0 Syllabus

- There is no required homework for this course, but it is recommended that you do the 2011-2022 applied qualifying exams for practice.
- Each session will have two parts. First we will review a topic. Next we will discuss the solutions to a particular year's qual.
- A tentative schedule is shown in Table 1. I'll update this schedule. We'll have to re-schedule one of the classes because of the Fourth of July holiday. The topics can also be adjusted based on your preferences.
- In the two weeks before your exams, you should take the 2022 quals in exam conditions. We'll schedule a time to go over the applied exam together.
- Many of you will find that you don't have time to fully write up solutions to every past qual you are asked to do. This is ok, but you will get a lot of reading every exam and at least writing down a sketch of the answer.
- I will update this document as we go along. I will follow Dan Kluger's notes which are already on Canvas.

Session	Date	Review Material	Past qualifying exam
1	6/27	Advice and sample problems	
2	6/29	Linear models	2011
3	7/03	Exponential families and GLMs	2012
4	7/06	The bootstrap	2013
5	7/11	EM	2014
6	7/13	Cross validation	2015
7	7/18	Linear models: additional topics	2016
8	7/20	Bayesian modelling	2017
9	7/25	CAVI	2018
10	7/27	PCA	2019
11	8/01	Survival Analysis	2020
12	8/03	Optimization and numerical linear algebra	2021
13	8/08	Divergence	
No Class	8/10	Come to the ice-cream social!	
14	8/14	Solutions to last-years exam	**2022**

Table 1: Coaching schedule for the applied qualifying exam. This schedule is open to suggestions. The **2022** qual should be done in exact exam conditions. We will schedule a time to go over the 2022 applied exam together.

1 Review: Optimization 101 [1]

Recognizing problems as convex optimization problems is a useful skill for the applied qualifying exam. A convex optimization problem is something of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \text{ for } i = 1, \ldots, k, \\ & Ax = b, \end{array} \tag{1}$$

such that the following hold:

- The objective $f : \mathbb{R}^n \to \mathbb{R}$ is convex.
- Each constraint $g_i : \mathbb{R}^n \to \mathbb{R}$ is convex.
- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

All of these conditions must hold for (1) to be a convex optimization problem. If you have non-linear equality constraints or non-convex inequality constraints, then you do not have a convex optimization problem.

If you can turn your exam question into solving (1), then you are done. Convex optimization can be done quickly and reliably, and you don't need to worry about the details. You should know a handful functions that are convex. The following can be helpful for the applied exam,

- Exponential family cumulant functions: If $g(y) = \exp(\eta^\top y - \psi(\eta))$ is a d dimensional exponential family, then $\psi : \mathbb{R}^d \to \mathbb{R}$ is convex.
- Log-sum-exp: the function $f(x) = \log\left(\sum_{i=1}^n \exp(x_i)\right)$ is convex on $\mathbb{R}^n$.
- Log-normal CDF: if Φ is the CDF for the standard normal distribution, then $-\log \Phi(x)$ is convex on $\mathbb{R}$.
- $-f(x)$ is concave if and only if $f(x)$ is convex.
- If $x \mapsto f(x)$ is convex, then $y \mapsto f(By + c)$ is convex for all matrices B and vectors c.

Sometimes you will be asked about the implementation details of a convex optimization problem. When this happens, it's typically for problems without constraints. The algorithms you might need to mention are

- **Gradient descent**: This applies when the loss function is differentiable.
- **Newton's method**: This applies when the loss function is twice differentiable. This method is used to fit the MLE in GLMs.

[1] Nikos Ignatiadis, Dan Kluger and M.H.

- **Coordinate descent**: This can be used for non-smooth problems like the LASSO.

You might also have to solve non-convex optimization problems. In this case you may have to use EM (Section **??**), CAVI (section **??**) or gradient descent. Since these methods find local-minima, you should mention doing multiple initialization. If the problem is low-dimensional, then you can suggest grid search which will find the global minima.

2 Numerical linear algebra[2]

[2]Stephen Bates, Dan Kluger and M.H.

3 Applied 2020: Solution[3]

Problem 1: Maximum Likelihood for Truncated Data

a) Both `t.test(Y)` and `lm(Y ~ 1)` compute the mean and standard deviation Y. This is not possible without observing the full dataset. By default, both of these functions will simply ignore the missing values. This will cause us to overestimate the mean of Y and produce a confidence interval that is biased upwards.

b) The methods in part a) give estimates that have a positive bias. To derive a likelihood based method let $Z_1, \dots, Z_n$ denote the latent values of the light emissions. We are told that $Z_i \sim N(\mu, \sigma^2)$ for some unknown values μ and σ^2, and we observe

$$Y_i = \begin{cases} Z_i, \text{ if } Z_i > C \\ \text{NA, otherwise} \end{cases} .$$

Let Φ and ϕ denote the cdf and pdf of $N(0,1)$, respectively. Let $\delta_i := \mathbb{1}_{Y_i \neq NA}$. Then, the likelihood of the observed data is

$$p(Y; \mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma} \phi\left(\frac{Y_i - \mu}{\sigma} \right) \right)^{\delta_i} \Phi\left(\frac{C - \mu}{\sigma} \right)^{1-\delta_i}$$

and, up to a constant, the log-likelihood is

$$\ell(\mu, \sigma) = \sum_{i=1}^{n} -\delta_i \left(\log(\sigma) - \frac{1}{2} \left(\frac{Y_i}{\sigma} - \frac{\mu}{\sigma} \right)^2 \right) + (1 - \delta_i) \log \Phi\left(\frac{C}{\sigma} - \frac{\mu}{\sigma} \right).$$

This is not a concave function of (μ, σ), but it is concave in the transformed parameter $\lambda = \frac{1}{\sigma}$ and $\nu = \frac{\mu}{\sigma}$. In these parameters, we have

$$\ell(\lambda, \nu) = \sum_{i=1}^{n} \delta_i \left(\log(\lambda) - \frac{1}{2} \left(\lambda Y_i - \nu \right)^2 \right) + (1 - \delta_i) \log \Phi\left(\lambda C - \nu \right).$$

This function is concave because is non-negative combination of concave functions. We can thus use Newton's method to find $\hat{\lambda}, \hat{\nu}$ which we can transform to get $\hat{\sigma} = \frac{1}{\hat{\lambda}}$ and $\hat{\mu} = \frac{\hat{\nu}}{\hat{\lambda}}$.

If you do not recognize the above transformation, a good alternative is to use EM. In the E-Step we would need to compute

$$\mathbb{E}_{\hat{\mu}^t, \hat{\sigma}^t}[\log(p(Z; \mu, \sigma)) | Y].$$

By expanding $\log(p(Z; \mu, \sigma))$ we will find that it is sufficient to compute

$$\mathbb{E}_{\hat{\mu}^t, \hat{\sigma}^t}[Z_i | Y_i] \quad \text{and} \quad \mathbb{E}_{\hat{\mu}^t, \hat{\sigma}^t}[Z_i^2 | Y_i].$$

[3] Isaac Gibbs, Dan Kluger and M.H.

If $Y_i \neq NA$ then $\mathbb{E}[Z_i|Y_i] = Y_i$ and $\mathbb{E}[Z_i^2|Y_i] = Y_i^2$. Otherwise, computing these expectations reduces to computing the mean and variance of a truncated Gaussian, which is a straightforward computation that we could carry out. Finally, in the M-step we have to optimize $\mathbb{E}_{\hat{\mu}^t,\hat{\sigma}^t}[\log(p(Z;\mu,\sigma))|Y]$ over μ and σ, which is a straightforward two dimensional calculus problem which yields:

$$\hat{\mu}^{t+1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\hat{\mu}^t,\hat{\sigma}^t}[Z_i|Y_i] \quad \text{and} \quad (\hat{\sigma}^{t+1})^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{\hat{\mu}^t,\hat{\sigma}^t}[Z_i^2|Y_i] - 2\hat{\mu}^{t+1}\mathbb{E}_{\hat{\mu}^t,\hat{\sigma}^t}[Z_i|Y_i] + (\hat{\mu}^{t+1})^2\right)$$

c) Intuitively we should have that the observed data $Y_1, \ldots, Y_n$ come from a truncated normal distribution. Thus, we could maximize the likelihood

$$p(Y;\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma}\phi\left(\frac{Y_i - \mu}{\sigma}\right)\left(1 - \Phi\left(\frac{C-\mu}{\sigma}\right)\right)^{-1}.$$

The log of this likelihood is no longer concave in $\lambda = \frac{1}{\sigma}$ and $\nu = \frac{\mu}{\sigma}$, and so we have to use an inexact method such a gradient descent with multiple initialization. Another option would be to do a grid search over (μ, σ).

