BE 275 Midterm, Fall 2021

Note that this is from the graduate version of the class. The questions are meant to be more challenging.

Question 1 (10 points)

Assuming a constant and random fixed rate, the number of neuron firings for a given neuron within a set period of time follows a Poisson distribution:

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

where k is the number of firings, and λ is the average number of firings in that given period. You collect recordings of firing numbers across a population of neurons using a fluorescent reporter and a microscope. Each neuron has the same firing rate. Each microscopy field has roughly 50 neurons with the reporter, and you image 20 different regions.

a) Provide an expression for the variance in the number of events across neurons (using the distribution above). You do not need to simplify or solve.

$$\sigma = \int_0^\infty (k-\mu)^2 p(k) dk$$

$$\sigma = \int_0^\infty (k-\lambda)^2 \frac{\lambda^k e^{-\lambda}}{k!} dk$$

$$\sigma = e^{-\lambda} \sum_{k=0}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!}$$

This happens to simplify to just λ.

b) What are at least three things you can say about the distribution of average firing numbers across the fields of view in your experiment (relative to the distribution for single neurons)?
(1) The average will tend toward the same average as across neurons. (2) It will have less variance. (3) It will be more like a normal distribution (central limit theorem). (Other answers possibly ok.)

c) You expect that there is actually a population of neurons not firing at all. Meanwhile, the active neurons fire at a rate of once per second. What is the probability of a neuron having no firing events, given it is active? How about if it is inactive/non-firing?

$$p(k = 0 \mid \lambda = 0) = 1$$

$$p(k = 0 \mid \lambda = rt = 1t) = e^{-t}$$

d) You expect that 5% of your neuron population is inactive/non-firing. For how many seconds do you need to collect data to be 99% sure that a neuron with no events is inactive?

$$p(I \mid k = 0) = \frac{p(k = 0 \mid I)p(I)}{p(k = 0)} = \frac{p(k = 0 \mid I)p(I)}{p(k = 0 \mid I)p(I) + p(k = 0 \mid A)p(A)}$$

$$\frac{0.05}{0.05 + 0.95 \cdot p(k = 0 \mid A)} = \frac{1}{1 + 19 \cdot e^{-t}}$$

