Biostat 216 Homework 8

Due Nov 29 Friday @ 11:59pm

Submit a PDF (scanned/photographed from handwritten solutions, or converted from RMarkdown or Jupyter Notebook) to Gracescope on BruinLearn.

- Q1. (**MLE of multivariate normal model**) Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^p$ be iid samples from a p-dimensional multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where the mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ are unkonwn parameters. The log-likelihood is
$$\ell(\boldsymbol{\mu}, \boldsymbol{\Omega}) = -\frac{n}{2} \log \det \boldsymbol{\Omega} - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) - \frac{n}{2} \log 2\pi.$$
Show that the maximum likelihood estimate (MLE) is
$$\begin{aligned} \widehat{\boldsymbol{\mu}} &= \frac{\sum_{i=1}^n \mathbf{y}_i}{n} \\ \widehat{\boldsymbol{\Omega}} &= \frac{\sum_{i=1}^n (\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'}{n}. \end{aligned}$$
That is to show that $\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Omega}}$ maximize ℓ.

  Hint: Use the first order optimality condition to find $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Omega}}$. To check the optimality of $\widehat{\boldsymbol{\Omega}}$, use its Cholesky factor.

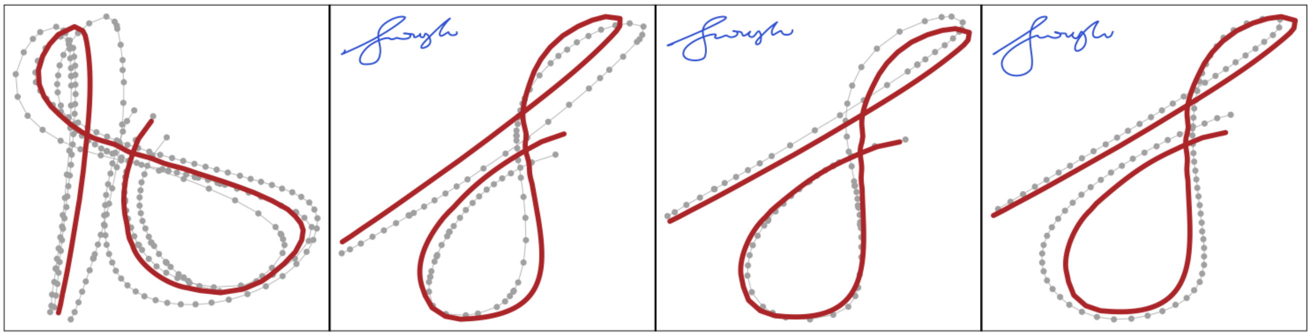
- Q2. (**Smallest matrix subject to linear constraints**) Find the matrix $\mathbf{X}$ with the smallest Frobenius norm subject to the constraint $\mathbf{XU} = \mathbf{V}$, assuming $\mathbf{U}$ has full column rank.