Without doing any rigorous calculations (assuming both devices take measurements for the same amount of time) we suspect that the device in b) should give a better estimate of μ and σ than the device in c). This is due to the fact that in b) we observe strictly more data than in c).

Aside: The question explicitly says that we do not need to do a rigorous calculation. However, seeing a rigorous calculation could still be informative. This calculation shows that the likelihood written out above is actually the conditional likelihood for the Y_i, conditional on the number of observations that we observe.

The calculation proceeds as follows. Let $Z_1, \ldots, Z_m$ denote the latent non-truncated measurements of light brightness. Then, we observed $Z_{i_1}, \ldots, Z_{i_n}$ where $1 \leq i_1 < i_2 < \cdots < i_n \leq m$ are the indices at which $Z_{i_j} > C$. Note that n is itself random here. Now, for any $x > C$ and $1 \leq i \leq n$ we have that

$$\begin{aligned}\mathbb{P}(Y_i > x|n = k) = \sum_{i_1,\ldots,i_k} &\mathbb{P}(Y_i > x|n = k, Z_{i_1}, \ldots, Z_{i_k} > C, Z_j \leq C, \forall j \notin \{i_1, \ldots, i_k\}) \\ &\cdot \mathbb{P}(Z_{i_1}, \ldots, Z_{i_k} > C, Z_j \leq C, \forall j \notin \{i_1, \ldots, i_k\}|n = k).\end{aligned}$$

Given fixed values for the indices $i_1, \ldots, i_k$ let j^* be the index such that $Y_i = Z_{i_{j^*}}$.

Then,

$$\begin{aligned}
&\mathbb{P}(Y_i > x | n = k, Z_{i_1}, \ldots, Z_{i_k} > C, Z_j \leq C, \forall j \notin \{i_1, \ldots, i_k\}) \\
&= \frac{\mathbb{P}(Z_{j^*} > x, Z_{i_1}, \ldots, Z_{i_k} > C, Z_j \leq C, \forall j \notin \{i_1, \ldots, i_k\})}{\mathbb{P}(Z_{i_1}, \ldots, Z_{i_k} > C, Z_j \leq C, \forall j \notin \{i_1, \ldots, i_k\})} \\
&= \frac{\left(1 - \Phi\left(\frac{x-\mu}{\sigma}\right)\right) \Phi\left(\frac{C-\mu}{\sigma}\right)^{m-k} \left(1 - \Phi\left(\frac{C-\mu}{\sigma}\right)\right)^{k-1}}{\Phi\left(\frac{C-\mu}{\sigma}\right)^{m-k} \left(1 - \Phi\left(\frac{C-\mu}{\sigma}\right)\right)^{k}} \\
&= \frac{1 - \Phi\left(\frac{x-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{C-\mu}{\sigma}\right)}.
\end{aligned}$$

Differentiating this last expression gives the claimed formula for the likelihood.

d) We are told to replace μ in the above likelihoods with $\mu_i = X_i^\top \beta$. Since we are given a function for the gradient of the log-likelihoods we can use gradient ascent to maximize the objective. Assuming that σ^2 is fixed and known the updates would look like

$$\beta_{t+1} = \beta_t + \eta \text{grad\_log}(Y, X\beta_t, \text{sigma\_sq}, C)^T X, \tag{2}$$

where $\eta > 0$ is a fixed step-size parameter. If σ^2 is unknown then we would still use the update (2), but in this update we would only use the first n coordinates of $\text{grad\_log}(Y, X\beta_t, \text{sigma\_sq}, C)$, we would replace sigma_sq with σ_t^2, and we would additionally have the update

$$\sigma_{t+1} = \sigma_t + \eta \Big[\text{grad\_log}(Y, X\beta_t, \sigma_t^2, C)\Big]_{n+1}.$$

Finally, we could use the function logl(Y,mu,sigma_sq,C) to judge the results of gradient ascent with multiple different random initializations.

Problem 2: Low-Rank Matrix Factorization

a) The goal is to find matrices $A \in \mathbb{R}^{T \times 13}$ and $B \in \mathbb{R}^{990 \times 13}$ that minimize $||X - AB^T||_2^2$. We know that the skinny SVD optimizes this objective. Namely, let

$$X = \sum_{i=1}^{990} \sigma_i u_i v_i^T$$

denote the SVD of X where $|\sigma_1| \geq \cdots \geq |\sigma_{990}| \geq 0$. Then, we can take

$$\hat{A} = U_{1:13} \text{diag}(\sigma_1, \ldots, \sigma_{13}) \quad \text{and} \quad \hat{B}^T = (V_{1:13})^T$$

where $U_{1:13}$ and $V_{1:13}$ denotes the matrices with columns $u_1, \ldots, u_{13}$ and $v_1, \ldots, v_{13}$, respectively.

b) The individual matrices A and B will not be unique. Namely, given an optimal solution (A, B) we have that for any orthogonal matrix O, $AB^T = AOO^TB^T$ and thus (AO, BO) is also a solution. On the other hand, AB^T will be unique iff $\sigma_{13} > \sigma_{14}$ since in this case the skinny SVD is unique.

c) Let $A_1, \ldots, A_T$ denote the rows of A and $B_1, \ldots, B_{990}$ denote the rows of B. Then,

$$||X - AB^T||_2^2 = \sum_{i=1}^{T} \sum_{j=1}^{990} (X_{ij} - B_j^T A_i)^2.$$

For each fixed value of $i \in \{1, \ldots, T\}$ we recognize $\sum_{j=1}^{990}(X_{ij} - B_j^T A_i)^2$ as a linear regression problem with feature-response pairs $\{(B_j, X_{ij})\}_{1 \leq j \leq 990}$. Using known results from linear regression this expression will be minimized by taking

$$A_i = (B^T B)^{-1} B^T x_i.$$

By the separability of the objective function in i it follows that for a fixed B, A is optimized when setting $A^T = (B^T B)^{-1} B^T X^T$. Note that this assumes that B is full-rank. If it is not full rank then you can replace the inverse by a pseudo-inverse.

d) Using the same reasoning as in part c) we have that the minimizing value for B will be given by

$$B^T = (A^T A)^{-1} A^T X.$$

Note that this assumes that A is full-rank. If it is not full rank then you can replace the inverse by a pseudo-inverse.

e) Letting $x_{T+1,\mathcal{O}} \in \mathbb{R}^{990-45}$, denote the vector of observations at time $T+1$ with the entries dropped and let $B_{\mathcal{O}} \in \mathbb{R}^{(990-45)\times 13}$ denote the matrix B with the rows which correspond to missing entries of x_{T+1} removed. We could estimate a_{T+1} be setting

$$\hat{a}_{T+1} = \arg\min_a ||x_{T+1,\mathcal{O}} - B_{\mathcal{O}} a||_2^2,$$

As in parts c) and d) solving this optimization problem is equivalent to solving a linear regression problem and obtain that $\hat{a}_{T+1} = (B_{\mathcal{O}}^T B_{\mathcal{O}})^{-1} B_{\mathcal{O}}^T x_{T+1,\mathcal{O}}$. Then, we could estimate the missing values in the full vector by looking at the corresponding entries of $\hat{x}_{T+1} = B\hat{a}_{T+1}$.

