BE 175 Midterm, Winter 2023

Question 1 (20 pts)

These questions are meant to be answered with short answers (less than three sentences should be plenty).

a) What is cross-validation and what does it estimate? Why are multiple folds necessary?
Cross-validation is a strategy for estimating the prediction error by creating a validation set from your original dataset. Multiple folds are necessary because, with just one fold, the specific data points held out would greatly influence the results.

b) What is the risk of applying your model in patient groups very different from your cross-validation dataset?
Cross-validation estimates your ability to predict data only if your dataset is a representative sample. If this is not the case, then the model could have much worse prediction ability than cross-validation estimates.

c) What is bootstrapping and what does it estimate? Where do the bootstrapped datasets come from?
Bootstrapping estimates the variance in your model. Bootstrapped datasets come from the original dataset itself, but the observations are resampled with replacement to create a "new" dataset of the same size.

d) During hypothesis testing, how can you determine the false negative and false positive rates? Be as specific as possible.
The false positive rate is equal to the p-value cutoff used, α. For the false negative rate, you have to perform a power analysis wherein you assume the null hypothesis is false, and then run repeated simulations to see how often you miss this fact.

e) What does regularization do to the degrees of freedom of a model, the fitting error, and the prediction error? Be sure to distinguish effects that always occur from those that sometimes occur.
Regularization always reduces the degrees of freedom of the model and increases the fitting error. It may decrease the prediction error, but that is not guaranteed.

Question 2 (20 pts)

Generally, prostate cancer screening is recommended for men (and those assigned male at birth) aged 45 and older. One form of screening is a PSA test. You are a 45-year-old man with a typical prostate cancer risk. You are screened, and your PSA level is 3.1 ng/mL. A PSA cutoff of >3.0 ng/mL has a sensitivity and specificity of 32% and 85%, respectively, for the detection of any prostate cancer. At 45 years old, the prostate cancer incidence rate is ~40 per 100,000.

a) Write out Bayes' law and then rewrite the equation for you having prostate cancer given the positive PSA test result.

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

$$p(\text{cancer} \mid \text{pos}) = \frac{p(\text{pos} \mid \text{cancer})\, p(\text{cancer})}{p(\text{pos})}$$

b) What is the chance you have prostate cancer, given your positive test?

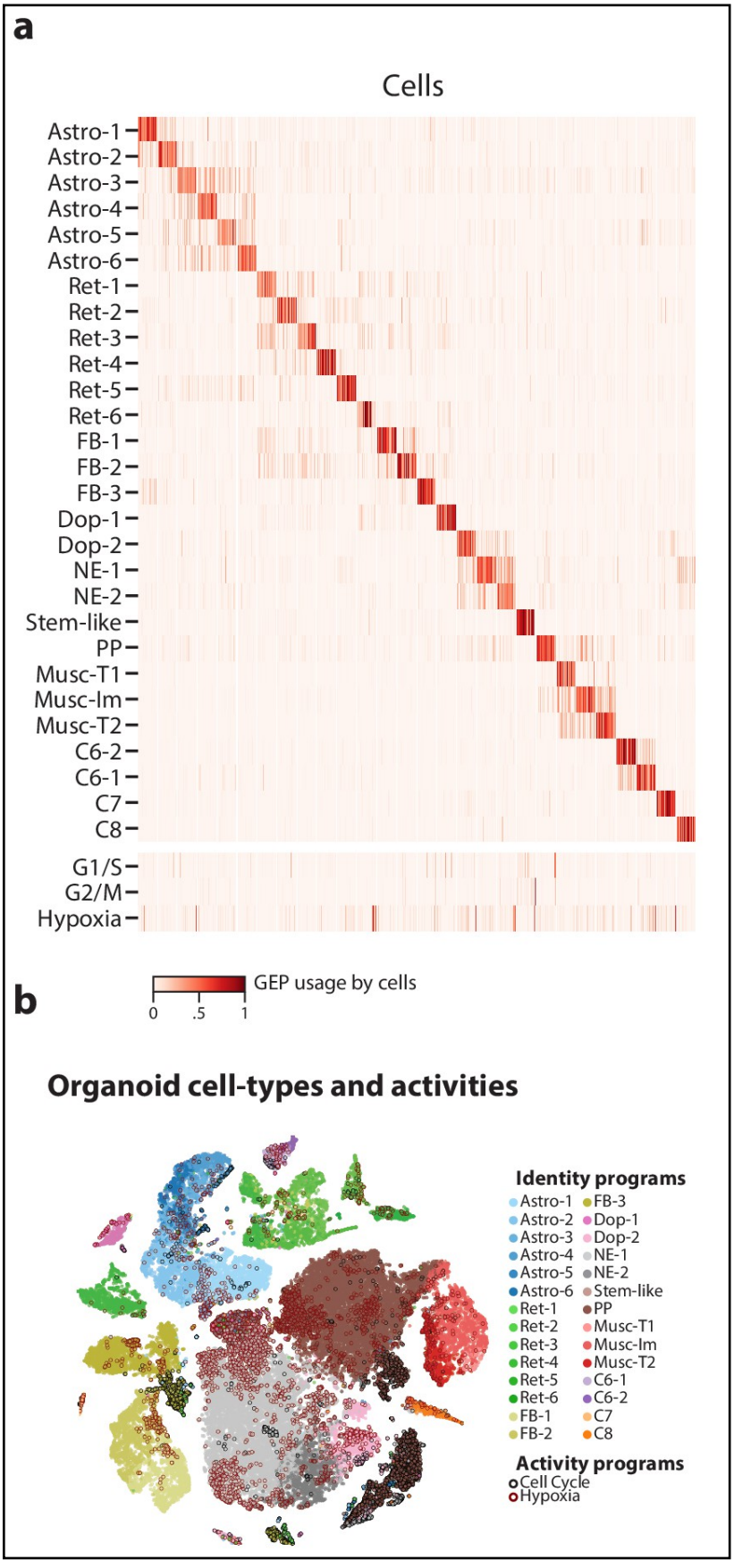
$$p(\text{cancer} \mid \text{pos}) = \frac{0.32 \times 0.0004}{0.32 \times 0.0004 + 0.15 \times 0.9996} = \frac{0.000128}{0.000128 + 0.14994} = 0.085\,\%$$

