BE C175/275 Midterm, Winter 2020

Question 1 (15 pts)

Chao *et al*, *Mol Syst Biol*, 2019 identified that the lengths of individual cell cycle phases are Erlang-distributed. This is a distribution that naturally arises for processes made up of k subprocesses in series. (That is, if $k = 10$, a cell cycle phase is made up of 10 steps for the cell to progress through that phase.) The Erlang distribution is defined by the equation:

$$p(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$$

where λ is the rate of progression through each subprocess.

Note: You can just setup each problem; you do not need to solve the integrals.

a) What is the mean of this distribution when $k = 3$ and $\lambda = 1$?
$$\int_0^\infty x\, p(x)\, dx = 3$$

b) What is the skew of this distribution when $k = 3$ and $\lambda = 1$? The skew of this distribution is always positive—what does this say about how the values are distributed?
$\int_0^\infty x^3 \left(p(x) - \mu\right)\, dx = 2/\sqrt{3}$ Positive skew means values extend further from the mean in the positive direction.

c) You perform an experiment where you watch cells for 48 hrs, and measure how long they take to progress through the cell cycle. Because you only watched for 48 hrs, you've truncated your distribution (made its range $0 \leq x \leq 48$. Renormalize your expression to make it into a truncated distribution. (Hint: You can add a scaling factor, $p_T(x) = C_1 p(x)$, but then need to figure out its value.)
$$\int_0^{48} C_1 p(x)\, dx = 1$$

d) What are three things (total) you can say about the sampling distributions of the mean for N=1 and N=5?
With N=1 it is just the distribution you started with, for N=5 it has lower variance and is more normal.

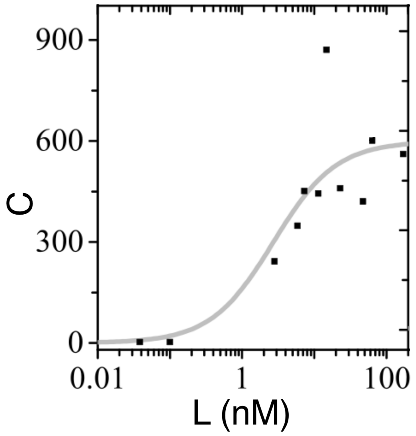
e) You want to test whether your measurements follow the truncated distribution you identified. How could you do this? (Very briefly describe the process.)
KS test.

Question 2 (20 pts)

You are asked to fit a series of binding measurements to a receptor-ligand binding model:

$$C(L) = \frac{R_T L}{K_D + L} + N_B L$$

Where K_D, N_B, and R_T are unknowns, L is the concentration of ligand in solution, and C is the measured amount of binding. N_B indicates the amount of nonspecific binding.



Example binding curve measurements.

a) What method should you use to fit these measurements?
Nonlinear least squares.

b) You colleague asks you if additional measurements of this curve would be helpful, or if these are enough to get an accurate measurement of K_D within a standard deviation of 1 nM. What method could you use to quantify whether this has been accomplished?
Bootstrap.

c) Say you have many fitting points (say N > 50). What can you say about how you would expect new calibration points to be distributed? How about how K_D would be distributed if you were to collect your data again many times?
New calibration points will be normally distributed around the prediction line.
K_D would be normally distributed according to the central limit theorem.

d) Your colleague uses a Scatchard plot to analyze the data in parallel to you. To do so, they plot the data as C versus $C/(R_T - C)$. This provides a linear binding relationship where the slope is K_D^{-1} and y-intercept R_T. What is the benefit of fitting the data this way? What is the problem with doing this?
Transforming your data makes it possible to apply ordinary least squares, which is numerically simpler, but doing this distorts your assumption of normally distributed error.

e) To measure binding, your team has been using the ratio of two wavelengths from spectroscopy data (because your protein's absorption changes with binding). You wonder whether the entirety of the spectroscopy data might be helpful. Therefore, rather you redesign your model as follows, for a ligand where you know the K_D (β is your unknown):

$$X\beta = \frac{L}{K_D + L}$$

To fit your model, you measure absorption at 200 wavelengths for 10 concentrations of ligand binding (X). Describe how you could calibrate your model. Justify your choices.
Must apply regularization or principal components regression, since you have more variables than data. Any form of regularization is acceptable.

Question 3 (15 pts)

A newly identified coronavirus has been spreading throughout the Wuhan region in China, and a few cases have been confirmed within the U.S. along with other countries. You are part of a rapid response team developing a blood-based assay for the virus. The goal is to deploy this assay within airports in the U.S. to identify individuals who are infected.