Problem 3: Poisson GLMs

a) You can take X to be the matrix with n_i copies of row x_i and $Y \in \mathbb{R}^N$ to be the vector with entries y_{ij} and then use the R call glm($Y \sim X - 1$, family=poission(link="log")).

b) The likelihood for the data is

$$p(Y; X, \beta) = \prod_{ij} \frac{\exp(y_{ij} x_i^T \beta - \exp(x_i^T \beta))}{y_{ij}!} \propto \prod_i \exp(\sum_j y_{ij} x_i^T \beta - n_i \exp(x_i^T \beta)).$$

Additionally, note that $T_i := \sum_j y_{ij} \sim \text{Poisson}(n_i \exp(x_i^T \beta))$. So, the likelihood for $T_1, \ldots, T_I$ is

$$p(T_1, \ldots, y_I; x_1, \ldots, x_I, \beta) \propto \prod_i \exp(T_i x_i^T \beta - n_i \exp(x_i^T \beta)).$$

In particular, we find that the likelihood for the data $\{T_i\}$ is proportional to the likelihood for $\{y_{ij}\}$. Thus, maximum likelihood based inference will be the same for these two datasets. Under, the current model we have that the mean of T_i is $\mu_{i\cdot}$ with

$$\log(\mu_{i\cdot}) = \log(n_i) + x_i^T \beta$$

We recognize this as a GLM with offsets. Let $T = (T_1, \ldots, T_I) \in \mathbb{R}^I$, $\tilde{X} \in \mathbb{R}^{I \times p}$ be the matrix with rows x_i, and $n = (n_1, \ldots, n_I)$. Then, we find that the model from part a) can be fit using the GLM call glm($T \sim \tilde{X} - 1$,family=poisson(link="log"),offset=log(n)).

Note that your justification above need not explicitly write out the likelihood and could use a sufficient statistics argument instead.

c) Extending our previous model it may now be reasonable to posit that the events for galaxy i come from a homogeneous Poisson process with mean parameter $\exp(x_i^T \beta)$. i.e. we have that y_{ij} are independent Poisson random variables with

$$\log(\mu_{ij}) = \log(\ell_{ij}) + x_i^T \beta.$$

We can fit this model using the call glm($Y \sim X - 1$, family=poission(link="log"),offset $= \log(\ell)$) where ℓ is the vector of lengths and Y and X are as in part a).

Problem 4: Estimating Starfish Diversity

a) S_i will be larger for less diverse antibody pools. One way to see this is to observe that

$$S_i = \sum_{j=1}^J \hat{p}_{ij}^2 \leq \sum_{j=1}^J \hat{p}_{ij} = 1,$$

with equality for vectors $\hat{p}_{i\cdot}$ that satisfy $\hat{p}_{ij} = 1$ for some j and $\hat{p}_{ij'} = 0$ for all other $j' \neq j$. i.e. S_i is maximized by the least diverse antibody pools. Moreover, by Jensen's inequality we have that

$$S_i = J \sum_{j=1}^J \frac{\hat{p}_{ij}^2}{J} \geq J \left(\sum_{j=1}^J \frac{\hat{p}_{ij}}{J} \right)^2 = \frac{1}{J}$$

with equality being obtained by the uniform distribution. i.e. S_i is minimized by the most diverse populations.

b) Perhaps the largest issue with S_H is that its value can be dominated by a single starfish that has very large counts for all antibodies. This is a major concern because we are told that "the absolute numbers n_{ij} depend a lot on how the sample was taken, and so the ratios are considered more useful." For a concrete example, suppose 9/10 of the starfish collected have very diverse antibodies and a relatively small total number of antibodies measured and the final remaining starfish does not have very diverse antibodies, but has a large total number of antibodies. Then, the value of S_H will be dominated by the one non-diverse starfish and we will estimate that the population does not have a very diverse set of antibodies even though 9/10s of the starfish have a large diversity. Cases like this where some samples have larger total counts than others are common in many types of biological data (e.g. RNA-seq).

   A better method would be to weight all of the starfish equally regardless of the magnitude of their total counts. To do this we could define

$$S_{H,i} = \sum_{j=1}^{J} \left(\frac{n_{ij}}{\sum_{j'=1}^{J} n_{ij'}} \right)^2$$

   to be the diversity measure for the i_{th} high-salinity starfish and then define the new estimator

$$\tilde{S}_H = \frac{1}{10} \sum_{i=1}^{10} S_{H,i}.$$

   We could do the same thing for the low salinity starfish and compute the estimate $\tilde{S}_H - \tilde{S}_L$.

c) Their bootstrap procedure does not appear to reflect the data generating mechanism and account for the clustered nature of their data. In particular, their bootstrap procedure treats each observed antibody in the high-salinity water as a sample, and samples the $\sum_j n_{H,j}$ antibodies with replacement. It does not account for the fact that the antibodies were measured by taking sampling antibodies in 10 different starfish (each of which could have different proportions of each antibody).

   A better approach would be to use clustered bootstrap which resamples entire starfish with replacement but does not resample data within starfish (See Strategy 1 of Section 3.8 in Davison and Hinkley [1997]). You can think of this as a block bootstrap where each starfish is a block. More specifically, for each $b = 1, \ldots, B$ we would obtain bootstrap datasets of starfish-level diversity scores $\{S^b_{H,i}\}_{1 \le i \le 10}$ and $\{S^b_{L,i}\}_{1 \le i \le 10}$ where $S^b_{H,i} \sim \text{Unif}(S_{H,1}, \ldots, S_{H,10})$ and

$S^b_{L,i} \sim \text{Unif}(S_{L,1}, \ldots, S_{L,10})$. We would then use these datasets to get new estimates of the difference in means given by

$$\tilde{S}^b_H - \tilde{S}^b_L = \frac{1}{10}\sum_{i=1}^{10} S^b_{H,i} - \frac{1}{10}\sum_{i=1}^{10} S^b_{L,i}$$

and then form a confidence interval by looking at the empirical quantiles of $\{\tilde{S}^b_H - \tilde{S}^b_L\}_{1\le b\le B}$ (see Section **??**).

An alternative to the previous procedure is to use a clustered bootstrap, where you sample starfish with replacement, and subsequently resample the antibodies within each starfish with replacement (See Strategy 2 of Section 3.8 in Davison and Hinkley [1997]). In particular, for $b = 1, \ldots, B$, to construct the bootstrap dataset for the high-salinity starfish, sample 10 high-salinity with replacement $i_1, \ldots, i_{10} \overset{IID}{\sim} \text{Unif}\{1, \ldots, 10\}$ then for $k = 1, \ldots, 10$, compute $S^b_{H,k}$, by resampling the antibodies from starfish i_k with replacement (this can be done by sampling a multinomial on $\{1, \ldots, J\}$ with probabilities $\hat{p}_{i_k,j}$ with $\sum_j n_{i_kj}$ trials for the multinomial) and then using the resampled antibody counts in starfish i_k to compute the diversity score $S^b_{H,k}$. You would do the same procedure to generate bootstrap samples of the diversity scores $S^b_{L,k}$ in the Low salinity waters. Finally you would let the bth bootstrap statistic be

$$\tilde{S}^b_H - \tilde{S}^b_L = \frac{1}{10}\sum_{k=1}^{10} S^b_{H,k} - \frac{1}{10}\sum_{k=1}^{10} S^b_{L,k},$$

and use the empirical quantiles of $\{\tilde{S}^b_H - \tilde{S}^b_L\}_{1\le b\le B}$ to compute a bootstrap confidence interval.

While this is out of scope for quals, Section 3.8 in Davison and Hinkley [1997] recommends Strategy 1 over Strategy 2, but either would be an acceptable answer for quals.