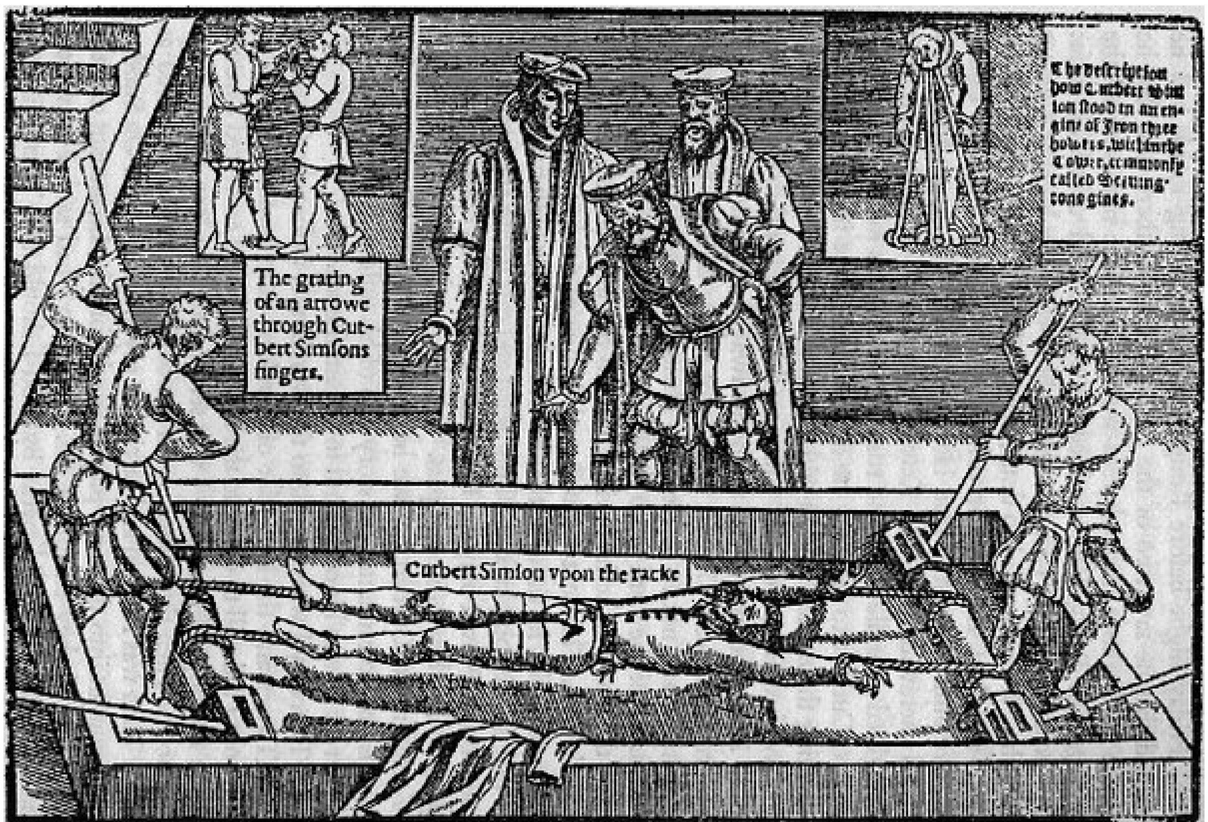
  Hint: Write down the optimization problem and use the method of Lagrange multipliers.

- Q3. (**Minimizing a convex quadratic form over manifold**) $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix. Find a matrix $\mathbf{U} \in \mathbb{R}^{n \times r}$ with orthonomal columns that maximizes $\text{tr}(\mathbf{U}'\mathbf{AU})$. That is to
$$\begin{array}{ll} \text{maximize} & \text{tr}(\mathbf{U}'\mathbf{AU}) \\ \text{subject to} & \mathbf{U}'\mathbf{U} = \mathbf{I}_r. \end{array}$$
This result generalizes the fact that the top eigenvector of $\mathbf{A}$ maximizes $\mathbf{u}'\mathbf{Au}$ subject to constraint $\mathbf{u}'\mathbf{u} = 1$.

  Hint: Use the method of Lagrange multipliers.

- Q4. (**Procrustes rotation** (https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.orthogonal_procrustes.html#scipy.linalg.ort
  The Procrustes problem studies how to properly align images.





'A true description of the racking and cruell handeling of Cutbert Simson in the Tower.'

Let matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times p}$ record n points on the two shapes. Mathematically we consider the problem

$$\text{minimize}_{\beta, \mathbf{O}, \mu} \quad \|\mathbf{X} - (\beta \mathbf{Y} \mathbf{O} + \mathbf{1}_n \mu^T)\|_{\text{F}}^2,$$

where $\beta > 0$ is a scaling factor, $\mathbf{O} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix, and $\mu \in \mathbb{R}^p$ is a vector of shifts. Here $\|\mathbf{M}\|_{\text{F}}^2 = \sum_{i,j} m_{ij}^2$ is the squared Frobenius norm. Intuitively we want to rotate, stretch and shift the shape $\mathbf{Y}$ to match the shape $\mathbf{X}$ as much as possible.

  - Q4.1 Let $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ be the column mean vectors of the matrices and $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ be the versions of these matrices centered by column means. Show that the solution $(\hat{\beta}, \hat{\mathbf{O}}, \hat{\mu})$ satisfies

    $$\hat{\mu} = \bar{\mathbf{x}} - \hat{\beta} \cdot \hat{\mathbf{O}}^T \bar{\mathbf{y}}.$$

    Therefore we can center each matrix at its column centroid and then ignore the location completely.
  - Q4.2 Derive the solution to

$$\text{minimize}_{\beta,\mathbf{O}} \quad \|\tilde{\mathbf{X}} - \beta\tilde{\mathbf{Y}}\mathbf{O}\|_{\text{F}}^2$$

using the SVD of $\tilde{\mathbf{Y}}^T\tilde{\mathbf{X}}$.

