


BE188 Midterm, Winter 2019

Question 1 (15 pts)

You decide to model your confidence in a medical procedure succeeding as a Beta distribution, where $p(x) = C_1 x^\alpha (1-x)^\beta$ for $0 \leq x \leq 1$. That is, x is the probability of a procedure succeeding, and $p(x)$ is the probability you have assigned to that particular x being true. So far, 10 procedures have succeeded, and none has failed, so $\alpha = 10$ and $\beta = 0$.

- Based on the properties of probability distributions, what is the scaling constant C_1 ? What is $p(x < 1/2)$?
- What is the mean of this distribution?
- 
- What are three things (total) you can say about the sampling distributions of the mean for $N=1$ and $N=5$?
- How could you test whether a set of points follow this distribution? (Very briefly describe.)

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.

- What method can you use to quickly determine this conversion factor from your calibration points?
- Your team asks you to provide design a scheme whereby the device provides feedback as to whether new calibration are needed. How could you determine this from the calibration points you have $([(V_1, O_1), (V_2, O_2), (V_3, O_3), \dots])$ so far?
- You have many calibration points (say $N > 50$), so you know that you can ignore variance in the model (i.e. if you ran bootstrapping, your β terms would have negligible variance). What can you say about where you would expect new calibration points to be distributed? How are the β terms distributed?
- A team member insists that the voltage-oxygenation relationship is log-linear instead of linear, and so suggests transforming V to $\log(V)$ instead. When would this be alright? What is an alternative approach? What are the concerns with either approach?
- In version 1 of the device you used a single value as input, calculated from the ratio of 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 could calibrate your model. What assumptions are you making? How could you compare performance of version 2 to version 1?

Question 3 (15 pts)

A PSA test is a diagnostic test for prostate cancer with a sensitivity and specificity of roughly 75% and 50%, respectively. A completely healthy, asymptomatic man shows a positive test and is recommended for a biopsy. The incidence of prostate cancer in the general population is 1 per 1000 men.

- Write out Bayes' law, and rewrite the equation for the probability of the man having a prostate tumor given his positive test.

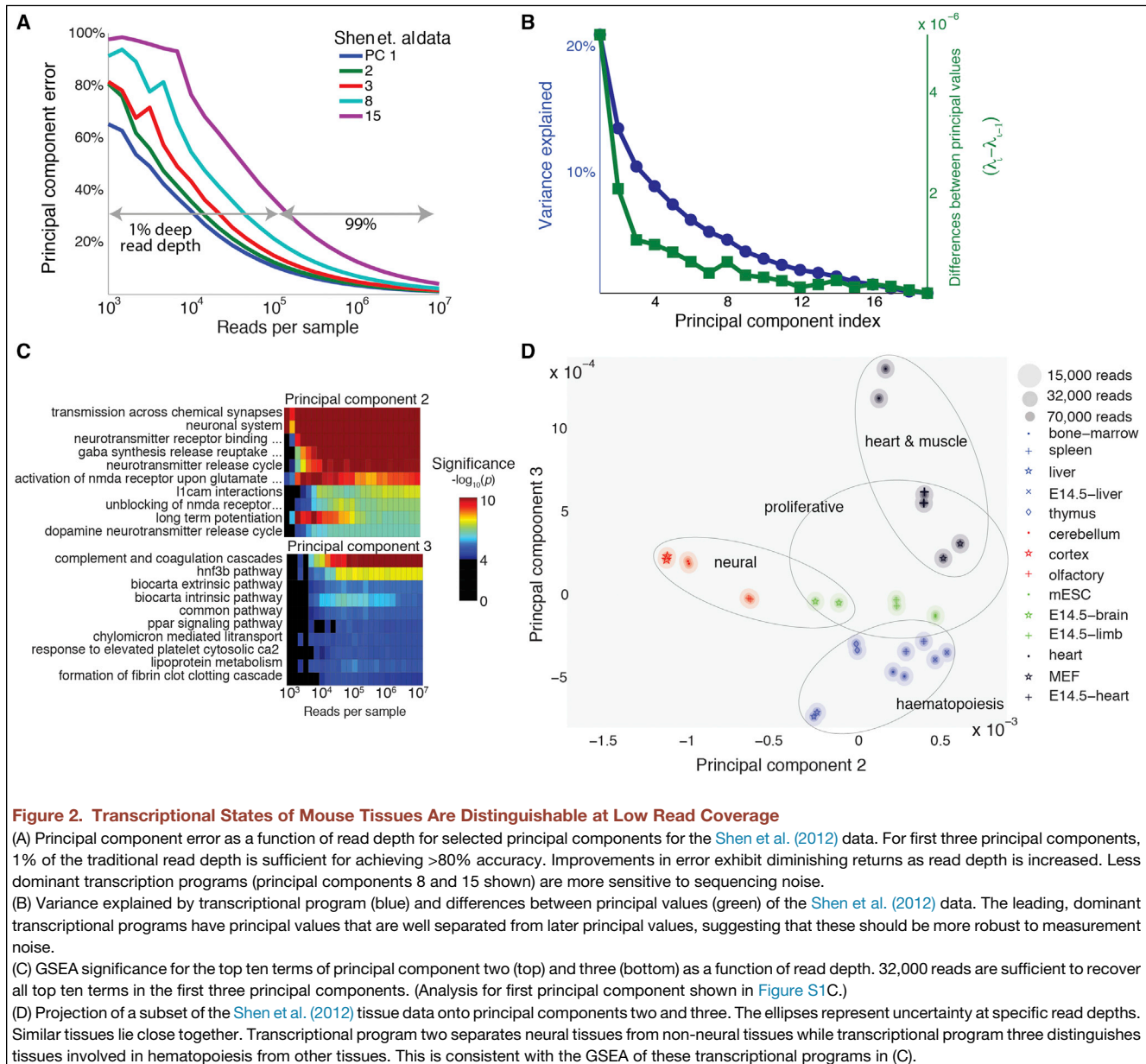
- b) Sensitivity is true positives over all positives, while specificity is true negatives over all negatives. Therefore, the false positive rate is $1 - \text{specificity}$ and the false negative rate is $1 - \text{sensitivity}$. How many false and true positives are expected in a cohort of 1000 tests?
- c) Calculate the probability of the man having prostate cancer, given his positive test result.
- d) What could we do to further ensure positive tests are giving us true results? (You can't improve the test itself.)
- e) A common form of prostate biopsy has a sensitivity and specificity of 50% and 90%, respectively. What the chance the **biopsy** is positive, given the information from above?
-