Problem 5: EM for a Mixture of Regressions

a) We have that

$$\log(p_\theta(Y)) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - a_j - b_j X_i\right)^2\right)\right)$$

and

$$\begin{aligned}\log(p_\theta(Y,Z)) &= \sum_{i=1}^n \log\left(\pi_{Z_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(Y_i - a_{Z_i} - b_{Z_i}X_i\right)^2\right)\right)\\ &= \sum_{i=1}^n \log(\pi_{Z_i}) - \frac{1}{2\sigma^2}\left(Y_i - a_{Z_i} - b_{Z_i}X_i\right)^2 - \frac{1}{2}\log(2\pi\sigma^2)\\ &= \sum_{i=1}^n\left(\sum_{j=1}^k\left(\log(\pi_j)Z_{ij} - \frac{1}{2\sigma^2}\left(Y_i - a_j - b_jX_i\right)^2 Z_{ij}\right) - \frac{1}{2}\log(2\pi\sigma^2)\right).\end{aligned}$$

There are k independent parameters for the a_j's, k independent parameters for the b_j's, $k-1$ independent parameters for the π_j, and one parameter for σ for a total of $3k$ independent parameters.

b) The notation here isn't very good. In order to run EM we will need to compute $\mathbb{E}_{\tilde{\theta}}[\log(p_\theta(Y,Z))]$ not $\mathbb{E}_\theta[\log(p_\theta(Y,Z))]$ where $\tilde{\theta}$ and θ are two potential values for the fitted parameters. Thus, I will re-define $\tau_{ij} = \mathbb{P}_{\tilde{\theta}}(Z_i = j|Y)$. This gives the expression

$$\mathbb{E}_{\tilde{\theta}}[\log(p_\theta(Y,Z))] = \sum_{i=1}^n\left(\sum_{j=1}^k\left(\log(\pi_j)\tau_{ij} - \frac{1}{2\sigma^2}\left(Y_i - a_j - b_jX_i\right)^2\tau_{ij}\right) - \frac{1}{2}\log(2\pi\sigma^2)\right). \tag{3}$$

c) The description of EM given in the problem statement is quite bad. The expression $\mathbb{E}_{\theta,\tau}[\cdot]$ doesn't really make any sense and should be $\mathbb{E}_{\hat{\theta},\tau}[\cdot]$. Regardless we will solve the problem ignoring the notational issues using the correct version of the EM algorithm. The goal is to optimize (3) over (a,b,π,σ) (strangely the question does not ask us to optimize π, but π is part of θ so we certainly need to optimize over it). Optimizing over π is a straightforward Lagrange multipliers calculation from which you should find that

$$\hat{\pi}_j = \frac{\sum_{i=1}^n \tau_{ij}}{\sum_{k=1}^K\sum_{i=1}^n \tau_{ik}} = \frac{\sum_{i=1}^n \tau_{ij}}{n},$$

where one easily checks that from the definitions we must have that $\sum_{j=1}^K \tau_{ij} = 1$ for all i. Optimizing a_j and b_j splits into K standard weighted least squares problems. Let $T_j = \mathrm{diag}(\tau_{1j},\ldots,\tau_{nj})$ and let $X = \begin{bmatrix}\mathbf{1} & (X_i)_{i=1}^n\end{bmatrix}$. Then, we have that

$$\begin{aligned}\begin{pmatrix}\hat{a}_j\\ \hat{b}_j\end{pmatrix} &= (X^TT_jX)^{-1}X^TT_jY\\ &= \frac{1}{(\sum_{i=1}^n \tau_{ij}X_i^2)(\sum_{i=1}^i \tau_{ij}) - (\sum_{i=1}^n \tau_{ij}X_i)^2}\\ &\quad\cdot\begin{pmatrix}(\sum_{i=1}^n \tau_{ij}X_i^2)(\sum_{i=1}^i \tau_{ij}Y_i) - (\sum_{i=1}^n \tau_{ij}X_i)(\sum_{i=1}^n \tau_{ij}X_iY_i)\\ (\sum_{i=1}^n \tau_{ij})(\sum_{i=1}^i \tau_{ij}X_iY_i) - (\sum_{i=1}^n \tau_{ij}X_i)(\sum_{i=1}^n \tau_{ij}Y_i)\end{pmatrix}\end{aligned}$$

Finally, by differentiating in terms of σ^2 and re-arranging one can easily compute that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(Y_i - \hat{a}_j - \hat{b}_j X_i)^2 \tau_{ij}.$$

d) The correct way to compute BIC is to use the **observed** data log-likelihood. This gives the value

$$\text{BIC} = 3k\log(n) + 2 \cdot 116.36 = 9 \cdot \log(39) + 2 \cdot 116.36 \approx 265.7.$$

e) In our setting there are $n = 39$ samples and $k = 3K$ independent parameters, so we can use the definition of AIC and BIC to compute the AIC and BIC values for $K \in \{1, \ldots, 5\}$ in the table presented below. AIC is minimized at $K = 5$ so AIC selects $K = 5$ (or it could select $K > 5$ depending on un-presented log likelihood values) . On the other hand the BIC is minimized at and selects $K = 4$. There are many different alternative ways to select the number of clusters. One option is to use information in the forestry literature to set a prior on K and also a prior on $\theta \mid K$ and then compute the mode of the posterior distribution of K given the data.

Method	Formula	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
AIC	$6K - 2\log p_{\hat{\theta}}(Y)$	294.08	270.38	250.72	241.56	239.16
BIC	$3K\log(39) - 2\log p_{\hat{\theta}}(Y)$	299.0707	280.3614	265.6921	261.5227	264.1134

Problem 6: Estimating the Mutation Rate

There are many possible models you could consider. Two obvious choices are a Poisson model and a linear model. In what follows I will answer parts a, b, and c separately for each of these two choices.

Solution using a Poisson model

One potential model is that the $d_{i,r}$ values are independent with

$$d_{i,r}|t_i - t_r \sim \text{Poisson}((t_i - t_r)\mu).$$

One nice aspect of this model is that the variance of $d_{i,r}$ scales linearly in the mean, which seems consistent with what we observe in the plot of the data. A straightforward calculation shows that this gives the maximum likelihood estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} d_{i,r}}{\sum_{i=1}^{n} t_i - t_r}.$$

This estimator is unbiased for μ and since it is a sum of independent random variables it should also be consistent under only mild assumptions on the times $t_i - t_r$. To do inference for μ we could form the standard confidence interval for the MLE. This would be valid as long as the Poisson model is accurate. If we want to avoid this parametric assumption an alternative is to use the fact that $\hat{\mu}$ is unbiased and an average of independent random variables and thus directly derive a CLT for $\hat{\mu}$. The main thing we will need here is an estimate of the variance of $\hat{\mu}$. We have that

$$\text{Var}(\hat{\mu}) = \frac{1}{(\sum_{i=1}^n t_i - t_r)^2} \sum_{i=1}^n \text{Var}(d_{i,r}).$$

We know that $\text{Var}(d_{i,r}^2) = \mathbb{E}[(d_{i,r} - (t_i - t_r)\mu)^2]$. Since $\hat{\mu}$ is consistent for μ a reasonable way to estimate this quantity is using the estimator

$$\hat{\sigma}^2 = \frac{1}{(\sum_{i=1}^n t_i - t_r)^2} \sum_{i=1}^n (d_{i,r} - (t_i - t_r)\hat{\mu})^2.$$

Finally we can use the normal approximation $\hat{\mu} \dot{\sim} N(\mu, \hat{\sigma}^2)$ to get a confidence interval for μ.

One flaw in our model is that it assumes a Poisson distribution for the mutations. However, above we derived a non-parameteric method for computing a confidence interval for μ and thus this modelling assumption is not critical. A potentially larger issue is the presence of outliers in the dataset. In particular, we see that at some of the larger time points there are a few outlying points with very large numbers of mutations (note that a Pois(10) Random variable is 40 or larger with probability less that 10^{-12}). These points could have a large influence on the estimator $\hat{\mu}$ and thus heavily impact our estimate of the mutation rate. As a first step we should investigate how the data was collected and try to determine if there were any possible errors in the data collection process that could lead to these outlying points. Alternatively, maybe there is a biological mechanism that would tell us that some rare cases will deviate from the Poisson model presented. In both of these cases we might want to remove these outlying points from the dataset before fitting the model. If we think these outlying points are true data points that cannot be ignored then we could still use the above estimator $\hat{\mu}$. However, we should be conscious of the fact that these points may have an outsized influence on the estimator that will invalidate our normal approximation. To try to judge the size of the influence of these points we could compute $\hat{\mu}$ both with and without removing the outliers and report both values.

Another issue with our model assumption is the assumption of independent observations. In particular, it could be that many observations had a recent ancestor that is also in the data. For example, an observation in January 2020 may be correlated with its direct descendants in the data (if a virus observed in January 2020 has a relatively high hamming distance from from the original strain, it is likely the descendants of that virus also have a relatively high hamming distance from the original strain). It

is hard to tell from the information given whether there will be a lot of dependency between the samples without domain knowledge or additional information beyond $d_{r,i}$, t_i and t_r. That being said, even if the samples are correlated, the proposed estimator for μ should still be unbiased and consistent, but the confidence intervals would likely be too small.

