb{E}_f[\log f/g]$$

$$= \mathbb{E}_f[-\log g/f]$$

Jensen's inequality, $\geq -\log \mathbb{E}_f[g/f]$

($-\log$ is a convex function) $= -\log \int g/f * f$.

$$= -\log 1$$

$$= 0$$

$D(\cdot\|\cdot)$ is like a distance (but no Δ inequality)

TOPIC:

DATE:

FILE UNDER:

PAGE:

Recall

$Q(\theta_{n+1}|\theta_n) \geq Q(\theta_n|\theta_n)$

assumption

by information inequality

$$\ell(\theta_{n+1}) - \cancel{\ell(\theta_n)} = \log f(y|\theta_{n+1}) - \cancel{\log f(y|\theta_n)}$$

$$= \log f(y|\theta_{n+1}) + \underbrace{Q(\theta_{n+1}|\theta_n) - Q(\theta_{n+1}|\theta_n)}_{0}$$

$$= Q(\theta_{n+1}|\theta_n) - \left[Q(\theta_{n+1}|\theta_n) - \log f(y|\theta_{n+1}) \right]$$

$$= Q(\theta_{n+1}|\theta_n) - \left[\mathbb{E}_p\left[\log g(y,z|\theta_{n+1}) \middle| y,\theta_n \right] - \log f(y|\theta_{n+1}) \right]$$

$$= Q(\theta_{n+1}|\theta_n) - \mathbb{E}_p\left[\log \frac{g(y,z|\theta_{n+1})}{f(y|\theta_{n+1})} \middle| y,\theta_n \right]$$

$$= Q(\theta_{n+1}|\theta_n) - \mathbb{E}_p\left[\log p(z|y,\theta_{n+1}) \middle| y,\theta_n \right]$$

$\hookrightarrow p(z|y,\theta_n)$

$$\geq Q(\theta_n|\theta_n) - \mathbb{E}_p\left[\log p(z|y,\theta_{n+1}) \middle| y,\theta_n \right]$$

$$\geq Q(\theta_n|\theta_n) - \mathbb{E}_p\left[\log p(z|y,\theta_n) \middle| y,\theta_n \right]$$

$$= Q(\theta_n|\theta_n) - \mathbb{E}_p\left[\log \frac{g(y,z|\theta_n)}{f(y|\theta_n)} \middle| y,\theta_n \right]$$

$$= Q(\theta_n|\theta_n) - \mathbb{E}_p\left[\log g(y,z|\theta_n) \middle| y,\theta_n \right] + \log f(y|\theta_n)$$

$$= Q(\theta_n|\theta_n) - Q(\theta_n|\theta_n) + \log f(y|\theta_n)$$

$$= \log f(y|\theta_n)$$

$$= \ell(\theta_n)$$

$\square$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Roughly,

If l is ~~convex~~ concave then $\{l(\theta_n)\}$ is a monotone increasing, bounded, seq. of numbers.

$\Rightarrow$ There is a limit.

~~But does $\theta_n \to \theta$~~

But does this mean $\theta_n \to \hat{\theta}$, MLE?

Not necessarily, but yes for exponential families.

For exp fam, there is always a unique maximizer.

<u>Note</u>: We do not <u>require</u> that

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta | \theta_n).$$

We <u>only</u> need $Q(\theta_{n+1} | \theta_n) \geq Q(\theta_n | \theta_n)$.

↳ This algorithm is Generalized EM (GEM).

Ex. ~~Mv~~ Bivariate Normal w/ missing data

$Y = (Y_1, Y_2) \sim N(\mu, \Sigma)$

$\mu = (\mu_1, \mu_2)$, $\Sigma = \begin{bmatrix} \sigma^2 & v \\ v & \tau^2 \end{bmatrix}$

$$\begin{array}{cc} Y_1 & Y_2 \end{array}$$
$$\begin{bmatrix} 3 & 5 \\ 6.1 & 8.3 \\ ? & 7.4 \\ 4 & ? \end{bmatrix}$$

