Suggestions, PS 13

14.14. Background: As usual we have a random variable (rv) $X = (X_1, ..., X_p)$ from which we have generated a new rv $S = S(_1, ..., S_p)^T$ satisfying the relationship $X = AS$, where A is a $p \times p$ matrix. The matrix A is chosen so that S has covariance I (i.e., its components are independent standard normal). The implied model thus is $X = AS$, or

$$\begin{aligned}&X_1 = a_{11}S_1 + a_{12}S_2 + ... + a_{1p}S_p\\ &\vdots\\ &X_p = a_{p1}S_1 + a_{p2}S_2 + ... + a_{pp}S_p\end{aligned}$$

We can also choose to keep only the first $q < p$ terms in each equation, interpreting the remainder as a random error :

$$\begin{aligned}&X_1 = a_{11}S_1 + a_{12}S_2 + ... + a_{1q}S_q + \epsilon_1\\ &\vdots\\ &X_p = a_{p1}S_1 + a_{p2}S_2 + ... + a_{pq}S_q + \epsilon_p\end{aligned}$$

The implied model is then

$$X = AS\ ,$$

where now

$$S = (S_1, ..., S_q)^T$$

and $A = A_q$ has only q of its original columns.

We then have (with $\epsilon = (\epsilon_1, ..., \epsilon_p)^T$)

$$\text{Cov}\, X = A\ \text{Cov}\ S\ A^T + \text{Cov}\,\epsilon = AA^T + \text{Cov}\ \epsilon = AA^{\mathrm{T}} + D_\epsilon\ , \quad (1)$$

where D_ϵ is the diagonal matrix with entries $V(\epsilon_i)$.

To formulate (1) as a correlation identity, form the diagonal matrix W with diagonal entries

$$W_{ii} = \sqrt{V(X_i)} = \sqrt{(AA^T)_{ii}V(\epsilon_i)}\ .$$

Thus the correlation matrix of X is

$$\rho(X) = BB^T + W^{-1}D_\epsilon W^{-1},$$

where $B = W^{-1}A$ (why?).

14.21. For a graph G with m connected components, (not connected to each other by any edges), we can number the vertices in the first component as $v_{11}, ..., v_{1n_i}$, the second $v_{21}, ..., v_{2n_2}$ through the m^{th} as $v_{m1}, v_{m2}, ..., v_{mn_m}$. Given function $f(v)$ on the vertices of G then $\mathbf{f} = \begin{bmatrix} f(v_{11}) \\ \vdots \\ f(v_{mn_m}) \end{bmatrix}$ has the ordered values in it. Show using this ordering of the vertices, the weight matrix $W = (w_{ij})$ will have a block structure

$W = \begin{bmatrix} W_1 & 0 & 0 & 0 \\ 0 & W_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & W_m \end{bmatrix}$, with W_m the matrix restricted to component . Recall ,

$L = G - W$ where $G = \begin{bmatrix} g_{11} & 0 & \dots & 0 \\ 0 & g_{12} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & g_{mn_m} \end{bmatrix}$ has the degrees g of all vertices in

all components. Recall for any vertex i, $g = \sum\limits_{j:j \neq i} w_{ij}$, - why? Show thus that the first row of L (along with other rows) sums to $g_{11} - \sum\limits_{j \neq 1} w_{ij} = 0$. [Note sometimes we are using just a single index i to number the vertices ν_i (rather than ν_{mn} referring also to the cluster m) when it's convenient $-$ in this case w_{ij} has the usual two indices i and j.]

Show also that L inherits the block structure of G and W, so

$$L = \begin{bmatrix} L_1 & 0 & 0 & 0 \\ 0 & L_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & L_m \end{bmatrix}.$$

For a vector $\mathbf{f}$ of an indicator function of the first component, show the first n_1 entries in $\mathbf{f}$ are 1 and the rest are 0. Divide $\mathbf{f}$ into blocks corresponding to graph components, so

$\mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_m \end{bmatrix}$ and show $L\mathbf{f} = \begin{bmatrix} L_1\mathbf{f}_1 \\ L_2\mathbf{f}_2 \\ \vdots \\ L_m\mathbf{f}_m \end{bmatrix}$. For this $\mathbf{f}$ the first block $\mathbf{f}_1 = \mathbf{1}_{n_1}$ (a column of 1's), while the remaining blocks $\mathbf{f}_{12}, ...\mathbf{f}_{n_m} = \mathbf{0}$. Thus (using the fact the rows of block L_1 add up to 0) show $L_1\mathbf{f} = \mathbf{0}$, and similarly $L_i\mathbf{f}_i = \mathbf{0}$ for $i > 1$. Show $L\mathbf{f} = \mathbf{0}$, so $\mathbf{f}$ is a 0-eigenvector of L, and the same holds for the indicator of any other component besides the first.