Solution using a linear model

An alternative to fit a linear model

$$d_{i,r} = (t_i - t_r)\mu + \epsilon_i.$$

Intuitively, it is reasonable to expect that the variance of ϵ_i will be larger for larger times as there is more time for mutation events to occur and thus more time to accumulate deviations from the mean. From the plot of the data it looks like a reasonable model would be that $\text{Var}(\epsilon_i)$ increases linearly with time. Thus, we could model that $\text{Var}(\epsilon_i) = (t_i - t_r)\sigma^2$. Using weighted least squares this would give the estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} d_{i,r}}{\sum_{t=1}^{n} t_i - t_r},$$

which is the same as the estimator from the Poisson model. All the considerations from the previous section still apply to this estimator. Here it is even more clear that we should not make standard modelling assumptions like $\epsilon_i \sim N(0, (t_i - t_r)\sigma^2)$ since we know that the $d_{i,r}$ are discrete. The normal approximation given in the previous section could still be a good way to form a confidence interval for μ.

4 Applied 2021: Solution[4]

Key Ideas/ Main Tools: Unmeasured confounders, GLMs, offsets

Problem 1: Modeling association between vaccination and death rates for Covid-19

a) It is essential to ask the researchers whether they know the population in each county because counties with higher populations will tend to have a higher death count irrespective of the vaccination rates. Also, if rural counties with low populations tend to have low vaccination rates, not accounting for county population can make vaccination seem less effective than it is.
Other questions worth asking the researchers are if there are any confounding variables that are likely to effect both vaccination rates and death rates, and if any of those confounding variables are measured. For example, the quality of the health care system in a county would effect both vaccination rates and deaths, so it would be worth asking if there are any measured variables that reflect the quality of the healthcare system in each county. Another example is the age demographics. The age demographics can influence both the death rate and the vaccination rate in a county, so it would be helpful to know if the researchers have any covariates that reflect age demographics (e.g. the percentage of the population above age 70 in each county). If the researchers mentions a large number of confounding that they have measurements for, I would ask the researchers to select the few most important ones based on their domain knowledge and would mention that they should choose much fewer than 20 variables to control for because there are only 20 samples.

In addition to asking about county population levels and whether there are measured confounder variables not presented in the table, it would be worthwhile to double check with the researcher that the deaths were counted after the vaccines were distributed (otherwise any analysis would be unable to say anything about the effect of vaccination rates on death rates).

b) Suppose the researchers are able to provide you with the population counts $N_1, \ldots, N_{20}$ of the 20 counties, and for each of the 20 they can give you a vectors $z_1, \ldots, z_{20}$ of the few most important measured confounder variables for each county (e.g. age demographics and health care system quality metrics). Also suppose that their death counts in each county, indeed only include deaths from a time period after most of the vaccinations were given.

Letting d_i denote the number of deaths in county i, v_i denote the vaccination rate, $x_i \equiv (v_i, z_i)$, I would fit the following Poisson GLM with offsets $\alpha_i \equiv \log(N_i)$

[4] D.K.

$$d_i \overset{\text{Ind}}{\sim} \text{Poisson}(\mu_i) \quad \log(\mu_i) = \alpha_i + x_i^T \beta \quad \text{for } i = 1, ..., 20.$$

After fitting the GLM (which can easily be done in R) using the glm function, I would look at the confidence interval for the first estimated coefficient $\hat{\beta}_1$. If the confidence interval only contains negative values, then we can conclude that the data suggests higher vaccination rates are associated with lower death rates when controlling for the confounders encoded in the z_i.

Problem 2: Finding an essential subset

Key Ideas/ Main Tools: Group Lasso

Defining the optimal essential subset as a solution to an optimization problem

Observe that one way to obtain an essential subset is to find the matrix $B \in \mathbb{R}^{750 \times 750}$ that minimizes $||R - RB||_F^2$ subject to the constraint that only 25 of the rows of B are allowed to have nonzero entries. More formally, letting $B[i,]$ denote the ith row of the matrix B we could get an essential subset by solving the following optimization problem on B:

$$\boxed{\text{minimize} \quad ||R - RB||_F^2 \quad \text{subject to} \quad \sum_{i=1}^{750} I\{B[i,] \neq \mathbf{0}\} \leq 25, B \in \mathbb{R}^{750 \times 750}}.$$

Let $\tilde{B}$ be the solution to the above optimization problem and $\mathcal{S} = \{i \in [750] \ : \ \tilde{B}[i,] \neq \mathbf{0}\}$, (that is let $\mathcal{S}$ be the indices of the nonzero rows of the solution $\tilde{B}$). $\mathcal{S}$ will give an essential subset, as each portfolio (represented by a column in R) will be reasonably well approximated by a linear combination of the portfolios of at most . [5]

Solving the boxed optimization problem and setting $\mathcal{S}$ to be the nonzero rows of the solution will recover an essential subset of size at most 25, but unfortunately the optimization problem is nonconvex (it has an l_0 type constraint).

A tractable approach using the Group Lasso

One way to induce a sparse number of nonzero rows of B is to use the Group Lasso. In particular, for each $\lambda > 0$, the Group Lasso can be used to solve the following convex optimization problem of finding:

[5] Technically, to be an essential subset, we only desire that portfolios not in $\mathcal{S}$ are well approximated by linear combinations of portfolios in $\mathcal{S}$, but trivially, portfolios in $\mathcal{S}$ can be written as exact linear combinations of portfolios in $\mathcal{S}$ and do not contribute to the loss function $||R - RB||_F^2$.

$$\boxed{\hat{B}_\lambda \in \underset{B\in\mathbb{R}^{750\times 750}}{\operatorname{argmin}} \Big(\frac{1}{2}||R-RB||_F^2 + \lambda \sum_{i=1}^{750} ||B[i,]||_2\Big).}$$

We can solve this group Lasso problem for many different λ values until we find a solution $\hat{B}_\lambda$ which has exactly 25 rows which have nonzero. In particular, letting $\hat{\mathcal{S}}_\lambda = \{i \in [750] \ : \ \hat{B}_\lambda[i,] \neq \mathbf{0}\}$, we can do a grid search on λ until we find a λ_* for which $|\hat{\mathcal{S}}_{\lambda_*}| = 25$. Then we can report $\hat{\mathcal{S}}_{\lambda_*}$ to our boss as an essential subset. Note that this may not be the optimal essential subset in the sense of minimizing $||R-RB||_F^2$ subject to 25 nonzero rows of B; however, it will still be an essential subset according to your bosses definition (that any portfolio in not in $\hat{\mathcal{S}}_{\lambda_*}$ can be well approximated by a linear combination portfolios in $\hat{\mathcal{S}}_{\lambda_*}$).

Additional References: In some years, the Group Lasso is covered in the 305 coursework's lecture notes[6], but it is not covered every year. See Obozinski et al. [2011] for a reference on the Group Lasso and some its theoretical guarantees in recovering a sparse set of rows.

Problem 3: Constructing Conformal Prediction Intervals

Key Ideas/ Main Tools: Prediction Intervals, Conformal Inference, Exchangeability

a) The defining property of the prediction interval is that $\mathbb{P}\big(Y_{n+1} \in [L,U]\big) \geq 1-\alpha$, where $\mathbb{P}$ is the joint distribution of the $n+1$ data points $(X_1,Y_1),\ldots,(X_n,Y_n),(X_{n+1},Y_{n+1})$

b) This procedure is not reasonable because the more you overfit the data, the smaller the predictions intervals will be. Ideally our prediction will not be over-confident about overfit predictions. In an extreme case suppose that $X_1,\ldots,X_n$ are all distinct and that $\hat{\mu}$ is the best fit $n-1$ degree polynomial to the first n datapoints. In this case, $\hat{\mu}(X_i) = y_i$ for all $i \in [n]$ implying that the residuals $r_1,\ldots,r_n$ are all equal to zero, further implying that the proposed prediction interval will have width zero. Clearly, if we overfit the data, the true prediction interval shouldn't have width zero for a new point X_{n+1}.

c) Let $\mathcal{S} = \{y \in \mathbb{R} \ : \ \pi(y) \leq (1-\alpha)(n+1)/n\}$ and let $L = \inf S$ and $U = \sup S$. To show that this gives a valid prediction interval first note that

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in [L,U]) &\geq \mathbb{P}(Y_{n+1} \in \mathcal{S}) \\
&= \mathbb{P}(\pi(Y_{n+1}) \leq (1-\alpha)(n+1)/n) \\
&= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{R_{Y_{n+1},i} \leq R_{Y_{n+1},n+1}\} \leq (1-\alpha)(n+1)/n\Big)
\end{aligned}$$