Question 4 (20 pts)

- a) What is crossvalidation and what does it evaluate?
- b) Outline the steps to performing crossvalidation.
- c) How do predictions from crossvalidation necessarily differ from fitting a full model?
- d) Why are multiple folds necessary?
- e) What does bootstrapping pretend to do with your data?
- f) Outline the steps for performing bootstrapping.

Question 5 (15 pts)

Heimberg *et al*, *Cell Systems*, 2016, proposed that even shallow or noisy sequencing, with very few mRNA reads from cells, can provide a reliable picture of sample variation in principal components space.



- What are three benefits decomposition methods like PCA provide?
- The dataset includes each cell type as a row (observation) and each gene as a column (variable). Is subfigure D a scores or loading plot?
- A gene is exclusively expressed in neural tissues. Would it be represented in the scores or loading plot? What can you say about where it would be?
- Does this plot indicate which group is most different from the others? If so, which one?
- Would the location of the other groups change if we removed the neural tissues from the data set? Justify your answer.
- Your colleague accidentally scaled the variables by standard error instead of standard deviation before running PCA. How would the loadings and scores change?

Question 6 (15 pts)

Kim *et al.* used partial least squares regression to interpret the relationship between signaling factors and mammary epithelial cell migration before and after epithelial-mesenchymal transition. To do so, they regress signaling measurements against migration speed (Y).

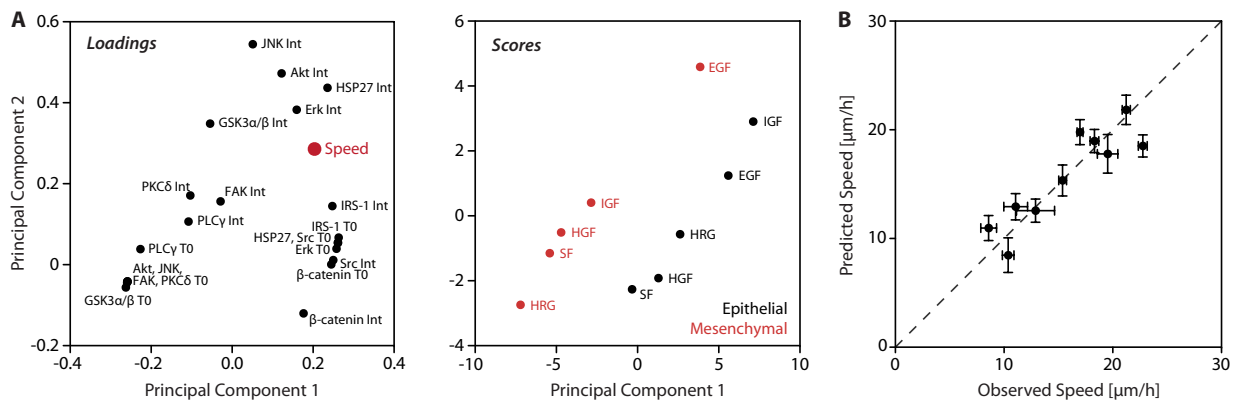


FIG. 4. **A multivariate partial least squares regression model captures signaling metrics contributing most to the prediction of both epithelial and mesenchymal cells.** A PLSR model has been constructed using the initial phosphorylation levels and those integrated over 60 min of the 14 signals described 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.

- What pre-processing was likely necessary before using the data to build the model?
- What effect do you predict an Erk inhibitor would have on measured cell speed?
- How do you expect levels of PKC δ activation to differ in the mesenchymal cells as compared to the epithelial ones?
- 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?
- 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?