$$0.99 = \frac{1}{1 + 19 \cdot e^{-t}}$$

t is approximately 7.5 seconds

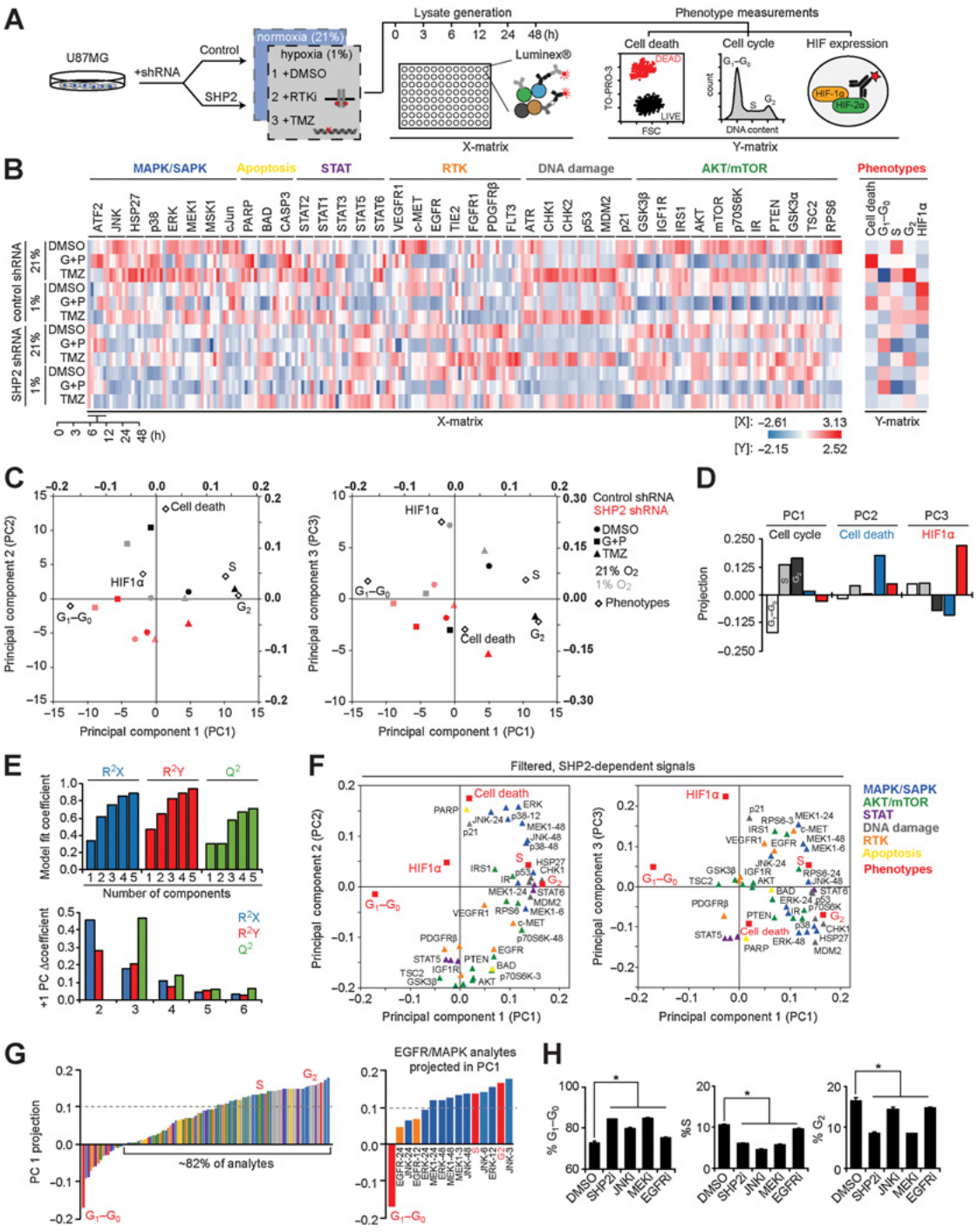


Figure 2.

A data-driven model predicts SHP2-regulated signaling governing GBM cell response to therapy. **A,** U87MG cells expressing control or SHP2 shRNA were cultured in 21% or 1% oxygen and treated with gefitinib (G) and PHA665752 (P), temozolomide (TMZ), or DMSO. Cells were lysed 0, 3, 6, 12, 24, and 48 hours after treatment, and lysates were analyzed via Luminex. **B,** Mean-centered, variance-scaled signaling and phenotypic measurements are represented by a heat map, with Luminex kits indicated. Supplementary Table S2 provides analyte posttranslational modifications. **C,** Bi-plots of scores (bold numbered axes; top right) for conditions and loadings (bottom left) of Y-matrix phenotypes for PC2 or PC3 versus PC1. **D,** Phenotype projections into PCs 1–3 are plotted. **E,** Model fit (R^2X, R^2Y) and predictive (Q^2) coefficients are plotted versus PC. Coefficient changes (Δ) with additional components are shown. **F,** Loadings are plotted of signals (triangles, color coded by Luminex kit) and phenotypes (red squares) for the three-component model, displaying only SHP2-regulated analytes. Times are indicated where two or more time points projected in opposite directions for an analyte. Loadings are listed in Supplementary Table S5. **G,** PC1 signal and phenotype projections are plotted, smallest to largest. Strongly projecting EGFR (orange) and MAPK (blue) analytes, and cell-cycle phases (red) are highlighted. **H,** U87MG cells were treated for 48 hours with 10 μmol/L SHP099 (SHP2i), 20 μmol/L SP600125 (JNKi), 5 μmol/L CI-1040 (MEKi), 10 μmol/L gefitinib (EGFRi), or DMSO, and cell-cycle distribution was analyzed. Error bars indicate mean ± SEM of three replicates; *, $P < 0.05$ for the indicated comparisons from Tukey *post hoc* testing following one-way ANOVA.

Question 2

Day *et al*, *Canc Res*, 2021 build a model for the signaling determinants of glioblastoma response to SHP2 inhibition. To do so, they measure a variety of signaling proteins over time. Their results are shown in Figure 2 on the next page.

a) You decide to analyze the molecular measurements (e.g., ATF2, JNK) by PCA as well. Would you expect the scores and loadings to be the same, similar, or completely different? Explain.
The scores and loadings will almost certainly be different. The degree of difference will come down to how different the variation in X is compared to the covariation between X and y.

b) Describe the effect of PC2 on cell phenotype (e.g., Cell death, S, G_2). How is PC2 related to SHP2 siRNA treatment?
PC2 strongly corresponds to an increase in cell death. It also corresponds to a small increase in HIF1α and S phase. SHP2 shRNA treatment corresponds to a decrease in PC2.