[6] https://web.stanford.edu/class/stats305c/notes/Regression/Sparse.html

To simplify the above expression with an exchangeability argument, first define $\tilde{\mu}$ to to the curve fit to the $n+1$ data points $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$. Next, define for $i = 1, \ldots, n+1$, $V_i \equiv |Y_i - \tilde{\mu}(X_i)|$. Observe that since $(X_1, Y_1), \ldots, (X_n, Y_n)$, (X_{n+1}, Y_{n+1}) are IID and $\tilde{\mu}$ is function of the collection of these $n+1$ data points, $V_1, V_2, \ldots, V_n, V_{n+1}$ is an exchangeable sequence of random variables. In addition, since $\tilde{\mu}(\cdot) = \hat{\mu}_{Y_{n+1}}(\cdot)$,

$$V_i \equiv |Y_i - \tilde{\mu}(X_i)| = |Y_i - \hat{\mu}_{Y_{i+1}}(X_i)| = R_{Y_{n+1},i}.$$

Combining this with a previous result and using the exchangeability of $(V_i)_{i=1}^{n+1}$ (and assuming that almost surely $V_i \neq V_j$ for $i \neq j$),

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in [L, U]) &\geq \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{R_{Y_{n+1},i} \leq R_{Y_{n+1},n+1}\} \leq (1-\alpha)(n+1)/n\Big) \\
&= \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq (1-\alpha)(n+1)/n\Big) \\
&= \mathbb{P}\Big(\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq (n+1)(1-\alpha)\Big) \\
&\geq \mathbb{P}\Big(\sum_{i=1}^{n} I\{V_i \leq V_{n+1}\} \leq \lfloor (n+1)(1-\alpha) \rfloor\Big) \\
&= \mathbb{P}\Big(\text{Unif}\{0, 1, \ldots, n-1, n\} \leq \lfloor (n+1)(1-\alpha) \rfloor\Big) \\
&= \frac{1 + \lfloor (n+1)(1-\alpha) \rfloor}{n+1} \\
&\geq 1 - \alpha.
\end{aligned}$$

Above the step where $\sum_{i=1}^{n} I\{V_i \leq V_{n+1} \sim \text{Unif}\{0, 1, \ldots, n-1, n\}$ follows from exchangeability of $(V_i)_{i=1}^{n+1}$ (and the assumption almost surely $V_i \neq V_j$ for $i \neq j$).

Note you can also cite Lemma's or Theorem's from Lecture 17 in Stats 300C to solve this problem.

d) If X_{n+1} is far outside the range of the training data, I would be concerned that the assumption that $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable from some distribution P is violated and that the intervals from part (c) are no longer valid. Even if X_{n+1} was technically a draw from P, the collaborator may have taken many draws from P and selected X_{n+1} as an outlier draw from P, in which case the exchangeability assumption would also be violated.

Despite failure to meet the exchangeability assumption, if we went ahead and constructed the prediction intervals defined in (c), we would get prediction intervals with undesirable behavior. In particular, if the curve $\hat{\mu}$ is fit based on

kernel smoothing or local linear regression (and only considers points with similar X values), then it would follow that for all y, $\hat{\mu}_y(X_{n+1}) = y$ implying that $R_{y,n+1} = 0$ for all y, further implying that $\pi(y) = 0$ for all y. Therefore, if $\hat{\mu}$ is fit based on kernel smoothing or local linear regression, the prediction interval would have infinite length. If on the other hand the curve $\hat{\mu}$ is fit based on a global polynomial regression, one would expect $R_{y,n+1}$ to be much larger than $R_{y,i}$ $(i < n)$ for most y values in which case the prediction interval would be very small. However, for a global model, since the global model is unlikely to hold for outliers we would want the prediction intervals to be very large. In summary, if we were to fit a curve with large extrapolation bias, the interval from part (c) would be very small and not reflect the extrapolation bias, but if we were to fit a local smooth model, the intervals from part (c) would be infinite length. The collaborator should therefore expect meaningless intervals if they went ahead and used prediction intervals for their outlier point.

Problem 4: PCA versus k-means

a) **Explanation for PCA:** Suppose that each of the features is centered such that columns of X each have mean 0. PCA can be thought of in terms of matrix factorizing X. In particular to implement PCA, you take the SVD of X, which is a matrix factorization given by $X = UDV^T$ where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $D_{ij} = 0$ for all $i \neq j$ and $|D_{11}| \geq |D_{22}| \geq \cdots \geq 0$. In PCA, the principal component directions are given by the columns of V, while the principal component scores are given by UD. Since $X = (UD)V^T$, principal component analysis can be thought of as factorizing X into matrix of the principal component scores (given by UD) and a matrix whose rows are the principal component directions (given by V^T).

Explanation for K-means: K-means can be thought of as an approximate matrix factorization of $X \in \mathbb{R}^{n\times p}$. In particular let

$$\mathcal{C}_K \equiv \Big\{ C \in \mathbb{R}^{n\times K} \ : \ C_{ij} \in \{0,1\} \ \forall_{i\in[n],j\in[K]}, \sum_{j=1}^{K} C_{ij} = 1 \ \forall_{i\in[n]} \Big\},$$

be the collection of $n \times K$ matrices whose rows contain exactly $K-1$ zeros and 1 one. We can then think of K-means as solving the following approximate matrix factorization problem

$$(\hat{C}, \hat{Z}) = \underset{C\in\mathcal{C}_K, Z\in\mathbb{R}^{K\times p}}{\operatorname{argmin}} ||X - CZ||_F^2.$$

In particular, if you solve the above approximate matrix factorization problem, the K rows of $\hat{Z}$ will give the K cluster centroids for K-means, and the column

index of the nonzero entry in each of the n rows of $\hat{C}$ will give the cluster assignment for each of the n points in the K-means algorithm (or put another way, $\hat{C}_{ij}$ is a indicator of whether the ith point is assigned to cluster j in the K-means algorithm).

b) For PCA, you can determine the matrix V, whose columns give the principal components dierections using X_{train}. Then, you can use the principal component scores on your test set given by $X_{\text{test}}V$ as features (or some subset of the principal component scores as features) for predicting Y_{test}. Note that if your prediction method is linear regression, and only the first k principal component scores are used as features, this approach would be principal components regression.

For K-means, you use the training data X_{train}, to determine the K cluster centroids. Once you have the cluster centroids, you can determine the cluster assignment of each observation in the test set by determining which of the K cluster centroids is closest (in Euclidean distance) to each row of the matrix X_{test}. Once you have the cluster assignments on the test set, you can use the cluster assignments as a categorical feature for prediction Y_{test} (you may want to use other features in addition to the cluster assignment).

You can evaluate the error in predicting Y for your choice of prediction algorithm and your choice features using cross-validation on the set $(X_{\text{test}}, Y_{\text{test}})$.

c) While my peer isn't wrong that the cluster structure they found is predictive of Y, it is just not such an impressive predictor of Y. In particular my peer's cluster assignments give a statistically significant improvement over just using the grand mean of Y for predicting Y. That being said, looking at the sum of squares in the ANOVA table, using the cluster means to predict Y doesn't give an impressive improvement over using the grand mean to predict Y. In a similar vein, from linear regression on the cluster we also see that the residuals from the linear regression tend to be much larger in absolute value than the estimates of Y for each cluster. So, I agree with my peer that the cluster structure found is predictive of Y, but it doesn't appear to be terribly important structure for predicting Y.

d) The results don't contradict each other so it does not cause me to doubt my peer's results. It is possible that the first first 5 principle components are simply not that associated (in a linear way) with the outcome variable Y. Perhaps the jth principal component for $j > 5$ is important in predicting Y. Perhaps the 1st principal component is important in explaining Y, but it has U shaped quadratic relationship with Y with a linear coefficient of 0. In either case the 3-means approach that my friend did would give a better prediction of Y than the principal components regression approach (with 5 components) that I took.

