Midterm, BE188

February 15, 2018

Question 1 (15 pts)

a) Suppose that the probability density function of a distribution is $p(x) = a(1 - x^3)$ for $0 < x < 1$ and $p(x) = 0$ otherwise. Based on the properties of probability distributions, what is a? What is $p(x < \frac{1}{2})$?

$$1 = \int_0^1 p(x)dx = a\int_0^1 (1 - x^3)dx, \qquad a = \frac{4}{3}$$

$$p(x < \frac{1}{2}) = \int_0^{1/2} p(x)dx = \frac{31}{48}$$

b) What is the mean of this distribution?

$$\mu = \int_0^1 xp(x)dx = \frac{2}{5}$$

c) Sketch the PDF. How would you qualitatively describe the skew of this distribution? What does this mean about values as compared to the mean? *Will show plot in class. Skewed positively. Most values are greater than the mean.*

d) What are three things (total) you can say about the sampling distributions of the mean for $N = 1$ and $N = 5$?

(1) The sampling distribution for N=1 is the same as the original distribution. (2) The sampling distribution for N=5 tends towards a normal distribution. (3) The variance of the sampling distribution for N=5 is less than that for N=1.

e) How could you test whether a set of points follow this distribution? (Very briefly describe.) *You could use a KS-test.*

Question 2 (20 pts)

You are designing a medical device to provide measurements of blood oxygenation from skin spectroscopy measurements performed on the wrist. You know that the device provides a voltage that is proportional to blood oxygenation, but have to calibrate it for each patient to values measured separately.

a) What method could you use to quickly determine this conversion factor from your calibration points? *Oridinary least squares.*

b) Your team asks you to provide design a scheme whereby the device provides feedback as to whether new calibration points would be helpful. How could you determine this from the calibration points you have ($[(V_1, O_1), (V_2, O_2), (V_3, O_3), ...]$) so far? *You could bootstrap your model using all of the points you have been given so far, and look at the distribution of conversion values you obtain. When this falls below a certain threshold you have enough calibration points.*

c) You have many calibration points (say N > 30), so you know that you can ignore variance in the model (i.e. if you ran bootstrapping, your β terms come out as virtually identical). What can you say about your confidence in where new calibration points will be distributed? *We can expect new points to fall within a normal distribution around the line of prediction. That normal distribution will have a mean of 0 and standard deviation equal to the standard deviation of the residual during fitting.*

d) A team member suggests that the voltage-oxygenation relationship is log-linear instead of linear, and so suggests using $\log(V)$ instead. When would this be alright? What is an alternative approach? Are there any concerns with calculating the answer in either case? *This would be alright iff the error can be expected to be log-normally distributed. If not, an alternative approach would be NNLSQ, using* $y = \log(V)$ *as the relationship. One concern with this is it's not guaranteed to give the globally optimal answer.*

e) In version 1 of the device you used a single value as input, calculated from two wavelengths outside of your model. In version 2 your team is interested in whether the full spectroscopy data (200 wavelengths simultaneously) can be used for a more reliable measurement. You're allowed to require up to 20 calibration points. Describe how you would use these to calibrate your model. What assumptions are you making? How would you compare version 2 to version 1? *You could use PLSR, in effect assuming that the covariance between the spectroscopy data and blood oxygenation will be most useful. You would build a model with the matrix of spectroscopy data by calibration point as input, and a vector of calibration points as output. You could compare the two model versions by evaluating their crossvalidation performance.*

Question 3 (15 pts)

A mammogram is a diagnostic imaging test for cancer with a sensitivity and specificity of roughly 80% and 95%, respectively. A completely healthy, asympomatic 40 year-old woman shows a positive test and is recommended for a biopsy. The incidence of breast cancer for her age is roughly 1 per 1000 women.

a) Write out Bayes' law, and rewrite the equation for the probability of the woman having a tumor given her positive test.

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

$$p(\text{tumor} \mid \text{positive test}) = \frac{p(\text{positive test} \mid \text{tumor})p(\text{tumor})}{p(\text{positive test})}$$

b) Sensitivity is true positives over all positives, while specificity is true negatives over all negatives. Therefore, the false positive rate is $1 -$ specificity and the false negative rate is $1 -$ sensitivity. How many false and true positives are expected in a cohort of 1000 tests?