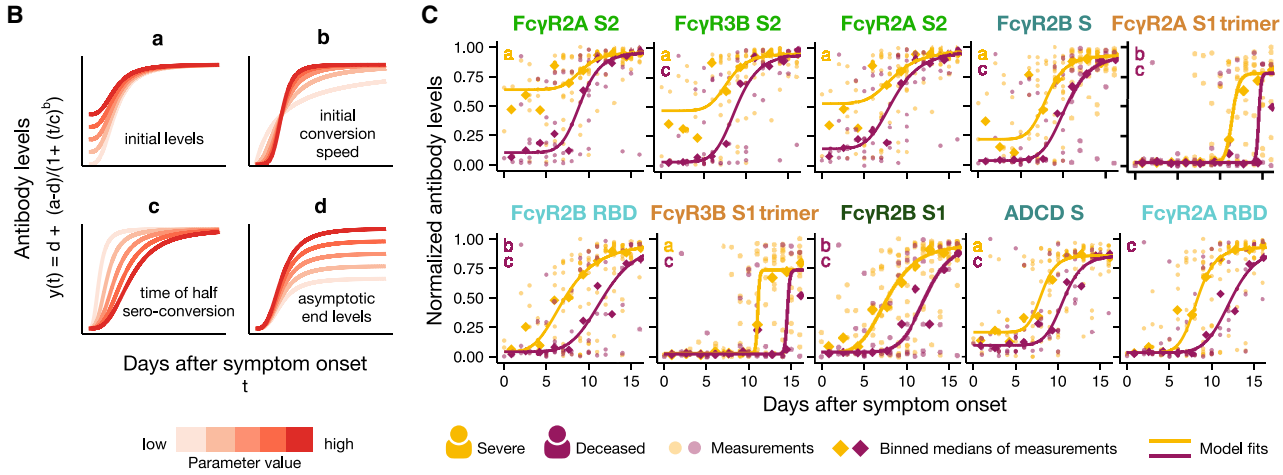
c) Let's say the authors decide to use a 4-component model instead of a 3-component one. How would the plots in (C) and (F) change?
They would not change. The preceding components of a PLSR model are not dependent on the subsequent ones.

d) How do R2X, R2Y, and Q2Y vary with the number of components? How would R2X look relative to R2X using PCA?
R2X and R2Y monotonically increase with the number of components. Q2Y likely increases and then eventually decreases (bias-variance tradeoff). R2X with PCA will always be higher than that for PLSR at a given number of components.

e) You reproduce Figure 2C, but the points look mirrored along the y-axis compared to the figure (all the negative PC1 points are at positive PC1 positions). What has happened? Are your results the same as theirs? Would any other plots also change?
PLSR models are sign-indeterminant. The PLSR results are still the same. The PC1 scores would also have to flip, so (F) should be mirrored along the y-axis.

Question 3

Zohar *et al*, *Cell*, 2020 analyzes the dynamics of antibody response during SARS-CoV-2 infection. Each of the plots below depicts a certain measure of antibody quality/quantity, separated by whether subjects survived after being admitted to the ICU. They summarize these dynamics by fitting the longitudinal data to a logistic curve model using non-linear least squares.



a) What are two algorithmic challenges (difficulties in arriving at a solution) that arise when using non-linear least squares but do not in the linear case (OLS)?
One has to use an iterative fitting scheme to arrive at an answer with non-linear least squares, which is much less efficient. You are also not guaranteed to find the globally optimal solution.

b) What do you need to provide to a non-linear least squares method to fit the model?
A function describing the fitting curve and the starting point for fitting.

c) You reimplement this analysis, and want to independently check that the fitting process successfully fit. What is something you could check to verify the solution converged?
You could check that the gradient for the sum of squared error is zero.

d) You want to test whether the b parameter is significantly different between severe and deceased outcomes. How can you do this?
You could bootstrap the curve fits for both outcomes and then quantify the fraction of times the b parameter crosses to derive a p-value.

e) You wish to cross-validate your model to verify that it is not over-fit to the data. In the study, each subject was measured at 2–3 timepoints. How should you decide which samples to leave out in each fold? Explain your reasoning.
The samples left out in each fold should be independent of those included for training. Therefore, I would cross-validate across subjects, not samples.