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

$$p(+\,|\,\text{pos test}) = \frac{p(+)\,p(\text{pos test}\,|\,+)}{p(\text{pos test})}$$

b) You estimate that your test has a 95% sensitivity and 90% specificity. The false positive rate is $1 -$ specificity and the false negative rate is $1 -$ sensitivity. Roughly 8,000 people enter the U.S. from China each day. Assuming one of those individuals is sick, what are the number of false and true positives you will have each day?

$TP = p(\text{pos test} \mid +)p(+) = 0.00012 \times 8000 = 0.95$

$FP = p(\text{pos test} \mid -)p(-) = 0.1 \times 7999 = 799.9$

c) Calculate the probability of a passenger having the virus, given a positive test result.

$$\frac{0.95 \times 1/8000}{0.95 \times 1/8000 \ + \ 0.1 \times 7999/8000} = \frac{0.95}{0.95 + 799.9} = 0.0012$$

d) What could we do to further ensure positive tests are giving us true results? (You can't improve the test itself.)

Only test people who have other sensitive symptoms, like fever.

e) A followup PCR test has a sensitivity and specificity of 50% and 90%, respectively. What is the chance that the **PCR test will be positive**, given someone is tested only after a positive result with your assay?

$p(\text{PCR}) = p(\text{PCR} \mid \text{virus})p(\text{virus}) + p(\text{PCR} \mid \sim \text{virus})p(\sim \text{virus})$

$p(\text{PCR}) = 0.5 \times 0.0012 + 0.1 \times (1 - 0.0012) = 0.10$

Question 4 (20 pts)

a) What is cross-validation and what does it evaluate?

Evaluates the prediction error by leaving out a portion of the dataset for validation.

b) Outline the steps to performing cross-validation.

Split, fit, quantify error. Repeat over folds then average.

c) How do predictions from cross-validation necessarily differ from fitting a full model?

Higher variance.

d) Why are multiple folds necessary?

Average out the effect of the left-out group.

e) What does bootstrapping pretend to do with your data?

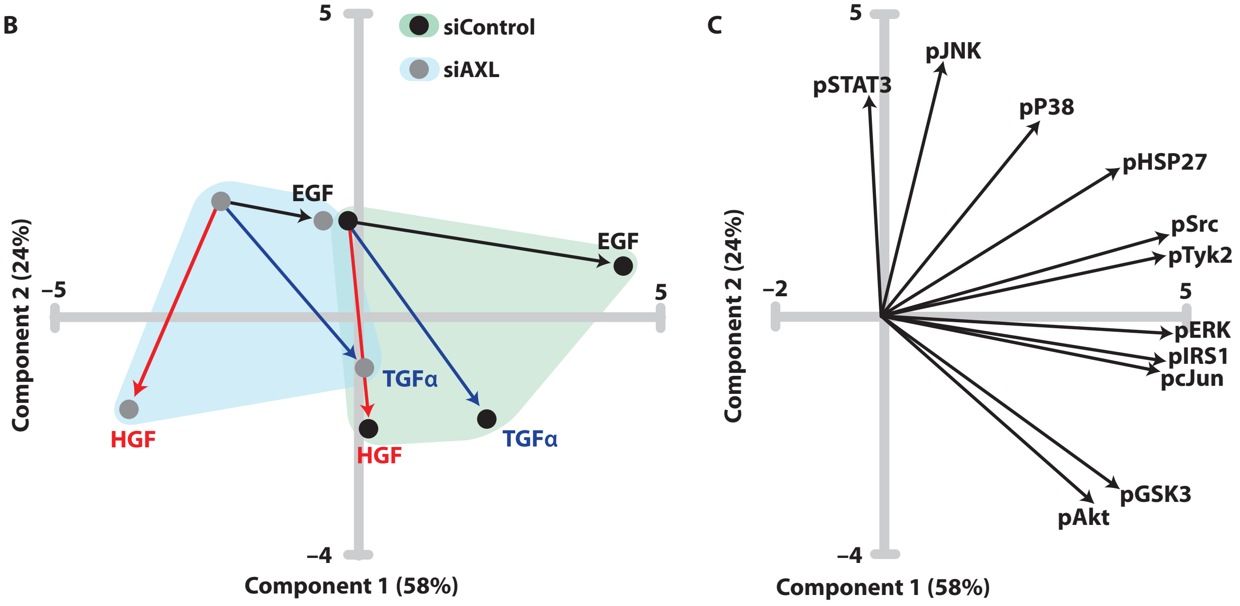
Pretends to make entirely new datasets of the same size.

f) Outline the steps for performing bootstrapping.

Resample with replacement, then fit. Repeat.