Now show the **only** 0-eigenvectors of L are the indicators of some component of the graph. Let $\mathbf{f}$ be an eigenvector of L (with positive or negative entries) with $L\mathbf{f} = 0$. Then show

$$\mathbf{f}^t L\mathbf{f} = \sum_i g_i\, f_i^2 - \sum_{i,j} w_{ij}\, f_i\, f_j = \sum_{i,j} w_{ij}(f_i^2 - f_i f_j)$$

$$= \sum_{i,j} w_{ij}\Big(\frac{1}{2}(f_i^2 + f_j^2) - f_i f_j\Big).$$

$$= -\frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2 .$$

Now show that if f_i and f_j represent points in the same component of the graph, then $f_i - f_j$. How does this prove the result?

18.2. Background: This problem shows that the shrunken centroids estimator, as is implied in its description, is a principled solution to an optimization problem. Recall that the point of shrunken centroids is to 'move' the centers of the LDA of a linear discriminant analysis (LDA) classifier closer together, so that only strongly separated features (dimensions) of the centers stay separated. The idea is that if the centers of two LDA classes are weakly separated along a given dimension, then they should be coalesced (along that dimension only), so that this dimension no longer plays a role in separating the two classes.

Shrunken centroids has a rule for moving the centroids toward each other that seems somewhat ad hoc. What this problem shows is that the new locations of the centroids (class centers) can in fact be chosen as a solution to an optimization problem, namely (18.55). In LDA the estimate $\widehat{\mu} + \widehat{\mu}_k$ of the centroid a given class k is just the mean $\overline{\mathbf{x}}_k = (\overline{x}_{k1}, ..., \overline{x}_{kp})$ of the data points x_i in class k. This mean can be thought of as obtained by minimizing the first part (the triple sum) in (18.55). However we wish here to treat the centroids just like the regression coefficients in Lasso, in which we try to make the centroids closer (i.e., make the $\widehat{\mu}_k$ smaller) by using a simple rule, here by introducing a penalty for the sizes of the $\widehat{\mu}_k$ (the double sum in (18.55)). The idea again is that noise will make some of the centroids large, so that shrinking them in a principled way will tend to quench the noisy ones but preserve the informative ones.

Notice the logic of this problem. We are assuming a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$ with feature vectors $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ that have components x_{ij} which are independent and come from an underlying fixed normal distribution $N(\mu_j + \mu_{jk}, \sigma_j^2)$ (with k denoting the class $k = y$ of the data point $\mathbf{x}_i$). The usual way of estimating the centroid $\mu_j + \mu_{jk}$ of the underlying normal for the class k is to take all the $\mathbf{x}_i$ in the class k and to find their mean. However, this method tries to 'shrink out' the noise from this estimate by imposing a penalty on large deviations μ_{jk} from the center. Thus we estimate our values $\widehat{\mu}, \widehat{\mu}_{jk}$ as

$$\widehat{\mu}_j, \widehat{\mu}_{jk} = (\arg\min_{\mu_j,\mu_{jk}} L = \frac{1}{2}\sum_j\sum_k\sum_{i\in C_k}\frac{(x_{ij}-\mu_j-\mu_{jk})^2}{s_j^2}$$

$$+\,\lambda\sum_j\sum_k\sqrt{N_k}\;\frac{|\mu_{jk}|}{s_j}$$

whose second term is a penalty for large μ_{jk} values. The goal of this problem is to show that the solution $(\widehat{\mu}_j, \widehat{\mu}_{jk})$ above is the same as the centroids that would have arisen from using the standard shrunken centroids method.

To minimize

$$L = \frac{1}{2}\sum_j\sum_k\sum_{i\in C_k}\frac{(x_{ij}-\mu_j-\mu_{jk})}{s_j^2} + \lambda\sum_j\sum_k\sqrt{N_k}\;\frac{|\mu_{jk}|}{s_j}, \qquad (*)$$

first solve for the minimum with respect to μ_j, assuming you know the μ_{jk}. Fixing all but μ_j, show

$$\mu_j = \frac{1}{N}\sum_k\sum_{i\in C_k}(x_{ij}-\mu_{jk)} = \overline{x}_j - \overline{\mu}_j \qquad (1a)$$

where $\overline{\mu}_j = \sum_k \frac{N_k}{N}\mu_{jk}$. Now to minimize with respect to μ_{jk}, with μ_j fixed: show that if

$$L_{jk} = \frac{1}{2}\sum_{i\in C_k}\frac{(x_{ij}-\mu_j-\mu_{jk})^2}{s_j^2} + \lambda\sqrt{N_k}\,\frac{|\mu_{jk}|}{s_j},$$