$\theta = (\mu_1, \mu_2, \sigma^2, \tau^2, v)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex: One-way Random effects

$$Y_{ij} \overset{iid}{\sim} N(\mu_i, \sigma_\epsilon^2)$$

$$Y_{ij} = \mu_i + \epsilon_i \sim N(0, \sigma^2)$$

$$\mu_i \overset{iid}{\sim} N(\alpha, \tau^2)$$

$$p(Y_{i1}, Y_{i2}, \ldots, Y_{in_i} | \alpha, \tau^2) \propto \int \left[\prod_{j=1}^{n_i} \varphi(Y_{ij} | \mu_i, \sigma_\epsilon^2)\right] \varphi(\mu_i | \alpha, \tau^2)\, d\mu_i$$

$$p(Y | \mu, \sigma, \alpha, \tau^2) \propto \prod_{i=1}^{G} \left[\prod_{j=1}^{n_i} \varphi(Y_{ij} | \mu_i, \sigma_\epsilon^2)\right] \varphi(\mu_i | \alpha, \tau^2)$$

$$\log p = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \log \varphi(Y_{ij} | \mu_i, \sigma_\epsilon^2) + \log \varphi(\mu_i | \alpha, \tau^2)$$

$$= \sum_{i=1}^{G} \left[\sum_{j=1}^{n_i} -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y_{ij} - \mu_i)^2\right] - \frac{1}{2}\log\tau^2 - \frac{1}{2\tau^2}(\mu_i - \alpha)^2$$

$$N = \sum_{i=1}^{G} n_i$$

$$\ell = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i\sum_j (Y_{ij} - \mu_i)^2 - \frac{G}{2}\log\tau^2 - \frac{1}{2\tau^2}\sum_i(\mu_i - \alpha)^2$$

$$\frac{\partial}{\partial\sigma} = \frac{-N}{\sigma} + \frac{1}{\sigma^3}\sum_i\sum_j (Y_{ij} - \mu_i)^2$$

$$-\frac{1}{2}\sigma^{-2}$$

$$+\sigma^{-3}$$

$$-\frac{N}{2\sigma^2} 2\sigma$$

$$Y_{ij}^2 - 2\hat{\mu}_i Y_{ij} + \hat{\mu}_i^2 + \hat{\sigma}_{\mu_i}^2$$

$$Y_{ij}^2 - 2\mu_i \sum Y_{ij} + \mu_i^2$$

$$\bar{Y}_i^2 - 2\mu_i\bar{Y}_i + \hat{\mu}_i^2 + \hat{\sigma}_{\mu_i}^2$$

$$p(\mu_i | \sim) \propto \left[\prod_{j=1}^{n_i} \varphi(Y_{ij} | \mu_i, \sigma^2)\right] \varphi(\mu_i | \alpha, \tau^2)$$

$$(Y_{ij} - \bar{Y}_i + \bar{Y}_i - \mu_i)^2$$

$$N\left(\hat\alpha + \frac{\hat\tau^2}{\hat\sigma^2/n_i + \hat\tau^2}(\bar{Y}_i - \hat\alpha),\ \frac{\hat\sigma^2/n_i \cdot \hat\tau^2}{\hat\sigma^2/n_i + \hat\tau^2}\right)$$

$$\sum S_i^2 + (\bar{Y}_i - \mu_i)^2$$

$$+ \text{ cm}$$

TOPIC: | DATE:

FILE UNDER: | PAGE:

Ex. Mixture Models

$Y_1, \ldots, Y_n$,

$$f(y_i) = \lambda \varphi(y_i \mid \mu_1, \sigma_1^2) + (1-\lambda)\varphi(y_i \mid \mu_2, \sigma_2^2)$$

$$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda), \quad \lambda \in (0,1)$$

$$\ell(\theta \mid Y) = \sum_{i=1}^{n} \log\left\{\lambda \varphi(y_i \mid \mu_1, \sigma_1^2) + (1-\lambda)\varphi(y_i \mid \mu_2, \sigma_2^2)\right\}$$

Suppose $Z \sim \text{Bernoulli}(\lambda)$, s.t. if

$z_i = 1$, then $Y_i \sim N(\mu_1, \sigma_1^2)$

$z_i = 0$, then $Y_i \sim N(\mu_2, \sigma_2^2)$.

CDL:

$$\cancel{L(\theta \mid Y) = \prod_{i=1}^{n} \varphi(Y_i \mid \mu_1, \sigma_1^2)^{z_i} \varphi(Y_i}$$

$$f(y_i \mid z_i) = \varphi(y_i \mid \mu_1, \sigma_1^2)^{z_i} \varphi(y_i \mid \mu_2, \sigma_2^2)^{1-z_i}$$

$$p(z_i) = \lambda^{z_i}(1-\lambda)^{1-z_i}$$

$$\cancel{L(\theta) = \prod_{i=1}}$$

$$\ell(\theta) = \sum_{i=1}^{n} \log f(y_i \mid z_i) p(z_i)$$

$$= \sum_{i=1}^{n} z_i \log \varphi(y_i \mid \mu_1, \sigma_1^2) + (1-z_i)\varphi(y_i \mid \mu_2, \sigma_2^2)$$

$$+ z_i \log \lambda + (1-z_i)\log(1-\lambda)$$

$$= \sum_{i=1}^{n} z_i \varphi_1 + (1-z_i)\varphi_2 + z_i \log \lambda + (1-z_i)\log(1-\lambda)$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$\frac{b}{x} = 1 - \frac{a}{x}$