c) As a follow-up, you have a biopsy taken. A biopsy has a sensitivity and specificity of 50% and 95%, respectively. What is the probability that ***the biopsy result is negative***?

$$p(b\_) = p(b\_ \mid \text{cancer})\, p(\text{cancer}) + p(b\_ \mid \text{not cancer})\, p(\text{not cancer})$$
$$p(b\_) = 0.5 \times 0.00085 + 0.95 \times 0.99915 = 0.9496$$

d) The incidence of prostate cancer rapidly drops off in younger individuals below age 45. PSA levels are routinely checked in blood panels but are not acted upon until age 45. Why is age useful to consider for this test?
Older individuals are much more likely to have cancer independent of the test result. This increases their prior likelihood of having cancer, and consequently the chance of them having cancer *given* a positive test result.

Question 3 (20 pts)



Kotliar *et al*, *eLife*, 2019 propose non-negative matrix factorization as a method to identify cells from single-cell RNAseq that share gene expression programs and cell identities. One of the paper's figures is partly reproduced here.

a) What does NNMF maximize? Under what constraints? Describe the benefit of using NNMF versus other matrix factorization schemes like PCA.

NNMF maximizes the variance explained with the added constraint that all values in the factorization must be positive. As a consequence, its factors are sparse, aiding interpretation.

b) Your colleague comes to you and says they are getting different results each time they fit with NNMF. Is something wrong? If not, what is going on here? Is there a step in the fitting process they could modify to get reproducible results?

No, this is expected. NNMF only performs a local optimization, so results will be different with different starting points. If they wanted to get reproducible results, they could set the random seed, or provide a starting point to the factorization.

c) The starting dataset in the figure is made up of a matrix of 5000 genes by 500 cells. One of the resulting factorization matrices is plotted in (a), and is made up of 31 components by 500 cells. What is the size and composition of the other matrix? Given this, how do you think they know component 1 is a type of astrocytes (a neuronal cell type)?

The other matrix is comprised of 31 components by 5000 genes. They likely looked at component 1 in the other matrix, and observed that it had heavier weights for several markers specific to astrocytes.

d) How is the data normalized before using NNMF?

You typically want to variance normalize the different variables, but you do not want to mean center them because all of the values need to be positive.

Question 4 (20 pts)

Carroll et al, Cancer Research, 2018 examined how cytokines released by activated macrophages (AAMs) in the peritoneum contribute to adhesion of ovarian cancer cells and thus metastasis from ascites fluid. To do this, they measured the abundance of cytokines with and without AAMs using several different cell lines. They then built a model predicting the adhesion of these cell lines in the same conditions.

