Applied Statistics Qualifying Exams Coaching

Michael Howes*

Summer 2023

Contents

[*] With lots of content credit given to previous applied quals coaches

0 Applied 2022[1]

Problem 1: R-squared and PCA

Key ideas:

- Connections between principal components analysis and the singular value decomposition.
- Orthogonality of principal components and principal component directions.

(a) We are given that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is standardized so that the columns have mean zero and variance one. We are also given that $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{X}$. Let $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ be the diagonal entries of $\boldsymbol{D}$. The columns of $\boldsymbol{V}$ are thus the principal component directions and $\frac{1}{n}d_j^2 = \frac{1}{n}\|Xv_j\|_2^2$ is the variance of X in the jth principal component direction. The cumulative percent variance explained sequence is thus,

$$\rho_k = 100 \times \frac{\sum_{s=1}^k \frac{1}{n}d_s^2}{\sum_{s=1}^p \frac{1}{n}d_s^2} = 100 \times \frac{\sum_{s=1}^k d_s^2}{\sum_{s=1}^p d_s^2},$$

for $k = 1, \ldots, p$.

(b) We are given a response $\boldsymbol{y} \in \mathbb{R}^n$ with mean zero and variance one. The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} \in \mathbb{R}^n$. We know that the fitted values $\hat{\boldsymbol{y}}$ are orthogonal to the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$. Thus,

$$\|\boldsymbol{y}\|_2^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}\|_2^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{y}}\|_2^2.$$

And so

$$\frac{1}{n}\sum_{i=1}^n y_i^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^n \hat{y}_i^2.$$

Since $\boldsymbol{y}$ has mean zero and variance one, this implies that

$$1 = \mathrm{MSE}_0 = \mathrm{MSE} + \mathrm{MSS}.$$

And so,

$$R^2 = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_0} = 1 - \mathrm{MSE} = \mathrm{MSS}.$$

(c) We now regress the jth column of $\boldsymbol{X}$ on the first k principal components of $\boldsymbol{X}$. Let $\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k] \in \mathbb{R}^{n \times k}$ be the matrix containing the first k columns of $\boldsymbol{U}$. Since $\boldsymbol{U}_k^\top \boldsymbol{U}_k = \boldsymbol{I}_k$, the fitted values from regressing $\boldsymbol{x}_j$ on $\boldsymbol{U}_k$ are

$$\hat{\boldsymbol{x}}_j = \boldsymbol{U}_k(\boldsymbol{U}_k^\top \boldsymbol{U}_k)^{-1}\boldsymbol{U}_k^\top \boldsymbol{x}_j = \boldsymbol{U}_k\boldsymbol{U}_k^\top \boldsymbol{x}_j.$$

[1] Michael Howes

The average regression sum of squares is thus,

$$\begin{aligned}\mathrm{MSS}_j &= \frac{1}{n}\hat{\boldsymbol{x}}_j^\top \hat{\boldsymbol{x}}_j \\ &= \frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{x}_j \\ &= \frac{1}{n}\boldsymbol{x}_j^\top \boldsymbol{U}_k \boldsymbol{U}_k^\top \boldsymbol{x}_j.\end{aligned}$$

Since $\boldsymbol{x}_j$ is the jth column of $\boldsymbol{X}$ we $\boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{e}_j$ where $\boldsymbol{e}_j \in \mathbb{R}^p$ is the jth standard basis vector.

$$\begin{aligned}\boldsymbol{U}_k^\top \boldsymbol{x}_j &= \boldsymbol{U}_k \boldsymbol{X}\boldsymbol{e}_j \\ &= \boldsymbol{U}_k^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{e}_j.\end{aligned}$$

We know that $\boldsymbol{U}_k^\top \boldsymbol{U} = [\boldsymbol{I}_k, \boldsymbol{0}_{k\times(p-k)}] \in \mathbb{R}^{k\times p}$ where $\boldsymbol{0}_{k\times(p-k)}$ is a matrix of all zeros of size $k \times (p-k)$. It follows that

$$\begin{aligned}\boldsymbol{U}_k^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top &= [\boldsymbol{I}_k, \boldsymbol{0}_{k\times(p-k)}]\boldsymbol{D}\boldsymbol{V}^\top \\ &= \boldsymbol{D}_k \boldsymbol{V}_k^\top,\end{aligned}$$

where $\boldsymbol{D}_k = \mathrm{diag}(d_1, \dots, d_k) \in \mathbb{R}^{k\times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{p\times k}$ is equal to the first k rows of $\boldsymbol{V}$. Thus,

$$\begin{aligned}\boldsymbol{U}_k^\top \hat{\boldsymbol{x}}_j &= \boldsymbol{D}_k \boldsymbol{V}_k^\top \boldsymbol{e}_j \\ &= \sum_{s=1}^k d_s \left(\boldsymbol{v}_s^\top \boldsymbol{e}_j\right) \boldsymbol{e}_s \\ &= \sum_{s=1}^k d_s v_{sj} \boldsymbol{e}_s\end{aligned}$$

where v_{sj} is the entry of $\boldsymbol{V}$ in row s and column j. We thus have

$$\begin{aligned}\mathrm{MSS}_j &= \frac{1}{n}\|\boldsymbol{U}_k^\top \hat{\boldsymbol{x}}_j\|_2^2 \\ &= \frac{1}{n}\left\|\sum_{s=1}^k d_s v_{sj}\boldsymbol{e}_s\right\|_2^2 \\ &= \frac{1}{n}\sum_{s=1}^k d_s^2 v_{sj}^2.\end{aligned}$$

