HW4
Colbert
Sesanker

①

1) The constraint based version that often arises in a frequentist setting: minimize $\|X\beta - y\|_2^2$ s.t. $\|\beta\|_1 \le t$

This constraint corresponds to the lagrange multiplier equation $\underset{\beta}{\text{Argmin}}\left(\|X\beta - y\|_2^2 + \lambda\|\beta_1\|\right)$, where λ is typically found via some cross validation procedure or prior knowledge. Note this is the form that arises naturally in a Bayesian setting assuming a double exponential prior. Lasso can be implemented in the constrained based form using a technique like LARS, Least Angle Regression.

b) The covariates in this model are LDL, heart rate, ...

let the design matrix $X = \text{individual}\begin{pmatrix} — X_1 — \\ \vdots \\ — X_n — \end{pmatrix} = \begin{pmatrix} LDL_1, \dots \\ \vdots \\ LDL_n, \dots \end{pmatrix}$

Solve for β: $\underset{\beta}{\text{Argmin}}\left(\|X\beta - y\|_2^2 + \lambda\|\beta\|\right)$, use crossvalidation for λ

c) WRT λ you could use a cross-validation scheme

d) You could link the response via $\mu = 1/(1+\exp(-X\beta))$ using logistic regression: $Y \sim Bin(\mu)$

e) You can use L_2 penalized regression, this assumes covariates are correlated and the matrix X^TX is ill-conditioned.

2) Random effects refer to population specific effects that don't represent population averages (fixed effects)

b) Genetic background is a random effect modeled as a random polygenic term.

c) $\langle FDR \rangle = FP/(FP+TP)$, Power $= p(T(x) \in R \mid H_0 = \text{False})$
R = reject region
$T(x)$ = test statistic

d) when the causatives are in the data: The power of MM and MLMM increases sharply then levels out, this sharp increase becomes slower (reverse for LM) as h^2 increases. LM follows a straight line for $h^2 = .25$ then the increase in power becomes more sensitive the increase in FDR.

when causatives dropped from data has a similar effect to increasing h^2. LM becomes more sensitive to FDR for large FDR. MM / MLMM power responds to FDR more linearly

3) b) $$p(x,z) = \prod_{i=1}^{2} \pi_i^{z_i} \mathcal{N}(x \mid \mu_i, \Sigma_i)^{z_i} \propto \prod_i e^{z_i \ln(\pi_i)} \exp\left(-\tfrac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$$ $(\pi_2 = 1-\pi_1,\ \pi_1 = \pi)$

$$= \exp\left(\sum_i z_i \ln(\pi_i) - \tfrac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)\right), \text{ thus } \prod_j^n p(x_j, z_j) =$$

$$= L(\theta; X, Z) = \exp\left[\sum_j^n \left(\sum_i^2 z_i \ln(\pi_i) - \tfrac{1}{2}(x_j - \mu_i)\Sigma_i^{-1}(X_j - \mu_i)\right)\right]$$

$Q(\theta \mid \theta^t) = \int_Z L(\theta; X, Z)\, p(Z \mid X, z)\, dZ$, (I like to think of this as marginalizing out Z, not computing an Expectation)

We can compute the posteriors as: $\tau_{ij}^t = P(Z_i = 1 \mid X_i = x_i, \theta^t) =$

$$Q(\theta \mid \theta^t) = \sum_j^n \sum_i^2 \tau_{j,i}^t \left(z_i \ln(\pi_i) - \tfrac{1}{2}(X_j - \mu_i)\Sigma_i^{-1}(X_j - \mu_i)\right)$$

$$\frac{\pi^t \mathcal{N}(x_i \mid \mu_i^t, \Sigma_i^t)}{\pi^t \mathcal{N}(x_i \mid \mu_1^t, \Sigma_1^t) + (1-\pi)^t \mathcal{N}(x_i \mid \mu_2^t, \Sigma_2^t)}$$

n observations

* missing $\frac{z_i}{2} \ln(|\Sigma^{-1}|)$ term in all exponentials above

3 b) $\frac{d}{d\theta} Q(\theta|\theta^*) = 0 \rightarrow \begin{cases} \frac{d}{d\pi} Q(\theta|\theta^*) = 0 & (1) \\ \frac{d}{d\mu} Q(\theta|\theta^*) = 0 & (2) \\ \frac{d}{d\Sigma} Q(\theta|\theta^*) = 0 & (3) \end{cases}$

(1) implies $\arg\max_{\pi} (Q(\theta|\theta^*) = \arg\max_{\pi} \left(\left[\sum_i \tau_{i,1}^t\right] \log \pi + \left[\sum_i (1-\tau_{i,1}^t)\right] \log(1-\pi) \right)$

is equivalent to the MLE for the binomial

thus $\pi_i^t = \dfrac{\sum_j^n \tau_{i,j}^t}{\sum_j^n (\tau_{i,j}^t - (1-\tau_{i,j}^t))} = \frac{1}{n} \sum \tau_{i,j}^t$

$\arg\max_{\mu,\Sigma} (Q(\theta|\theta^*)) = \arg\max_{\mu,\Sigma} \sum_i \tau_{i,1}^t \left(-\frac{1}{2}\ln|\Sigma^{-1}| - \frac{1}{2}(x_i-\mu_1)^T \Sigma^{-1}(x_i-\mu_1)\right)$

This is the MLE equation for multivariate gaussian (weighted)

so $\mu_1^{t+1} = \dfrac{\sum_i \tau_{i,1} x_i}{\sum_i \tau_{i,1}}$, $\Sigma_1^{t+1} = \dfrac{\sum_i \tau_{i,1} \left(\bigotimes_{j=1}^{2} (x_i-\mu_1)_j\right)}{\sum_i \tau_{i,1}}$

$\mu_2^{t+1}, \Sigma_2^{t+1}$ same as above, with 2 replacing 1

$\otimes$ is the outer product between vectors, this is what I use in my R code

c) EM gives a generative model and finds the elliptical shape, K-means is faster but cant use information about the shape of the clusters.