False positives: $999 \times 0.05 = 49.95$. True positives: $1 \times 0.8 = 0.8$.

c) Calculate the probability of the woman having breast cancer, given her positive test result.

$$p(\text{tumor} \mid \text{positive test}) = \frac{1 \times 0.8}{1 \times 0.8 + 999 \times 0.05} = \frac{0.8}{0.8 + 49.95} = 1.57\%$$

If a data point y follows the Poisson distribution with rate parameter θ, then the probability of a single observation y is $\frac{\theta^y e^{-\theta}}{y!}$, for $y = 0, 1, 2, \ldots$. You are given data points $[0, 1, 0, 0, 1]$ independently drawn from a Poisson distribution with parameter θ. Your prior for θ is $p(\theta) = \theta^{-2}$.

d) Write down the log-likelihood of the data as a function of θ.

$$p(\text{data} \mid \theta) = \left(e^{-\theta}\right)^3 \left(\theta e^{-\theta}\right)^2 = \theta^2 e^{-5\theta}$$

$$\log(p(\text{data} \mid \theta)) = 2\log(\theta) - 5\theta$$

e) Write out the expression for the Bayesian expectation for the next value to be α, given these previous observations. You only need to write down the expression, not integrate it.

$$p(\theta \mid \text{data}) = p(\text{data} \mid \theta)p(\theta) = \theta^2 e^{-5\theta}\theta^{-2} = e^{-5\theta}$$

Normalizing the probability by:

$$p(\forall) = b\int_0^\infty e^{-5\theta}d\theta = 1, \qquad b = 5$$

$$p(\alpha) = b\int_0^\infty \frac{\theta^\alpha e^{-\theta}}{\alpha!}e^{-5\theta}d\theta$$

$$p(\alpha) = \frac{5}{\alpha!}\int_0^\infty \theta^\alpha e^{-6\theta}d\theta$$

Question 4 (20 pts)

a) What is crossvalidation and what does it evaluate? *Crossvalidation is the process by which one simulates the existance of new data by leaving out a portion of a data set, training a model on the remaining portion, then evaluating the model's ability to predict data not previously observed (held out). In this way it evaluates the prediction performance of a model.*

b) Outline the steps to performing crossvalidation. *(1) Leave out a portion of data. (2) Fit a model from scratch, in no way based on the left out data. (3) Compare the model to the left out portion. (4) Repeat with a new portion of data left out.*

c) How do predictions from crossvalidation necessarily differ from fitting a full model? *When performing crossvalidation, one's model will necessarily be trained on a reduced number of data points. Therefore, crossvalidation will overestimate the prediction error.*

d) Why are multiple folds necessary? *Without multiple folds, the model error is dependent upon exactly which points were left out as the validation set. Averaging over multiple left out sets minimizes the contribution of test set selection variance.*

e) What does bootstrapping pretend to do with your data? *Bootstrapping pretends to repeatedly build an entirely new dataset of identical size from the same underlying distribution.*

f) Outline the steps for performing bootstrapping. *(1) Resample one's original dataset with replacement (allowing for duplicate observations). (2) Build a model. (3) Record the built model. (4) Repeat the process a large number of times to build a distribution of models.*

Question 5 (15 pts)

Lek *et. al.* examined the protein-coding variation in 60,706 humans. In part of their analysis they presented their data as a principal components plot as shown.

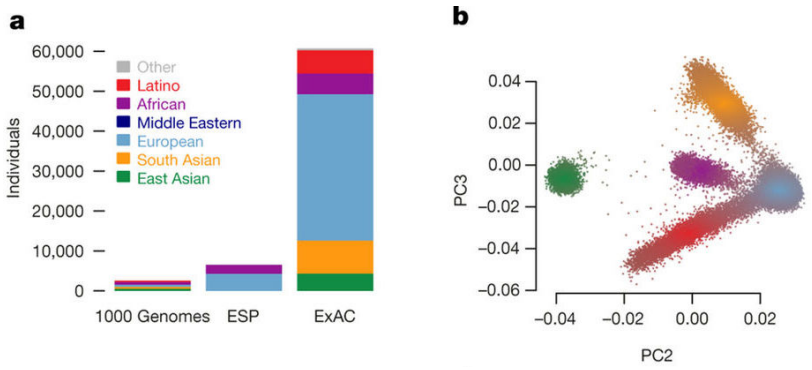

Figure 1: Lek *et al*, *Nature*, 2016; Figure 1

a) What are three benefits decomposition methods provide? *(1) Improved interpretability. (2) Reduced noise in most cases. (3) p is reduced relative to n.*

b) The dataset includes each individual as a row (observation) and each protein-coding variant as a column (variable). Is this a scores or loading plot? *Scores plot.*