Since $\boldsymbol{x}_j$ has mean zero and variance one, we are in the setting of part (b) and hence

$$R_j^2 = \mathrm{MSS}_j = \frac{1}{n}\sum_{s=1}^k d_s^2 v_{sj}^2.$$

(d) Note that

$$\begin{aligned}\sum_{j=1}^{p} \mathrm{MSS}_j &= \sum_{j=1}^{p} \frac{1}{n} \sum_{s=1}^{k} d_s^2 v_{sj}^2 \\ &= \sum_{s=1}^{k} \sum_{j=1}^{p} \frac{1}{n} d_s^2 v_{sj}^2 \\ &= \sum_{s=1}^{k} \frac{1}{n} d_s^2 \sum_{j=1}^{p} v_{sj}^2 \\ &= \sum_{s=1}^{k} \frac{1}{n} d_s^2 \|\boldsymbol{v}_s\|_2^2 \\ &= \frac{1}{n} \sum_{s=1}^{k} d_s^2,\end{aligned}$$

since all rows of $\boldsymbol{V}$ have norm one. We know that each column of $\boldsymbol{X}$ has variance one and mean zero. Thus,

$$\begin{aligned}p &= \sum_{j=1}^{p} \frac{1}{n} \boldsymbol{x}_j^\top \boldsymbol{x}_j \\ &= \frac{1}{n} \sum_{j=1}^{p} \mathrm{tr}\left(\boldsymbol{x}_j^\top \boldsymbol{x}_j\right) \\ &= \frac{1}{n} \sum_{j=1}^{p} \mathrm{tr}\left(\boldsymbol{x}_j \boldsymbol{x}_j^\top\right) \\ &= \frac{1}{n} \mathrm{tr}\left(\sum_{j=1}^{p} \boldsymbol{x}_j \boldsymbol{x}_j^\top\right) \\ &= \frac{1}{n} \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \\ &= \frac{1}{n} \mathrm{tr}\left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\right) \\ &= \frac{1}{n} \mathrm{tr}\left(\boldsymbol{D}^2\right) \\ &= \frac{1}{n} \sum_{s=1}^{p} d_s^2.\end{aligned}$$

Thus,

$$\frac{100}{p} \sum_{j=1}^{p} \mathrm{MSS}_j = \frac{100}{np} \sum_{s=1}^{k} d_s^2 = 100 \times \frac{\sum_{s=1}^{k} d_s^2}{\sum_{s=1}^{p} d_s^2} = \rho_k.$$

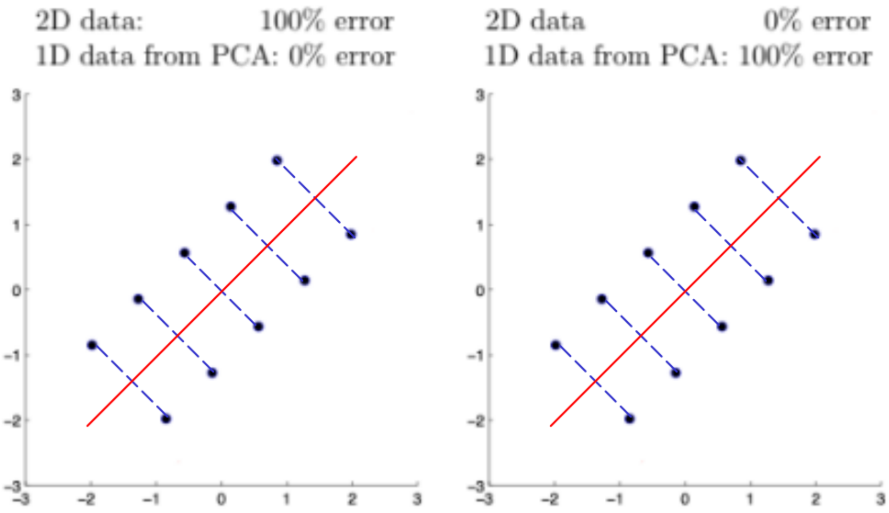
So $\frac{100}{p}\sum_{j=1}^{p}\text{MSS}_j$ is exactly the cumulative percent variance explained sequence.

Problem 2: LOO, PCA and 1NN

Key ideas:

- "Eyeballing" principal component directions.
- Working out nearest neighbor classifiers.

1. We are asked to draw a line corresponding to the first principal component. This does not have to be exact, but we can see that a roughly 45 degree line gives the direction with the most variance. Here the principal component direction is represented as a red solid line. In both plots I have also included dashed lines showing the projection onto the first principal direction.