Question 5 (15 pts)

Meyer *et al* used PCA to evaluate how the AXL receptor alters signaling in response to other RTK ligands like EGF, TGFα and HGF. To do so, they measured a panel of phosphorylation sites in cells with or without AXL knocked down by siRNA, in response to stimulation with each ligand.



a) What are three benefits decomposition methods like PCA provide?
Visualization, compression, noise reduction.

b) What are two things you can always say about PC2 in relation to PC1?
PC2 always explains a smaller percentage of the data variance, and is orthogonal to PC1.

c) The dataset includes each phosphorylation site (e.g., pJNK) as a column and each treatment (e.g., EGF) as a row/observation. Does the first plot (B) show a scores or loadings plot?
Scores plot.

d) You measure a new phosphorylation site that is only induced by HGF stimulation, and is not affected by siAXL treatment. Where would you expect it to be on the loadings plot?
Negative along PC2. Probably not much weighting on PC1.

e) Would adding this new phosphorylation site affect the position of the other points? Explain.
Yes. PCA captures the variation in the dataset, so changing any of the data will (at least subtly) change all of the decomposition.

f) Your colleague accidentally scaled each variable to twice the standard deviation, rather than the standard deviation. How would the scores and loadings change?
Either the scores or loadings will be scaled to twice the magnitude, but the relative position of the points will not 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).

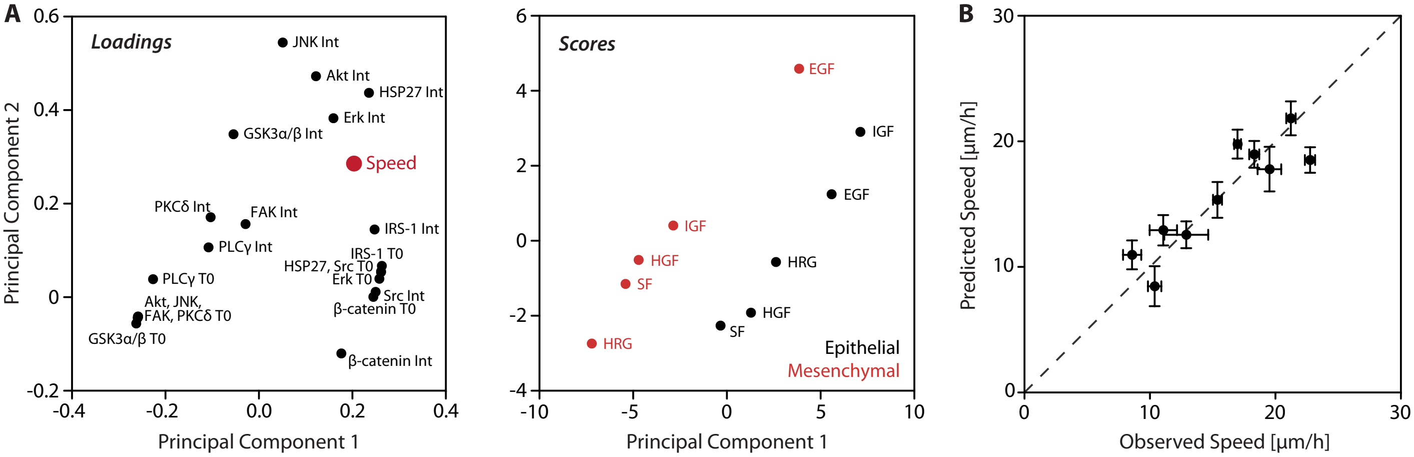


FIG. 4. **A multivariate partial least squares regression model captures signaling metrics contributing most to the prediction of both epithelial and mesenchymal cells.** A PLSR model has been constructed using the initial phosphorylation levels and those integrated over 60 min of the 14 signals described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

a) How could you determine whether *epithelial IGF*'s score is significantly positive on PC1, or positive within the variance of the model? (Hint: How would this point's score change if we collected a new dataset?)
Bootstrapping.

b) How would the model predict a JNK inhibitor would affect cell speed (use *JNK Int*, ignore *JNK T0*)?
Reduce cell speed.

c) How would you expect β-catenin phosphorylation to differ between epithelial and mesenchymal cells?
Higher in epithelial cells.

d) What do R2Y and Q2Y refer to? What can you say about how each varies with respect to the number of components?
The % variance captured by the model upon fitting (R2Y) or cross-validation (Q2Y). R2Y will increase and asymptotically approach 1 with more components. Q2Y will likely increase and then decrease after a point.

e) You built a PLSR model and prepare the data by z-scoring each column/variable, then wish to cross-validate the model. Do you need to z-score again for each fold? Why/why not?
Must z-score within each fold to get rid of the effect of the left-out data.