- Q5. (**Ridge regression** (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html)) One popular regularization method in machine learning is the ridge regression, which estimates regression coefficients by minimizing a penalized least squares criterion

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2.$$

Here $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ are fixed data. $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients to be estimated.

  - Q5.1 Show that, regardless the shape of $\mathbf{X}$, there is always a unique global minimum for any $\lambda > 0$ and the ridge solution is given by

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

  - Q5.2 Express ridge solution $\widehat{\boldsymbol{\beta}}(\lambda)$ in terms of the singular value decomposition (SVD) of $\mathbf{X}$.
  - Q5.3 Show that (i) the ℓ_2 norms of ridge solution $\|\widehat{\boldsymbol{\beta}}(\lambda)\|_2$ and corresponding fitted values $\|\widehat{\mathbf{y}}(\lambda)\|_2 = \|\mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2$ are non-increasing in λ and (ii) the ℓ_2 norm of the residual vector $\|\mathbf{y} - \widehat{\mathbf{y}}(\lambda)\|_2$ is non-decreasing in λ.
  - Q5.4 Let's address how to choose the optimal tuning parameter λ. Let $\widehat{\boldsymbol{\beta}}_k(\lambda)$ be the solution to the ridge problem

$$\text{minimize} \ \frac{1}{2}\|\mathbf{y}_{-k} - \mathbf{X}_{-k}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2,$$

where $\mathbf{y}_{-k}$ and $\mathbf{X}_{-k}$ are the data with the k-th observation taken out. The optimal λ can to chosen to minimize the cross-validation square error

$$C(\lambda) = \frac{1}{n}\sum_{k=1}^{n}[y_k - \mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_k(\lambda)]^2.$$

However computing n ridge solutions $\widehat{\boldsymbol{\beta}}_k(\lambda)$ is expensive. Show that, using SVD $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$,

$$C(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left[\frac{y_k - \sum_{j=1}^r u_{kj}\tilde{y}_j\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)}{1 - \sum_{j=1}^r u_{kj}^2\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)}\right]^2,$$

where $\tilde{\mathbf{y}} = \mathbf{U}^T\mathbf{y}$.

- Q6. (**Factor analysis** (https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.FactorAnalysis.html)) Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^p$ be iid samples from a multivariate normal distribution $N(\mathbf{0}_p, \mathbf{F}\mathbf{F}' + \mathbf{D})$, where $\mathbf{F} \in \mathbb{R}^{p \times r}$ and $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix with positive entries. We estimate the factor matrix $\mathbf{F}$ and diagonal matrix $\mathbf{D}$ by maximizing the log-likelihood function
$$\ell(\mathbf{F}, \mathbf{D}) = -\frac{n}{2} \ln \det(\mathbf{F}\mathbf{F}' + \mathbf{D}) - \frac{n}{2} \operatorname{tr}\left[(\mathbf{F}\mathbf{F}' + \mathbf{D})^{-1}\mathbf{S}\right] - \frac{np}{2} \ln 2\pi,$$
where $\mathbf{S} = n^{-1} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i'$.
  - Q5.1 We first show that, for fixed $\mathbf{D}$, we can find the maximizer $\mathbf{F}$ explicitly using SVD by the following steps.
    - Step 1: Take derivative with respect to $\mathbf{F}$ and set to 0 to obtain the first-order optimality condition.
    - Step 2: Reparameterize $\mathbf{H} = \mathbf{D}^{-1/2}\mathbf{F}$ and $\tilde{\mathbf{S}} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, and express the first-order optimality condition in terms of $\mathbf{H}$ and $\tilde{\mathbf{S}}$.
    - Step 3: Let $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ be its SVD. Show that columns of $\mathbf{U}$ must be r eigenvectors of $\tilde{\mathbf{S}}$.
    - Step 4: Identify which r eigenvectors of $\tilde{\mathbf{S}}$ to use in $\mathbf{U}$ and then derive the solution to $\mathbf{F}$.
  - Q5.2 Show that, for fixed $\mathbf{F}$, we can find the maximizer $\mathbf{D}$ explicitly. (Hint: first-order optimality condition.)

    Combining Q5.1 and Q5.2, a natural algorithm for finding the MLE of factor analysis model is to alternately update $\mathbf{F}$ and $\mathbf{D}$ until convergence.