2D data: 100% error
1D data from PCA: 0% error

2D data 0% error
1D data from PCA: 100% error

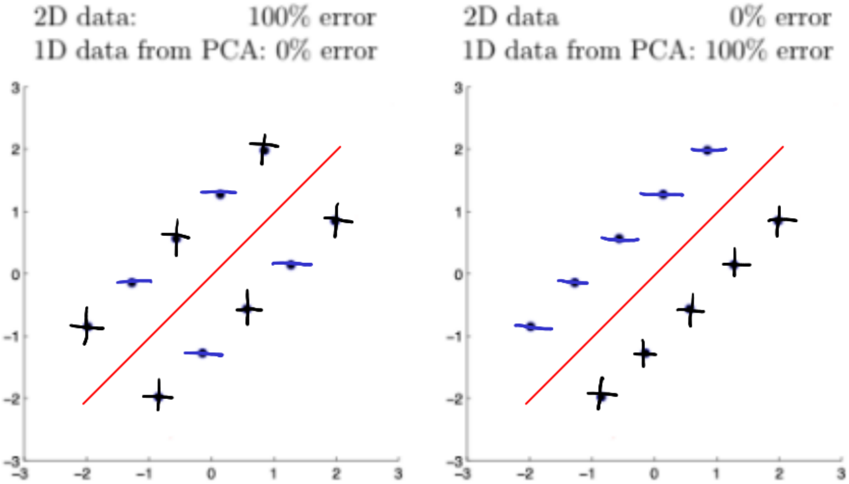
2. Looking at the previous figure we can make two observations.

   - Using the 2D data, the nearest neighbor of a point is one of the adjacent points on the line of points parallel to the PCA direction. For example, for the point at roughly $(-, 1, -2)$, the nearest neighbor is the point at roughly $(0, -1.25)$.
   - Using the projected 1D data from PCA, the nearest neighbor of a point is the point across the PCA direction. This because these points get projected onto the same value when we perform PCA. For example, for the point at roughly $(-1, -2)$, the 1D projected nearest neighbor is the point at roughly $(-2, -1)$.

To have 100% error on the 2D data we want an alternating sequence of "+"'s and "−"'s along the two lines of points parallel to the PCA direction. If we use the same sequence of "+"'s and "−"'s on both lines of points, then we will also have 0% error when using the 1D data. This is because the 1D-nearest neighbor points will have the same labels.

To have 0% error on the 2D data we want adjacent points on the two lines to all have the same label. If we label only line of points with "+"'s and the other with "−"'s, then the 1D data will also have 100% error. This is because the 1D nearest neighbors will have the opposite labels.

In summary, the labelling below has the specified error rates.



2D data: 100% error
1D data from PCA: 0% error

2D data 0% error
1D data from PCA: 100% error

Question 3: Financial Data Science

Question 4: Infectious Disease Survival Times

Key ideas:

- Testing coefficients in the proportional hazard model.
- Testing independence in $2 \times 2 \times K$ tables with the Mantel–Haenszel test.

(a) Since our data is right censored, we will use tools from survival analysis. Since we want to control for affects from each region, we can use a proportional hazard model. Suppose we have N subjects and for each subject $i = 1, \ldots, N$ we observe

- An observed time $O_i = \min\{C_i, T_i\}$. The time T_i is the event time (in this case the time at which the subject died after infect). The time C_i is the censoring time which we assume is independent of T_i.
- A censor-indicator $\delta_i = I_{\{O_i = T_i\}} \in \{0, 1\}$ with $\delta_i = 0$ meaning subject i is censored.
- The vaccination status $V_i \in \{0, 1\}$, we'll assume that $V_i = 1$ means the subject is vaccinated.
- The region $R_i \in \{1, \ldots, J\}$ where J is the number of regions (assumed to be small).

A proportional hazard models that include effects from both vaccination status and subject regions is

$$T_i \mid V_i, R_i \overset{\text{ind}}{\sim} h_i(t), \quad h_i(t) = h_0(t) \exp(\alpha V_i + \beta_{R_i}), \tag{1}$$

where we are modelling the distribution of T_i in terms of T_i's hazard function $h_i(t)$. To make this model identifiable, we must add a constraint such as

$$\sum_{j=1}^{J} \beta_J = 0 \quad \text{or} \quad \beta_J = 0.$$

If the number of regions J is not too large, then we will be able to fit the above model by maximizing the partial likelihood. If the number of regions is large, but we have measured features X_i for subjects i's region, then we could use the model

$$T_i \mid V_i, R_i \overset{\text{ind}}{\sim} h_i(t), \quad h_i(t) = h_0(t) \exp(\alpha V_i + \gamma^\top X_i).$$

Assuming the dimension of X_i is less than $J - 1$, then fitting this second model would be easier than fitting (1). We are told that there are only a "handful" of possible value of R, thus we will assume we have enough data to fit (1). To test the effectiveness of the vaccine, we can test the null hypothesis $\alpha = 0$. If we get evidence that $\alpha < 0$, then we can conclude that vaccination likely has a positive effect after controlling for regions.