c) A coding variant is exclusively present in individuals of east asian descent. Would it be represented in the scores or loading plot? Where would it be? *This would be represented on the loadings plot, since that is the plot that shows input variables. We should expect that the variant would have a negative loading along PC2, and probably little loading along PC3.*

d) Does this plot indicate which group is most different from the others? If so, which one? *No. We don't know how much variation in the dataset is presented here. We definitely know that this plot does not even show the axis along which the majority of variation occurs, since PC1 is absent.*

e) Would the location of each group change if there were 100X fewer individuals of east asian descent present in the data? Justify your answer. *Yes. Both the scores and loadings matrices are dependent upon all the points in the original matrix, since changing any one point will change the directions of maximal variance. Reducing representation of east asians in this analysis would reduce the contribution of the axis that is PC 2 here.*

f) Your colleague accidentally scaled the variables by standard error instead of standard deviation before running PCA. How would the loadings and scores change? *This would not change the directions of maximal variance, and so the directions of the loadings would not change. The loadings magnitudes would change since the magnitude of variance has changed. The scores matrix would not change.*

Question 6 (15 pts)

Kim *et. al.* use partial least squares regression to interpret the relationship between signaling factors and mammary epithelial cell migration before and after a gene expression program. To do so, they regress signaling measurements against migration speed (Y).

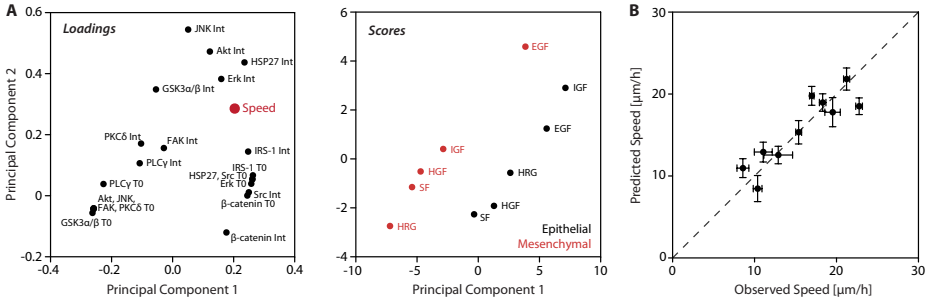

FIG. 4. **A multivariate partial least squares regression model captures signaling metrics contributing most to the prediction of both epithelial and mesenchymal cells.** A PLSR model has been constructed using the initial phosphorylation levels and those integrated over 60 min of the 14 signals described in Fig. 3 across serum-free, EGF, HRG, IGF, and HGF treatments. *A*, Projection of loadings (*left*) and scores (*right*) onto the first two principal components. Loadings of individual signaling metrics (Int = integral of phosphorylation; T0 = initial phosphorylation) are plotted in *black*. Loading of cell speed metric is plotted in *red*. Scores of each growth factor treatment are plotted *black* for epithelial and *red* for mesenchymal cells. *B*, Leave-one-out cross-validation of the PLSR model with cell speeds predicted by the two principal component model *versus* experimentally measured cell speeds.

Figure 2: Kim *et al*, *Mol Cell Prot*, 2011; Figure 4

a) What processing was likely necessary before using the data to build the model? *The dataset needed to be mean centered and unit variance scaled (z-scored).*

b) What effect to you predict an HSP27 inhibitor would have on measured cell speed? *This should decrease "HSP27 Int" and "HSP27 T0".*

An HSP27 inhibitor should decrease cell speed, assuming no other variables change.

c) How do you expect EGF stimulation to influence Erk activation ("Erk Int") as compared to control ("SF")? *EGF stimulation moves positively along both PC1 and PC2 as compared to SF. Because "Erk Int" is positively weighted along both PC1 and PC2, I expect that it increases upon EGF stimulation.*

d) How do the R2Y and Q2Y quantities differ? What can you say about how each quantity varies in general with respect to the number of components? *R2Y evaluates the Y variance explained by the model when directly fit, while Q2Y evaluates it upon cross-validation. R2Y will always increase with more components, while Q2Y may increase or decrease.*

e) You built a PLSR model and prepare the data by z-scoring each column/variable, then wish to crossvalidate the model. Do you need to z-score again for each fold? Why/why not? *You do need to z-score the separately within each fold. This is because z-scoring only before cross-validation "leaks" information about the left out observations. For example, if you leave out an observation that is lower than the average, the average of the training data will be above zero.*