then $L = \sum_{j,k} L_{jk}$ and it suffices to minimize L_{jk} one at a time. First minimize over the range where $\mu_{jk} \geq 0$. Without the absolute value in (*), you can differentiate L $a$with respect to μ_{jk} to get

$$\frac{\partial}{\partial\mu_{jk}}L_{jk} = -N_k\,\frac{\overline{x}_{jk}-\mu_j-\mu_{jk}}{s_j^2} + \lambda\sqrt{N_k}\,\frac{1}{s_j} = 0$$

so

$$\mu_{jk} = \overline{x}_{jk} - \mu_j - \frac{s_j\lambda}{\sqrt{N_k}}. \qquad (1)$$

On the other hand, if (1) is non-positive (i.e less than or equal to 0), show

$$\mu_{jk} = 0 \qquad (2)$$

minimizes L. Indeed if the point in (1) is non-positive show the derivative $\frac{\partial L}{\partial \mu_{jk}}$ of L *without* the absolute value must be positive at all points to the right of this point, i.e., $\frac{\partial L}{\partial \mu_{jk}} L_{jk}$ is positive for *all* values of $\mu_{jk} \geq 0$. Thus show where $\mu_{jk} \geq 0$, $\mu_{jk} = 0$ minimizes L. Thus show L_{jk} is minimized by (1) if its right side is non-negative, and by 0 otherwise.

Similarly over the range $\mu_{jk} \leq 0$, show (1) is replaced by

$$\mu_{jk} = \overline{x}_{jk} - \mu_j + \frac{s_j \lambda}{\sqrt{N_k}} \tag{3}$$

and again if the value μ_{jk} in (3) is positive, it is replaced by 0 for the same reason as above. Thus conclude,

$$\mu_{jk} = \begin{cases} \overline{x}_{jk} - \mu_j - \frac{s_j\lambda}{\sqrt{N_k}} & \text{if this is positive} \\ \overline{x}_{jk} - \mu_j + \frac{s_j\lambda}{\sqrt{N_k}} & \text{if this is negative} \\ 0 & \text{otherwise} \end{cases}$$

$$= \text{sign}(\overline{x}_{jk} - \mu_j)\left(|\overline{x}_{jk} - \mu_j| - \frac{s_j \lambda}{\sqrt{N_k}}\right)_+ . \tag{4}$$

This solves the above optimization problem, but it is not yet in terms of the original data x_{ij}. The original system of equations for μ_j and μ_{jk} has been replaced by the system consisting of (4) together with the original equation

$$\mu_j = \overline{x}_j - \overline{\mu}_j \tag{4aa}$$

Now solve this system (4), (4aa) for μ_j and μ_{jk} in terms of the original data x_{ij}. Since we know that any solution of (4) and (1a) gives a minimum value of L in (*), we seek any solution. If you can find a candidate solution and verify that it satisfies (is consistent with) the system (4) and (1a), you are done.

To do this make a choice of μ_j and μ_{jk} that will satisfy both (4) and (4aa). Try using

$$\mu_j = \text{w-med}_{k,\sqrt{N_k}}(\overline{x}_{jk}). \tag{4a}$$

Here $\text{w-med}_{k,\sqrt{N_k}}\,(\overline{x}_{jk})$ denotes the weighted median over all k of the points $\overline{x}_{jk}$, with weights $\sqrt{N_k}$ (with j fixed). This is the point μ_j such that the sum of the weights $\sqrt{N_k}$ of the points $\overline{x}_{jk}$ to its right and the sum of the weights to its left are equal. Notice that if the N_k are all the same then μ_j is the standard median (with respect to k) of the points $\overline{x}_{jk}$. Then choose μ_{jk} as in (4). Now show both (4) and (4aa) are satisfied, so equations (4a) and (4) solve the optimization.

To do this it is clear (4) holds since it defines μ_{jk} . Now to show (1a) holds, i.e. that

$$\mu_j + \overline{\mu} = \overline{x}_j \; ; \qquad (5)$$

note that the $\frac{N_k}{N}$-weighted average of both sides of (4) over k gives $\overline{\mu}_j = \sum\limits_k \frac{N_k}{N}\mu_{jk}$ on the left. On the right, comparing (4) to its two possible values (1) and (3), show that because $\mu_j =$ w-med$_{k,\sqrt{N_k}}(\overline{x}_{jk})$ you have on the right

$$\sum_k \frac{N_k}{N}\left[(x_{jk} - \mu_j) \pm \frac{s_j \lambda}{\sqrt{N_k}}\right] = \sum_k \frac{N_k}{N}(x_{jk} - \mu_j) + \frac{1}{N}\sum_k \pm \sqrt{N_k}\, s_j \lambda$$

$$= \overline{x}_j - \mu_j + \frac{1}{N}\, s_j \lambda \sum_k \pm \sqrt{N_k}\, .$$

Show by the choice of μ_j the last sum over k vanishes since by definition the sum of the weights of the points to the right of μ_j (for which the sign in the sum is -) and to the left of (for which it is +) is 0. Thus show the sum over the right sides of (4) is $\overline{x}_j - \mu_j$. Thus show averaging (4) on both sides with weights $\frac{N_k}{N}$ gives $\overline{\mu}_j = \overline{x}_j - \mu_j$, verifying (5) holds.