An alternative model would be to include an interaction affect between region and vaccination status. This model would be appropriate if we expect the effectiveness of the vaccine to vary across regions. The corresponding proportional hazard model is,

$$T_i \mid V_i, R_i \overset{\text{ind}}{\sim} h_i(t), \quad h_i(t) = h_0(t) \exp(\alpha_{R_i} V_i + \beta_{R_i}).$$

Again, testing $\alpha_j = 0$ across the regions j will test for the vaccine's effectiveness in region j. This probably is not a good choice of model. Firstly, we might not have enough data to fit both α_j and β_j. Secondly, while it is reasonable to expect health outcomes to vary across regions (represented by β_j) it is unlikely that the effectiveness of the vaccine will vary across regions (represented by α_j). We will thus stick with model (1)

(b) In this question, we will ignore the variation across regions. The corresponding Cox proportional hazards model is

$$T_i \mid V_i, R_i \overset{\text{ind}}{\sim} h_i(t), \quad h_i(t) = h_0(t)\exp(\alpha V_i)$$

This model is fit by maximizing the log partial likelihood,

$$\ell(\alpha) = \sum_{i:\delta_i=1} \alpha V_i - \log\left(\sum_{k\in R(O_i)} \exp(\alpha V_k)\right), \tag{2}$$

where $R(O_i) = \{k : O_k \geq O_i\}$ is the risk-set at time O_i. For the null $H_0 : \alpha = 0$, the score-test statistic is

$$T = \frac{\ell'(0)^2}{-\ell''(0)},$$

which has asymptotic distribution χ^2_1. We know that the score-test in a Cox proportional hazards model is equivalent to the log-rank test. The log-rank test is itself a special case of the Mantel–Haenszel test. For each i such that $\delta_i = 1$, we can make the following contigency table based on the risk set at time O_i. Specifically, we consider all individuals with $O_k \geq O_i$ and put them into one of the four below categories The null hypothesis that vaccination does not affect survival time implies that the row variable and column variable in the above 2×2 contingency table are equivalent. If there was just one such table, then we could condition on the row and column sums and use Fisher exact test. However, we actually have $K = |\{i : \delta_i = 1\}|$ such tables, and we need a way of combining them. The Mantel–Haenszel test uses the hypergeometric distribution from Fisher's exact test to combine these tables into a test statistic that is asymptotically distributed according to χ^2_1. It turns out that this statistic is exactly the score statistic T. This can be proved using the formula for the partial likelihood (2) and the formula for the Mantel–Haenszel test.

	$V_k = 0$	$V_k = 1$
$O_k = O_i$ and $\delta_k = 1$	Number of unvaccinated subjects that died at time O_i	Number of vaccinated subjects that died at time O_i
$O_k > O_i$ or $\delta_k = 0$	Number of unvaccinated subjects that survived beyond time O_i	Number of vaccinated subjects that survived beyond time O_i

(c) Suppose we are now using model (1) which allows variation across regions. If we wish to do inference on the affect of vaccination, we can still use the score test. The two hypothesis we are comparing are now,

$$H_0 : \alpha = 0 \quad \text{vs} \quad H_1 : \alpha \text{ unconstrained.}$$

Thus, we wish to test if the sub-model which does not include vaccination status is sufficient to explain our data. To make both the null and alternative models identifiable we will add the constraint $\beta_J = 0$ and think of β as a vector in $\mathbb{R}^{J-1}$. The log partial likelihood is equal to

$$\ell(\alpha, \beta) = \sum_{i:\delta_i = 0} \alpha V_i + \beta_{R_i} - \log\left(\sum_{k \in R(O_i)} \exp\left(\alpha V_k + \beta_{R_k}\right)\right).$$

Let

$$U(\alpha, \beta) = \nabla_{\alpha,\beta}\ell(\alpha, \beta) \in \mathbb{R}^J,$$
$$I(\alpha, \beta) = -\nabla^2_{\alpha,\beta}\ell(\alpha, \beta) \in \mathbb{R}^{J \times J},$$

be the gradient and negative Hessian of ℓ evaluated at (α, β). Let $\hat{\beta}_0 \in \mathbb{R}^{J-1}$ be the MLE of the Cox proportional hazards model under the constraints $\alpha = 0$ and $\beta_J = 0$. The score statistic for testing H_0 is

$$T = U(0, \hat{\beta}_0)^\top I(0, \hat{\beta}_0)^{-1} U(0, \hat{\beta}_0) \mathrel{\dot\sim} \chi^2_1, \tag{3}$$

where the approximation above is asymptotic under the null. As in part (b), we could again use contingency tables to test H_0. Again we have a 2×2 table for uncensored observation i. However, now our table should only include people in the same region of subject i. Specifically, if $\delta_i = 1$ for some i, then we make the following table This is the same as the contingency table in part (b) but now we only included individuals in the same region as subject i. Under the null hypothesis $\alpha = 0$, the row and column variables are independent. Thus, conditional on the row and column, the value in the top square is hypergeometric as in Fisher's exact test. We can thus combine these table via the Mantel–Haenszel test and get a test statistic which will be asymptotically χ^2_1 under the null. However, I do not believe that this is the same test statistic as (3).

	$V_k = 0$ and $R_k = R_i$	$V_k = 1$ and $R_k = R_i$
$O_k = O_i$ and $\delta_k = 1$	Number of unvaccinated subjects that died at time O_i	Number of vaccinated subjects that died at time O_i
$O_k > O_i$ or $\delta_k = 0$	Number of unvaccinated subjects that survived beyond time O_i	Number of vaccinated subjects that survived beyond time O_i

When adding the region covariates, we have two ways of testing $\alpha = 0$. One is based on the score test and the other is a variant of the Mantel–Haenszel test. While these tests are likely to be related, I believe that they are different.