Question 4

A mammogram is a diagnostic test for breast cancer with a sensitivity and specificity of roughly 97% and 64.5%, respectively. A completely healthy, asymptomatic woman shows a positive test and is recommended for a biopsy. The incidence of breast cancer in the general population is 1 per 8 women.

a) Write out Bayes' law, and then rewrite it as the probability of the woman having a breast tumor given her positive test.

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

$$p(T \mid P) = \frac{p(P \mid T)\, p(T)}{p(P)}$$

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?

$$FP = (7/8)(1 - 0.645)(1000) = 310.6$$

$$TP = (1/8)(0.97)(1000) = 121.3$$

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

$$p(T \mid P) = \frac{p(P \mid T)\, p(T)}{p(P)} = \frac{p(P \mid T)\, p(T)}{p(P \mid T)\, p(T) + p(P \mid \sim T)\, p(\sim T)}$$

$$\frac{(0.97)(1/8)}{(0.97)(1/8) + (1 - 0.645)(7/8)} = 0.28$$

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

We could run a second independent test, or select for women who are more likely to have breast cancer before they receive a mammogram.

e) A common form of breast biopsy has a sensitivity and specificity of 91% and 98%, respectively. What the chance **the biopsy is positive**, given the information from above?

$$p(P) = p(P \mid T)\, p(T) + p(P \mid \sim T)\, p(\sim T)$$

$$p(P) = (0.91)(0.28) + (0.02)(0.72) = 0.27$$

Question 5

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

Cross-validation is the process of iteratively leaving out one piece of a dataset, training on the remaining data, and then quantifying your ability to predict the left-out portion. By looping over all possible left-out data, you can estimate the ability of your model to predict new data.

b) Outline the steps for performing cross-validation in order, including normalizing your data by z-scoring.

(1) Leave out the test data. (2) Normalize the training data by z-scoring. (3) Fit the model to the

training data. (4) Predict the test data (correcting for the effect of z-scoring). (5) Loop over other folds.

c) How do predictions from cross-validation systematically differ from the actual model?
The cross-validated model is necessarily built on less data than the full model, and so it is an over-estimate of the prediction error.

d) Why are multiple folds necessary? What are two trade-offs when selecting a number of folds?
If one only leaves out a single fold, the prediction error estimate is dependent on the properties of those samples left out. This can lead to wildly different estimates. We can get a better estimate by averaging over folds. Two tradeoffs are: (1) Compute time. A large number of folds takes longer to calculate. (2) Smaller number of folds will increase the effect in part (c).

e) What does bootstrapping provide?
An estimate of the variance in our model, if we were to go and collect entirely new datasets on which to train it.

f) Outline the steps for bootstrapping a PLSR model. In the end, what are the resulting outputs of doing this?
(1) Resample the dataset (with replacement). (2) Build the model. (3) Repeat a certain number of times. The output of this will be a family of models (number equal to the number of bootstrap samples.)

Question 6

You are developing a glucose monitor that automatically dispenses insulin. You develop a *very* simple model of insulin and glucose regulation in the body:

$$\frac{\delta I}{\delta t} = -\gamma I + \theta G$$
$$\frac{\delta G}{\delta t} = \alpha - \beta I G$$

I and G indicate the amount of insulin and glucose, respectively. α indicates some constitutive production of glucose, θ some production of insulin by the body when glucose is present, β the removal of glucose by insulin, and γ clearance of insulin.

a) Is there a steady-state amount of insulin and glucose? If so, how much?
Yes. Set equations equal to 0 and solve (don't need to simply).

b) Can this model ever oscillate? If so, under what parameter values?
$$J = \begin{pmatrix} -\gamma & \theta \\ -\beta G & -\beta I \end{pmatrix}$$
$$\Delta = \gamma\beta I + \beta\theta G$$
$$\tau = -\gamma - \beta I$$
If $\tau^2 - 4\Delta < 0$, there will be imaginary or complex eigenvalues and therefore an oscillatory component to the model.

c) Sketch out what the function would look like that you would pass to an ODE solver. Describe the inputs and outputs.
The function needs to take in the current point and time as an argument, and then pass back the derivatives at that point.

d) Say you obtain the solution on the right, plotted in phase space. What does this say about solutions that begin at the red dot? Explain your reasoning.
Solutions beginning at the red dot will not be able to cross the line. This arises from the uniqueness property of continuously differentiable ODE models. If we were to cross the line, at some point we would have to be on the line, which would mean we would have to progress with the solution indicated by the line.