Thus you have shown that (4a) and (4) solve the optimization of minimizing L. Notice this gives the same shrinkage as shrunken centroids, but toward the weighted median μ_j of the $\overline{x}_{kj}$ instead of the mean.

Now to interpret solution (4), show that if we interpret the $\mu_j + \mu_{jk}$ as the new centroids to be used in our regular discriminant functions $\delta_k(\cdot)$ (the regular centroids would be $\overline{x}_{kj}$), then in terms of the notation of the text (Section 18.2) you would have (with the suggested setting of $s_0 = 0$ and $m_k^2 = \frac{1}{N_k}$, along with the replacement of $\overline{x}_j$ by μ_j):

$$d_{jk} = \frac{\overline{x}_{kl} - \mu_j}{s_j/\sqrt{N_k}} \qquad (5a)$$

Recall that the shrinkage in 18.2 replaces d_{jk} in (5a) by

$$d^{'}_{kj} = \text{sign } (d_{kj})(|d_{kl}| - \Delta)_+ \; .$$

Show this means that we are replacing the distance $\overline{x}_{kl} - \overline{x}_j = d_{kj} s_j/\sqrt{N_k}$ (with the suggested choice of $s_0 = 0$ and $m_k^2 = 1/N_k$) between the old centroids $\overline{x}_{kj}$ and the mean $\overline{x}_j$ with a new distance

new centroid displacement

$$= d^{'}_{kj}\, s_j/\sqrt{N_k} = \text{sign}(d_{kj})(|d_{kj}| - \Delta)_+\, s_j/\sqrt{N_k}$$

$$= \text{sign } (\overline{x}_{kj} - \overline{x}_j)(|\overline{x}_{kj} - \mu_j| - \Delta\, s_j/\sqrt{N_k}) \; . \qquad (5b)$$

Show this compares with our calculation earlier, which gives the replacement of the old centroid displacement $\overline{\mu}_{jk} = \overline{x}_{kj} - \overline{x}_j$ with the new one

$$\text{new centroid displacement } = \mu_{jk} = \text{sig}n(\overline{x}_{jk} - \mu_j)\left(|\overline{x}_{jk} - \mu_j| - \frac{s_j\lambda}{\sqrt{N_k}}\right)_+ \qquad (6)$$

Comparing (5b) and (6), show we must have $\Delta = \lambda$, and indeed with the replacements $s_0 = 0$ and $m_k^2 = 1/N_k$, the new implied centroids are the same (however, replacing the mean $\overline{x}_j$ by the median μ_j) using both methods.

18.4. Background: We are effectively trying to reduce the nominal dimensionality of a dataset from p dimensions to N dimensions. Recall that N datapoints within a feature space $F = \mathbb{R}^p$ of p dimensions effectively live in a hyperplane of $N-1$ dimensions $-$ so rotating a hyperplane of N dimensions into the proper position in $\mathbb{R}^p$ should be sufficient to contain the entire dataset. Then re-representing the dataset within this N-dimensional hyperplane is what is accomplished in the new data matrix R. That is, the row vectors $\mathbf{r}_i^T$ of are a re-representation of the data row vectors $\mathbf{x}^T$ making up the data matrix X , so that their geometry (and hence resulting inferences) are preserved.

The new $N \times N$ data matrix R is obtained via the singular value decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{R}\mathbf{V}^T ,$$

with $\mathbf{R} = \mathbf{U}\mathbf{D}$.

If $\widehat{\beta}$ represents a ridge regression coefficient based on the original dataset, then we have shown that

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}, \qquad (1)$$

where $\mathbf{y}$ is the set of measured values in the dataset. We wish to show that this equals

$$. \qquad \widehat{\beta} = (V(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\,\mathbf{R}^T\mathbf{y} \qquad (2)$$

To show that (1) = (2), show from the form of $\mathbf{V}$ that

$$\mathbf{X}^T\mathbf{R}^{T^{-1}}\,(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{V},$$

and right multiply by $(\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\mathbf{R}^T$ and left multiply by $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$.