Question 5: EM for a mixture of Student-t distributions

(a) Consider the following augmented model. Independently for $1 \leq n \leq N$,

$$\begin{aligned} z_n &\sim \boldsymbol{\pi}, \\ \tau_n \mid z_n &\sim \mathrm{Ga}\left(\nu_{z_n}/2, \nu_{z_n}/2\right), \\ \boldsymbol{x}_n \mid \tau_n, z_n &\sim \mathcal{N}\left(\boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right), \end{aligned}$$

the joint density for $\{z_n, \tau_n, \boldsymbol{x}_n\}_{n=1}^N$ corresponding to this model is

$$\begin{aligned} & p\left(\{z_n, \tau_n, \boldsymbol{x}_n\}_{n=1}^N; \boldsymbol{\theta}\right) \\ &= \prod_{n=1}^N p(z_n, \tau_n, \boldsymbol{x}_n; \boldsymbol{\theta}) \\ &= \prod_{n=1}^N p(z_n; \boldsymbol{\theta}) p(\tau_n \mid z_n; \boldsymbol{\theta}) p(\boldsymbol{x}_n \mid \tau_n, z_n; \boldsymbol{\theta}) \\ &= \prod_{n=1}^N \pi_{z_n} \mathrm{Ga}\left(\tau_n; \nu_{z_n}/2, \nu_{z_n}/2\right) \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right). \end{aligned}$$

The marginal model for $\{z_n, \boldsymbol{z}_n\}_{n=1}^N$ is

$$\begin{aligned} & p\left(\{z_n, \boldsymbol{x}_n\}; \boldsymbol{\theta}\right) \\ &= \int_{\mathbb{R}^N_{>0}} p\left(\{z_n, \tau_n, \boldsymbol{x}_n\}; \boldsymbol{\theta}\right) d\boldsymbol{\tau} \\ &= \int_{\mathbb{R}^N_{>0}} \prod_{n=1}^N \pi_{z_n} \mathrm{Ga}\left(\tau_n; \nu_{z_n}/2, \nu_{z_n}/2\right) \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right) d\boldsymbol{\tau} \\ &= \prod_{n=1}^N \int \pi_{z_n} \mathrm{Ga}\left(\tau_n; \nu_{z_n}/2, \nu_{z_n}/2\right) \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right) d\tau_n \\ &= \prod_{n=1}^N \pi_{z_n} \int \mathrm{Ga}\left(\tau_n; \nu_{z_n}/2, \nu_{z_n}/2\right) \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right) d\tau_n \\ &= \prod_{n=1}^N \pi_{z_n} \mathrm{St}\left(\boldsymbol{x}_n; \nu_{z_n}, \boldsymbol{\mu}_{z_n}, \tau_n^{-1}\boldsymbol{\Sigma}_{z_n}\right), \end{aligned}$$

where the last line follows from the equation describing the Student-t distribution as a scale mixture of Gaussian. This joint probability corresponds to the original student mixture model.

(b) We are asked to compute,

$$\omega_{nk} = p(z_n = k \mid \boldsymbol{x}_n; \boldsymbol{\theta}),$$

for $1 \leq k \leq K$ and $1 \leq n \leq N$. By Bayes rule,

$$\begin{aligned}\omega_{nk} &= \frac{p(\boldsymbol{x}_n \mid z_n = k; \boldsymbol{\theta}) p(z_n = k; \boldsymbol{\theta})}{\sum_{j=1}^K p(\boldsymbol{x}_n \mid z_n = j; \boldsymbol{\theta}) p(z_n = j; \boldsymbol{\theta})} \\ &= \frac{\mathrm{St}(\boldsymbol{x}_n; \nu_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^K \mathrm{St}(\boldsymbol{x}_n; \nu_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j},\end{aligned}$$

since $\boldsymbol{x}_n \mid z_n \sim \mathrm{St}\left(\nu_{z_n}, \boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}\right)$.

(c) Next we are asked to compute the conditional distribution,

$$p(\tau_n \mid z_n = k, \boldsymbol{x}_n; \boldsymbol{\theta}).$$

To do this, we will again use Bayes rule conditional on $z_n = k$,

$$\begin{aligned}&p(\tau_n \mid z_n = k, \boldsymbol{x}_n; \boldsymbol{\theta}) \\ &\propto p(\tau_n \mid z_n = k; \boldsymbol{\theta}) p(\boldsymbol{x}_n \mid \tau_n, z_n = k; \boldsymbol{\theta}) \\ &= \mathrm{Ga}(\tau_n; \nu_k/2, \nu_k/2) \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \tau_n^{-1}\boldsymbol{\Sigma}_k) \\ &\propto \tau_n^{\nu_k/2-1} \exp\left(-\nu_k \tau/2\right) \det(\tau_n^{-1}\boldsymbol{\Sigma}_k)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \left(\tau_n^{-1}\boldsymbol{\Sigma}_k\right)^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right).\end{aligned}$$