While it doesn't sound like my peer did anything wrong, I'm not terribly impressed by his results because the structure he found is quite a weak predictor of

Y (admittedly, I am even less impressed by my own results). I would suggest to our supervisor that we continue seeking an alternative to the 3-means approach, as there likely is a better choice of features out there. Perhaps it would be an 8-means approach or involve more than the first 5 principal components or it would involve interaction terms. Also neither of our approaches leveraged the Y_{train} data, even though we are told that it exists. I would therefore try to convince our supervisor that it is worth digging deeper before sticking to the 3-means approach of my peer.

Problem 5: Testing and inference on a censored Gaussian draw

Key Ideas/ Main Tools: Score test, maximum likelihood estimation, Bayesian inference

a) Z is distributed as $N(\mu, 1)$ variable constrained to be at least 2. Therefore, letting Φ denote the standard Gaussian CDF and ϕ denote the standard Gaussian pdf the likelihood is given by

$$L(\mu) = \frac{\frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(Z-\mu)^2\big)}{\int_2^\infty \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(z-\mu)^2\big) \mathrm{d}z} = \frac{\phi(Z-\mu)}{1-\Phi(2-\mu)} = \frac{\phi(Z-\mu)}{\Phi(\mu-2)}.$$

The loglikelihood is thus given by

$$l(\mu) = \log\big(L(\mu)\big) = \log\big(\frac{1}{\sqrt{2\pi}}\big) - \frac{1}{2}(Z-\mu)^2 - \log\big(\Phi(\mu-2)\big).$$

The score function is therefore given by

$$U(\mu) = l'(\mu) = (Z-\mu) - \frac{\phi(\mu-2)}{\Phi(\mu-2)},$$

and the Fisher information is given by

$$\begin{aligned} I(\mu) &= -\mathbb{E}[l''(\mu) \mid \mu] \\ &= -\mathbb{E}\Big[-1 - \frac{\Phi(\mu-2)\phi'(\mu-2) - [\phi(\mu-2)]^2}{[\Phi(\mu-2)]^2} \mid \mu\Big] \\ &= 1 + \frac{(2-\mu)\Phi(\mu-2)\phi(\mu-2) - [\phi(\mu-2)]^2}{[\Phi(\mu-2)]^2}. \end{aligned}$$

To conduct a score test of the hypothesis $H_0 : \mu = 0$, one would use the test statistic,

$$T = \frac{[U(0)]^2}{I(0)} = \frac{\big(Z - \frac{\phi(-2)}{\Phi(-2)}\big)^2}{1 + \frac{2\Phi(-2)\phi(-2) - [\phi(-2)]^2}{[\Phi(-2)]^2}} = \frac{(Z-r)^2}{1+2r-r^2} \quad \text{where } r = \frac{\phi(-2)}{\Phi(-2)} \approx 2.373.$$

The score test will reject whenever T exceeds $c_{1-\alpha}$, where $c_{1-\alpha}$ is the $1-\alpha$ quantile of a chi-squared distribution with one degree of freedom chosen to satisfy $\mathbb{P}(\chi_1^2 \leq c_{1-\alpha}) = 1-\alpha$. (Note $c_{1-\alpha} \approx 3.84$ for $\alpha = 0.05$). Thus the score test rejects whenever

$$T > c_{1-\alpha} \Leftrightarrow \frac{(Z-r)^2}{1+2r-r^2} > c_{1-\alpha} \Leftrightarrow Z > r+\sqrt{(1+2r-r^2)c_{1-\alpha}} \text{ or } Z < r-\sqrt{(1+2r-r^2)c_{1-\alpha}}.$$

Since $Z < 2$ cannot be observed, at level $\alpha = 0.05$, noting that $c_{0.95} \approx 3.84$, the score test of H_0 rejects whenever

$$Z > r + \sqrt{(1+2r-r^2)c_{1-\alpha}} \approx 3.036.$$

b) Since we just have to optimize over a 1-dimensional parameter you can use grid search to find the MLE. While the following argument is probably unneccasary for the applied qual, to double check that grid search can be used we will find $a, b \in \mathbb{R}$ such that we know $\text{argmax}_{\mu\in\mathbb{R}}\, l(\mu) \in [a,b]$. Note that whenever, $\mu > Z$, $l'(\mu) < 0$, so the MLE must be at most Z, and we can set $b = Z$. To find a lower endpoint a for the grid search observe that as a consequence of Theorem 1.2.3 in Durrett, for $x < -1$, $\frac{\phi(x)}{\Phi(x)} \leq (\frac{1}{-x} - \frac{1}{-x^3})^{-1}$

and hence for $\mu < 1$,

$$\begin{aligned}
l'(\mu) &= Z - \mu - \frac{\phi(\mu-2)}{\Phi(\mu-2)} \\
&\geq Z - \mu - \Big(\frac{1}{-(\mu-2)} - \frac{1}{-(\mu-2)^3}\Big)^{-1} \\
&= Z - \mu - \frac{(2-\mu)^3}{(2-\mu)^2 - 1} \\
&= Z - \mu - \frac{(2-\mu)^2}{(3-\mu)(1-\mu)}(2-\mu) \\
&= Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu(2-\mu)^2 - \mu(3-\mu)(1-\mu)}{(3-\mu)(1-\mu)} \\
&= Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu}{(3-\mu)(1-\mu)}.
\end{aligned}$$

Since the above inequality holds for any $\mu < 1$, it is easy to see that $\liminf_{\mu\downarrow-\infty} l'(\mu) \geq Z - 2 > 0$. Further we can use the lower bound above to find an a such that for all $\mu < a$,

$$l'(\mu) \geq Z - 2\frac{(2-\mu)^2}{(3-\mu)(1-\mu)} + \frac{\mu}{(3-\mu)(1-\mu)} > 0.$$

Hence we have an interval $[a,b]$ for which $l'(\mu) > 0$ when $\mu < a$ and $l'(\mu) < 0$ when $\mu > b$. It follows that the MLE $\text{argmax}_{\mu\in\mathbb{R}}\, l(\mu)$ must lie in $[a,b]$. Hence we can simply preform grid search over $[a,b]$ to find the MLE.

c) Suppose $\mu \sim \pi$ and an $Z \mid \mu \sim N(\mu, 1)$. We can estimate μ by considering the posterior distribution of μ given Z. In particular, if we observe some $Z > 2$, by Bayes' rule

$$p(\mu \mid Z, Z > 2) = p(\mu \mid Z) \propto \pi(\mu)\frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}(Z-\mu)^2\big) \propto \pi(\mu) \exp\big(-\frac{1}{2}(Z-\mu)^2\big).$$

Since the posterior distribution of μ is proportional to $\pi(\mu) \exp\big(-\frac{1}{2}(Z-\mu)^2\big)$, we can estimate μ with the MAP estimate given by

$$\hat{\mu}_{\text{MAP}} = \underset{\mu \in \mathbb{R}}{\text{argmax}}\Big\{\pi(\mu) \exp\big(-\frac{1}{2}(Z-\mu)^2\big)\Big\}.$$

As in part (b) this estimator can be found using 1-dimensional grid search. An alternative estimate for μ would be to estimate the posterior mean

$$\hat{\mu} = \mathbb{E}[\mu \mid Z] = \frac{\int_{-\infty}^{\infty} \mu\pi(\mu) \exp\big(-\frac{1}{2}(Z-\mu)^2\big)\mathrm{d}\mu}{\int_{-\infty}^{\infty} \pi(\mu) \exp\big(-\frac{1}{2}(Z-\mu)^2\big)\mathrm{d}\mu}.$$

The numerator and denominator of the above expression can each be approximated numerically using a Gauss-Hermite quadrature. Alternatively, the above posterior mean can be estimated using the Metropolis-Hastings algorithm.

d) I'm not entirely sure what the question means by "conflict", but the answer in part (c) was a Bayesian approach based on one observation for which $Z > 2$, whereas part (a) and part (b) describe frequentist approaches where Z is drawn until $Z > 2$. The answer in part (c) used a different likelihood than was used in parts (a) and (b). In particular the likelihood in items (a) and (b) was $p(Z \mid \mu, Z > 2) = \phi(Z-\mu)/\Phi(\mu-2)$ whereas the likelihood in part (c) that was used was $p(Z \mid \mu) = \phi(Z-\mu)$. The easiest way to see why it was not a mistake to use a different likelihood in part (c), is to note that we could have used the same likelihood in part (c) when applying Bayes' rule as the likelihood used in (a) and (b), but doing so would have made a more difficult calculation. In particular, conditioning on $Z > 2$, we could have used Bayes' rule as follows

$$p(\mu \mid Z, Z > 2) = \frac{p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2)}{\int_{-\infty}^{\infty} p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2)\mathrm{d}\mu} \propto p(\mu \mid Z > 2)p(Z \mid \mu, Z > 2),$$

and used the likelihood from parts (a) and (b), but $\pi(\mu \mid Z > 2)$ is more difficult to work with than $\pi(\mu)$ and requires applying Bayes' rule. In fact, if we apply Bayes' rule to $\pi(\mu \mid Z > 2)$ in the above expression, we simply recover the approach used in part (c):

$$p(\mu \mid Z, Z > 2) \propto \pi(\mu)p(Z > 2 \mid \mu)p(Z \mid \mu, Z > 2) = \pi(\mu)\Phi(\mu-2)\frac{\phi(Z-\mu)}{\Phi(\mu-2)} = \pi(\mu)\phi(Z-\mu).$$

Problem 6: Cross-validation in the normal linear model

Key Ideas/ Main Tools: Cross validation, properties of the normal linear model.