$\frac{b}{x} = \frac{x-a}{x}$

$x = a + b$

$$p(z_i|y_i) \propto p(y_i|z_i)\,p(z_i)$$

$$= \varphi_1^{z_i}\varphi_2^{1-z_i}\lambda^{z_i}(1-\lambda)^{1-z_i}$$

$$= (\lambda\varphi_1)^{z_i}(\varphi_2(1-\lambda))^{1-z_i}$$

$$= \text{Bernoulli}\left(\frac{\lambda\varphi_1}{\lambda\varphi_1 + (1-\lambda)\varphi_2}\right)$$

$$\parallel$$

$$\pi_i$$

$$E[z|y] = \pi_i = \frac{\lambda_0\varphi(y_i|\mu_{10},\tau_{10}^2)}{\lambda_0\varphi(y_i|\mu_{10},\tau_{10}^2) + (1-\lambda_0)\varphi(y_i|\mu_{20},\tau_{20}^2)}$$

$$Q(\theta|\theta_n) = E\left[\sum_{i=1}^n z_i\log\varphi_1 + (1-z_i)\log\varphi_2 + c\right]$$

$$= \sum_{i=1}^n \pi_i\log\varphi_1 + (1-\pi_i)\log\varphi_2 + \pi_i\log\lambda + (1-\pi_i)\log(1-\lambda)$$

$$= \sum_{i=1}^n \pi_i\left[-\frac{1}{2}\log 2\pi\sigma_1^2 - \frac{1}{2\sigma_1^2}(y_i-\mu_1)^2\right] + (1-\pi_i)\left[-\frac{1}{2}\log 2\pi\sigma_2^2 - \frac{1}{2\sigma_2^2}(y_i-\mu_2)^2\right]$$

$$\Rightarrow \hat{\mu}_1 = \frac{\sum \pi_i y_i}{\sum \pi_i} \qquad \hat{\sigma}_1^2 = \frac{\sum (y_i - \hat{\mu}_1)^2\pi_i}{\sum \pi_i}$$

$$\hat{\mu}_2 = \frac{\sum (1-\pi_i)y_i}{\sum \pi_i} \qquad \hat{\sigma}_2^2 = \frac{\sum (1-\pi_i)(y_i-\hat{\mu}_2)^2}{\sum (1-\pi_i)}$$

$$\hat{\lambda} = \frac{1}{n}\sum \pi_i$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

For a regular exponential family,

Complete data pdf

$$g(X|\theta) = h(x)\exp\left(\theta^T t(X)\right)/a(\theta)$$

$$\log g(x|\theta) = \cancel{\log h(x)} + \theta^T t(x) - \log a(\theta)$$

$$Q(\theta|\hat{\theta})\ \mathbb{E}(\log g(X|\theta) = \theta^T\,\mathbb{E}\left[t(x)\,|\,\hat{\theta},y\right] - \log a(\theta)$$

$$Q'(\theta|\tilde{\theta}) = \mathbb{E}\left[t(x)\,|\,\hat{\theta},y\right] - \mathbb{E}_\theta\left[t(x)\right] = 0$$

$$\Rightarrow \mathbb{E}_\theta\left[t(x)\right] = \mathbb{E}\left[t(x)\,|\,\hat{\theta},y\right]$$

Ex: $y_1, y_2, \ldots, y_m, z_{m+1}, \ldots, z_n \sim N(\mu,\sigma^2)$

$$g(X|\mu,\sigma^2) \propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2\right)$$

$$\log g(x|\mu,\sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum\left(X_i^2 - 2X_i\mu + \mu^2\right)$$

$$= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left[\sum X_i^2 - 2\mu\sum X_i + \mu^2\right]$$

$$\mathbb{E}\left(\begin{matrix}\sum x_i^2 \\ \sum X_i\end{matrix}\,\middle|\,Y\right) = \left(\begin{matrix}\sum_{i=1}^{m} Y_i^2 + \left[\tilde{\mu}^2 + \tilde{\sigma}^2\right](n-m+1) \\ \sum_{i=1}^{m} Y_i + \tilde{\mu}(n-m+1)\end{matrix}\right)$$

$$\mathbb{E}\left(\begin{matrix}\sum x_i^2 \\ \sum X_i\end{matrix}\right) = \left(\begin{matrix}(\mu^2+\sigma^2)n \\ \mu n\end{matrix}\right)$$