Note that,

$$\det(\tau_n^{-1}\boldsymbol{\Sigma}_k) = \det(\tau_n^{-1}\boldsymbol{I}_D \boldsymbol{\Sigma}_k^{-1}) = \tau_n^{-D} \det(\boldsymbol{\Sigma}_k^{-1}),$$

and

$$-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \left(\tau_n^{-1}\boldsymbol{\Sigma}_k\right)^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) = -\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\tau_n.$$

Thus,

$$\begin{aligned}&p(\tau_n \mid z_n = k\boldsymbol{x}_n; \boldsymbol{\theta}) \\ &\propto \tau_n^{\nu_k/2-1} \exp\left(-\nu_k \tau/2\right) \tau^{D/2} \det(\boldsymbol{\Sigma}_k)^{-1/2} \exp\left(\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\tau_n\right) \\ &= \tau_n^{(\nu_k+D)/2-1} \exp\left(-\frac{1}{2}\left(\nu_k + (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right)\tau_n\right) \\ &\propto \mathrm{Ga}(\tau_n; \alpha_{nk}, \beta_{nk}),\end{aligned}$$

where

$$\begin{aligned}\alpha_{nk} &= \frac{\nu_k + D}{2} \\ \beta_{nk} &= \frac{1}{2}\left(\nu_k + (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right).\end{aligned}$$

(d) We will first simplify the expected complete log-likelihood,

$$\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\theta}_{\text{old}}) = \sum_{n=1}^{N} \mathbb{E}_{p(z_n,\tau_n|\boldsymbol{x}_n;\boldsymbol{\theta}_{\text{old}})}\left[\log p(z_n,\tau_n,\boldsymbol{x}_n;\boldsymbol{\theta})\right].$$

First recall that,

$$\begin{aligned}
&\log p(z_n,\tau_n,\boldsymbol{z}_n;\boldsymbol{\theta})\\
&= \sum_{k=1}^{K} I[z_n=k]\log p(z_n=k,\tau_n,\boldsymbol{z}_n;\boldsymbol{\theta})\\
&= \sum_{k=1}^{K} I[z_n=k]\left(\log\pi_k + \log\text{Ga}(\tau_n;\nu_k/2,\nu_k/2) + \log\mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_k,\tau_n^{-1}\boldsymbol{\Sigma}_k)\right)\\
&= \sum_{k=1}^{K} I[z_n=k]\Bigg(\log\pi_k + \frac{\nu_k}{2}\log\left(\nu_k/2\right) - \log\Gamma(\nu_k/2) + (\nu_k/2-1)\log(\tau_n)\\
&- \frac{\nu_k}{2}\tau_n - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log\det(\tau_n^{-1}\boldsymbol{\Sigma}_k) - \frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Bigg).
\end{aligned}$$

If we drop terms that don't depend on $\boldsymbol{\theta}$, we have

$$\begin{aligned}
&\log p(z_n,\tau_n,\boldsymbol{z}_n;\boldsymbol{\theta})\\
&= \sum_{k=1}^{K} I[z_n=k]\Bigg(\log\pi_k + \frac{\nu_k}{2}\log\left(\nu_k/2\right) - \log\Gamma(\nu_k/2) + (\nu_k/2-1)\log(\tau_n)\\
&- \frac{\nu_k}{2}\tau_n - \frac{1}{2}\log\det(\tau_n^{-1}\boldsymbol{\Sigma}_k) - \frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Bigg).
\end{aligned}\tag{4}$$

On the exam, we are only asked to do the M-step for $\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K$. We can therefore drop terms that do not depend on $\boldsymbol{\mu}_k$ or $\boldsymbol{\Sigma}_k$. This gives,

$$\begin{aligned}
&\log p(z_n,\tau_n,\boldsymbol{z}_n;\boldsymbol{\theta})\\
&= \sum_{k=1}^{K} I[z_n=k]\left(-\frac{1}{2}\log\det(\tau_n^{-1}\boldsymbol{\Sigma}_k) - \frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\\
&= \sum_{k=1}^{K} I[z_n=k]\left(\frac{D}{2}\log\tau_n - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_k) - \frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\\
&= \sum_{k=1}^{K} I[z_n=k]\left(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k) - \frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right),
\end{aligned}$$

where in the last line we dropped the term $\frac{D}{2}\log\tau_n$. We will now take an expectation with respect to $p(z_n,\tau_n \mid \boldsymbol{x}_n;\boldsymbol{\theta}_{\text{old}})$. By parts (b) and (c) we have,

$$\mathbb{E}_{p(z_n,\tau_n|\boldsymbol{x}_n;\boldsymbol{\theta}_{\text{old}})}\left[I[Z_n=k]\right] = \omega_{nk}^{\text{old}},$$

and

$$\begin{aligned}\mathbb{E}_{p(z_n,\tau_n|\boldsymbol{x}_n;\boldsymbol{\theta}_{\mathrm{old}})}\left[I[Z_n=k]\tau_n\right] &= \omega_{nk}^{\mathrm{old}}\mathbb{E}_{\mathrm{Ga}(z_n;\alpha_{nk}^{\mathrm{old}},\beta_{nk}^{old})}\left[\tau_n\right] \\ &= \frac{\omega_{nk}^{\mathrm{old}}\alpha_{nk}^{\mathrm{old}}}{\beta_{nk}^{\mathrm{old}}}.\end{aligned}$$