This problem is based on results from Bates et al. [2022]. Note that the 2021 qual was open internet, so I think that those who were aware of the paper or were able to find the paper found it the problem quite straightforward, but those who didn't found part (b) especially tricky.

a) Fix any $x_1, x_2, \ldots, x_n, y_1, \ldots, y_n, \kappa$ and u. Now for any subset $S \subset [n]$, let $\hat{\theta}_{S,0}$ denote the OLS estimator trained on the points $\{(x_i, y_i)\}_{i \in S}$ and let $\hat{\theta}_{S,\kappa}$ denote the OLS estimator trained on the shifted points $\{(x_i, y_i + x_i^T \kappa)\}_{i \in S}$. Also let $\mathcal{X}_S \in \mathbb{R}^{|S| \times p}$ be the design matrix for these OLS regressions whose rows consist of $\{x_i \ : \ i \in S\}$ and letting $\mathcal{Y}_S = (y_i)_{i \in S}$ be the vector of outcomes for the OLS regression to obtain $\hat{\theta}_{S,0}$. Observe that by the formula for an OLS estimator

$$\hat{\theta}_{S,\kappa} = (\mathcal{X}_S^T \mathcal{X}_S)^{-1} \mathcal{X}_S^T \big[\mathcal{Y}_S + \mathcal{X}_S \kappa\big] = (\mathcal{X}_S^T \mathcal{X}_S)^{-1} \mathcal{X}_S^T \mathcal{Y}_S + \kappa = \hat{\theta}_{S,0} + \kappa.$$

For any $j \notin S$, if we let $\hat{y}_{S,0,j} = x_j^T \hat{\theta}_{S,0}$ be the OLS prediction at point j when training on the subset S for the raw data (x_i, y_i) and if we let $\hat{y}_{S,\kappa,j} = x_j^T \hat{\theta}_{S,\kappa}$ be the OLS prediction at point j when training on the subset S of the translated data $(x_i, y_i + x_i^T \kappa)$, it follows that

$$\ell(\hat{y}_{S,\kappa,j}, y_j + x_j^T \kappa) = \ell(x_j^T \hat{\theta}_{S,\kappa}, y_j + x_j^T \kappa) = \ell(x_j^T \hat{\theta}_{S,0} + x_j^T \kappa, y_j + x_j^T \kappa) = \ell(x_j^T \hat{\theta}_{S,0}, y_j) = \ell(\hat{y}_{S,0,j}, y_j),$$

where the 2nd last steps holds because ℓ is the squared error loss function. It is clear that above argument holds for any subset S and $j \notin S$.

Letting $S_1(u), \ldots, S_K(u)$ be the K training subsets of $[n]$ that define the cross-validation (which depend on the random draw U which we are fixing to be u), note that applying the previous result,

$$\begin{aligned}
\widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((x_1, y_1), \ldots, (x_n, y_n), u\Big) &\equiv \frac{1}{n} \sum_{k=1}^{K} \sum_{j \notin S_k(u)} \ell(\hat{y}_{S_k(u),0,j}, y_j) \\
&= \frac{1}{n} \sum_{k=1}^{K} \sum_{j \notin S_k(u)} \ell(\hat{y}_{S_k(u),\kappa,j}, y_j + x_j^T \kappa) \\
&= \widehat{\mathrm{Err}}^{(\mathrm{CV})}\Big((x_1, y_1 + x_1^T \kappa), \ldots, (x_n, y_n + x_n^T \kappa), u\Big).
\end{aligned}$$

Since this argument holds for any fixed $x_1, x_2, \ldots, x_n, y_1, \ldots, y_n, \kappa$ and u, it follows that $\widehat{\mathrm{Err}}^{(\mathrm{CV})}$ is linearly invariant by definition (2).

b) Recalling from the problem statement that $\hat{\theta}$ is the OLS estimator based on the all of the observed data, and let $(R_1, \ldots, R_n)$ be the residuals for the OLS estimate on the observed data (i.e. $R_i = Y_i - X_i^T\hat{\theta}$ for all $i \in [n]$). By part (a), if we let $\kappa = -\hat{\theta}$,

$$\begin{aligned}\widehat{\text{Err}}^{(\text{CV})}\Big((X_1, Y_1), \ldots, (X_n, Y_n), U\Big) &= \widehat{\text{Err}}^{(\text{CV})}\Big((X_1, Y_1 - X_1^T\hat{\theta}), \ldots, (X_n, Y_n - X_n^T\hat{\theta}), U\Big) \\ &= \widehat{\text{Err}}^{(\text{CV})}\Big((X_1, R_1), \ldots, (X_n, R_n), U\Big).\end{aligned}$$

It follows conditional on $X = (X_1, \ldots, X_n)$, $\widehat{\text{Err}}^{(\text{CV})}$ is a function of only the residuals $(R_1, R_2, \ldots, R_n, U)$. Also observe that conditional on X, Err_{XY} is only a function of $\hat{\theta}$. By a property of linear regression under the homoskedastic linear model, $(R_1, R_2, \ldots, R_n) \perp\!\!\!\perp \hat{\theta}|X$ (to see this, one can check using the hat matrix that conditional on X, the residuals are uncorrelated with the estimator $\hat{\theta}$ and note that for multivariate Gaussian's zero correlation implies independence). Since U is independent of the data, this further implies that

$$(R_1, R_2, \ldots, R_n, U) \perp\!\!\!\perp \hat{\theta}|X.$$

Because conditional on X, we have shown that $\widehat{\text{Err}}^{(\text{CV})}$ is only a function of $(R_1, R_2, \ldots, R_n, U)$ and because conditional on X, Err_{XY} is only a function of $\hat{\theta}$ the conditional independence result displayed above implies that

$$\widehat{\text{Err}}^{(\text{CV})} \perp\!\!\!\perp \text{Err}_{XY}|X.$$

c) The previous result from item (b) does not imply that $\widehat{\text{Err}}^{(\text{CV})}$ is a useless estimate of prediction error. In particular, $\widehat{\text{Err}}^{(\text{CV})}$ is still a good estimate of $\text{Err} \equiv \mathbb{E}[\text{Err}_{XY}]$, which is the expected prediction loss across all training sets (see Chapter 7.12 in Hastie et al. [2009] and Bates et al. [2022]). Even if we are truly interested in estimating Err_{XY} rather than Err, just because $\widehat{\text{Err}}^{(\text{CV})}$ is uncorrelated with Err_{XY}, it does not mean that it is a bad approximation of Err_{XY}: for large n, the random variable Err_{XY} will likely be concentrated closely about its mean $\text{Err} = \mathbb{E}[\text{Err}_{XY}]$.

d) Sample splitting into a training set and a test set would also give a linearly invariant estimate of the prediction error by a similar argument in part (a) and the same argument in part (b). Another commonly used estimate of prediction error that is linearly invariant, so that (3) holds is Mallow's C_p. See Bates et al. [2022] for discussion about Mallow's C_p and other commonly used linearly invariant estimates of prediction error.

References

Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2022.

A. C. Davison and D. V. Hinkley. *Further Ideas*, page 70–135. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.004.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1 – 47, 2011. doi: 10.1214/09-AOS776. URL `https://doi.org/10.1214/09-AOS776`.