Thus, up to a constant independent of $\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K$ we have

$$\begin{aligned}&\mathbb{E}_{p(z_n,\tau_n|\boldsymbol{x}_n;\boldsymbol{\theta}_{\mathrm{old}})}\left[\log p(z_n,\tau_n,\boldsymbol{z}_n;\boldsymbol{\theta})\right] \\ &= \sum_{k=1}^K \mathbb{E}_{p(z_n,\tau_n|\boldsymbol{x}_n;\boldsymbol{\theta}_{\mathrm{old}})}\left[I[z_n=k]\Big(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k)-\frac{\tau_n}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\Big)\right] \\ &= \sum_{k=1}^K \omega_{nk}^{\mathrm{old}}\left(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k)-\frac{\alpha_{nk}^{\mathrm{old}}}{2\beta_{nk}^{\mathrm{old}}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\end{aligned}$$

And so, up to terms independent of $\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K$, we have

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\theta}_{\mathrm{old}}) \\ &= \sum_{n=1}^N\sum_{k=1}^K \omega_{nk}^{\mathrm{old}}\left(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k)-\frac{\alpha_{nk}^{\mathrm{old}}}{2\beta_{nk}^{\mathrm{old}}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right) \\ &= \sum_{k=1}^K\sum_{n=1}^N \omega_{nk}^{\mathrm{old}}\left(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k)-\frac{\alpha_{nk}^{\mathrm{old}}}{2\beta_{nk}^{\mathrm{old}}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right).\end{aligned}$$

We see that the M-step for $\{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}_{k=1}^K$ splits over k. And so,

$$\boldsymbol{\mu}_k^\star,\boldsymbol{\Sigma}_k^\star \leftarrow \operatorname*{argmax}_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}\sum_{n=1}^N \omega_{nk}^{\mathrm{old}}\left(-\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k)-\frac{\alpha_{nk}^{\mathrm{old}}}{2\beta_{nk}^{\mathrm{old}}}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right).$$

If we differentiate with respect to $\boldsymbol{\mu}_k$, we see that $\boldsymbol{\mu}_k^\star$ solves,

$$\sum_{n=1}^N \frac{\omega_{nk}^{\mathrm{old}}\alpha_{nk}^{\mathrm{old}}}{\beta_{nk}^{\mathrm{old}}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k^\star)=0\in\mathbb{R}^D.$$

And so,

$$\boldsymbol{\mu}_k^\star\left(\sum_{n=1}^N \frac{\omega_{nk}^{\mathrm{old}}\alpha_{nk}^{\mathrm{old}}}{\beta_{nk}^{\mathrm{old}}}\right)=\sum_{n=1}^N \frac{\omega_{nk}^{\mathrm{old}}\alpha_{nk}^{\mathrm{old}}}{\beta_{nk}^{\mathrm{old}}}\boldsymbol{x}_n.$$

If we define

$$\gamma_{nk}=\frac{\omega_{nk}^{\mathrm{old}}\alpha_{nk}^{\mathrm{old}}/\beta_{nk}^{\mathrm{old}}}{\sum_{m=1}^N \omega_{mk}^{\mathrm{old}}\alpha_{mk}^{\mathrm{old}}/\beta_{mk}^{\mathrm{old}}},$$

then

$$\boldsymbol{\mu}^\star = \sum_{n=1}^N \gamma_{nk} \boldsymbol{x}_n.$$

Likewise, plugging in $\boldsymbol{\mu}^\star$ and differentiating with respect to $\boldsymbol{\Sigma}_k^{-1}$ gives the first order condition,

$$\frac{1}{2}\sum_{n=1}^N \omega_{nk}^{\text{old}} \boldsymbol{\Sigma}_k^\star - \frac{\omega_{nk}^{\text{old}} \alpha_{nk}^{\text{old}}}{\beta_{nk}^{\text{old}}} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^\star)(\boldsymbol{x}_n - \boldsymbol{\mu}_k^\star)^\top = 0 \in \mathbb{R}^{D \times D}.$$

And so

$$\boldsymbol{\Sigma}_k^\star = \frac{1}{\sum_{n=1}^N \omega_{nk}^{\text{old}}} \sum_{n=1}^N \frac{\omega_{nk}^{\text{old}} \alpha_{nk}^{\text{old}}}{\beta_{nk}^{\text{old}}} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^\star)(\boldsymbol{x}_n - \boldsymbol{\mu}_k^\star)^\top.$$

Optional: In terms of exam requirements, we are now done. We have calculated the M-step for $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$ which we did by isolating the part of $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}})$ that depended on these parameters. We can do a similar thing for $\boldsymbol{\pi}$. Using (4) we can compute $\log p(z_n, \tau_n, \boldsymbol{x}_n; \boldsymbol{\theta})$ as a function of only $\boldsymbol{\pi}$ specifically,

$$\log p(z_n, \tau_n, \boldsymbol{x}_n; \boldsymbol{\theta}) = C + \sum_{k=1}^K I[Z_n = k] \log(\pi_k).$$

Thus, as a function of $\boldsymbol{\pi}$,

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{old}}) \\
&= \sum_{n=1}^N \mathbb{E}_{p(z_n, \tau_n | \boldsymbol{x}_n; \boldsymbol{\theta}^{\text{old}})} \left[\sum_{k=1}^K I[Z_n = k] \log(\pi_k) \right] \\
&= \sum_{n=1}^N \sum_{k=1}^K \omega_{nk}^{\text{old}} \log(\pi_k).
\end{aligned}$$

Thus,

$$\pi_k^\star = \frac{\sum_{n=1}^N \omega_{nk}}{\sum_{j=1}^K \sum_{n=1}^N \omega_{nj}}.$$

We can be daring and attempt the ν_k updates as well. Up to a constant that does not depend on ν_k we have

$$\begin{aligned}
&p(z_n, \tau_n, \boldsymbol{x}_n; \boldsymbol{\theta}) \\
&= C + \sum_{k=1}^K I[z_n = k] \left(\frac{\nu_k}{2} \log(\nu_k/2) - \log\Gamma(\nu_k/2) + \nu_k/2 \log(\tau_n) - \frac{\nu_k}{2}\tau_n \right).
\end{aligned}$$

Let $v_k = \nu_k/2$. Thus, as a function of v_k we have

$$\begin{aligned} &p(z_n, \tau_n, \boldsymbol{x}_n; \boldsymbol{\theta}) \\ &= C + \sum_{k=1}^{K} I[z_n = k] \left(v_k \log(v_k) - \log \Gamma(v_k) + v_k \log(\tau_n) - v_k \tau_n \right). \end{aligned}$$

We also have

$$\begin{aligned} \mathbb{E}_{p(z_n, \tau_n | \boldsymbol{x}_n; \boldsymbol{\theta}^{\text{old}})} \left[I[Z_n = k] \right] &= \omega_{nk}^{\text{old}}, \\ \mathbb{E}_{p(z_n, \tau_n | \boldsymbol{x}_n; \boldsymbol{\theta}^{\text{old}})} \left[I[Z_n = k] \tau_n \right] &= \omega_{nk}^{\text{old}} \alpha_{nk}^{\text{old}} / \beta_{nk}^{\text{old}}, \\ \mathbb{E}_{p(z_n, \tau_n | \boldsymbol{x}_n; \boldsymbol{\theta}^{\text{old}})} \left[I[Z_n = k] \log(\tau_n) \right] &= \omega_{nk}^{\text{old}} \mathbb{E}_{\text{Ga}(\tau_n; \alpha_{nk}^{\text{old}}, \beta_{nk}^{\text{old}})} \left[\log(\tau_n) \right] \\ &= \omega_{nk}^{\text{old}} \left(\psi\left(\alpha_{nk}^{\text{old}}\right) - \log\left(\beta_{nk}^{\text{old}}\right) \right), \end{aligned}$$

where

$$\psi(y) = \frac{d}{dy} \log \Gamma(y),$$

is the digamma function. Thus, as a function of $\boldsymbol{v} = (v_k)_{k=1}^{K}$, we have

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{nk}^{\text{old}} \left(v_k \log(v_k) - \log \Gamma(v_k) + v_k \left(\psi\left(\alpha_{nk}^{\text{old}}\right) - \log\left(\beta_{nk}^{\text{old}}\right) \right) - v_k \alpha_{nk}^{\text{old}} / \beta_{nk}^{\text{old}} \right). \end{aligned}$$

And so $v_k^\star$ maximizes

$$(v_k \log(v_k) - \log \Gamma(v_k)) \sum_{n=1}^{N} \omega_{nk} + v_k \sum_{n=1}^{N} \omega_{nk}^{\text{old}} \left(\psi\left(\alpha_{nk}^{\text{old}}\right) - \log\left(\beta_{nk}^{\text{old}}\right) - \alpha_{nk}^{\text{old}} / \beta_{nk}^{\text{old}} \right).$$

Define,

$$\begin{aligned} A_k &= \sum_{n=1}^{N} \omega_{nk}, \\ B_k &= \sum_{n=1}^{N} \omega_{nk}^{\text{old}} \left(\psi\left(\alpha_{nk}^{\text{old}}\right) - \log\left(\beta_{nk}^{\text{old}}\right) - \alpha_{nk}^{\text{old}} / \beta_{nk}^{\text{old}} \right). \end{aligned}$$

We are trying to maximize

$$(v_k \log(v_k) - \log \Gamma(v_k)) A_k + v_k B_k.$$

Which has the first order condition

$$\left(\log(v_k^\star) - 1 - \psi(v_k^\star) \right) A_k + B_k = 0$$

So

$$\log(v_k^\star) - \psi(v_k^\star) = \frac{A_k - B_k}{A_k}.$$

A plot in Mathematic shows that $v \mapsto \log(v) - \psi(v)$ is strictly decreasing and so the above equation has a unique solution $v_k^\star$. This means that, with a bit of work, we can implement a full EM algorithm for all the parameters $\boldsymbol{\theta}$.

Question 6: Principal component regression with ℓ_1-penalties

References