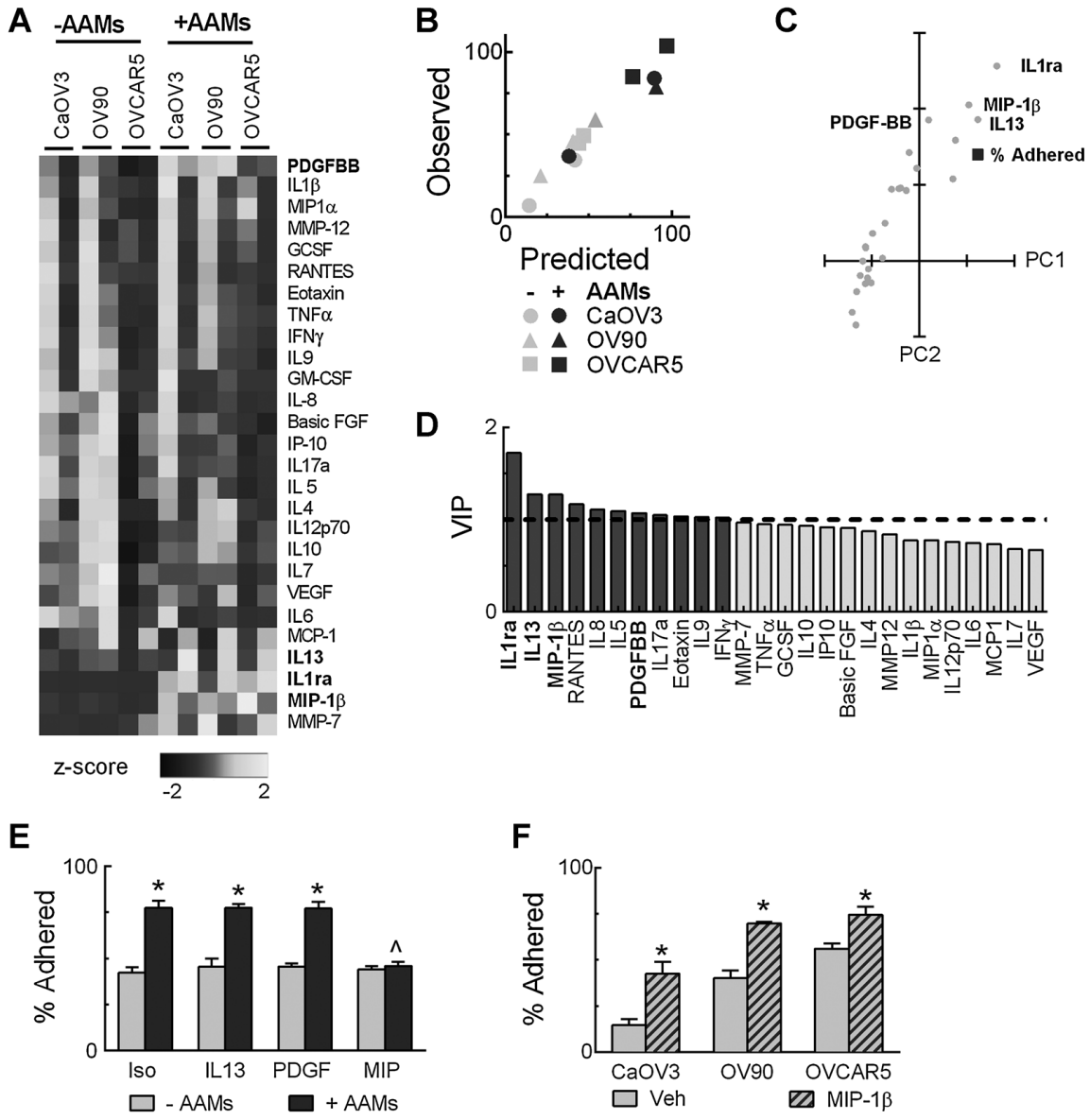


Figure 2.

PLSR prediction and experimental validation of role for MIP-1β in increased HGSOC adhesion. **A,** Ligands (*z*-score normalized) detected in the absence or presence of AAMs. Data are average of $n = 3$ replicates per donor; each column represents a unique donor/cell line combination. **B,** Comparison of PLSR-predicted to experimentally observed HGSOC adhesion to LP-9. **C,** Correlations of ligands and observed adhesion (% Adhered) with PC1 and PC2 from the PLSR model. **D,** VIP > 1 (dark gray) indicate important variables to predict adhesion. Those that positively correlated with HGSOC adhesion are shown in bold [and bolded in the heatmap (**A**) and labeled in **C**]. **E,** OV90 cocultures were treated with neutralizing antibodies against IL13, PDGF-BB (PDGF), MIP-1β (MIP), or isotype (Iso) during coculture; $n = 3$ replicates, one AAM donor. **F,** HGSOC adhesion to LP-9 treated with vehicle or 100 ng/mL MIP-1β, $n = 3$. Data is average ± SD; *, $P < 0.05$ vs.–AAMs of same isotype/antibody (**E**) or vehicle (**F**); ^, $P < 0.05$ vs. +AAMs/isotype (**E**) by two-sided *t* test (**F**), with Bonferroni correction (**E**).

a) What properties of PLSR make it an especially useful model for biological data?
Variables in biological data tend to have a high amount of inter-correlation, and so it is useful to think about their changes in terms of patterns (principal components). PLSR additionally allows us to hone in on those patterns related to an outcome prediction of interest.

b) Carroll *et al* does not label Figure 2C quite right in the figure caption—they call the plotted information the "correlations." Based on the context information, that they are predicting the percentage of cells adhered and that they are measuring variation in the abundance of each cytokine, what is this plot showing from the PLSR model?
These are the X loadings (light circles) and Y loadings (dark square).

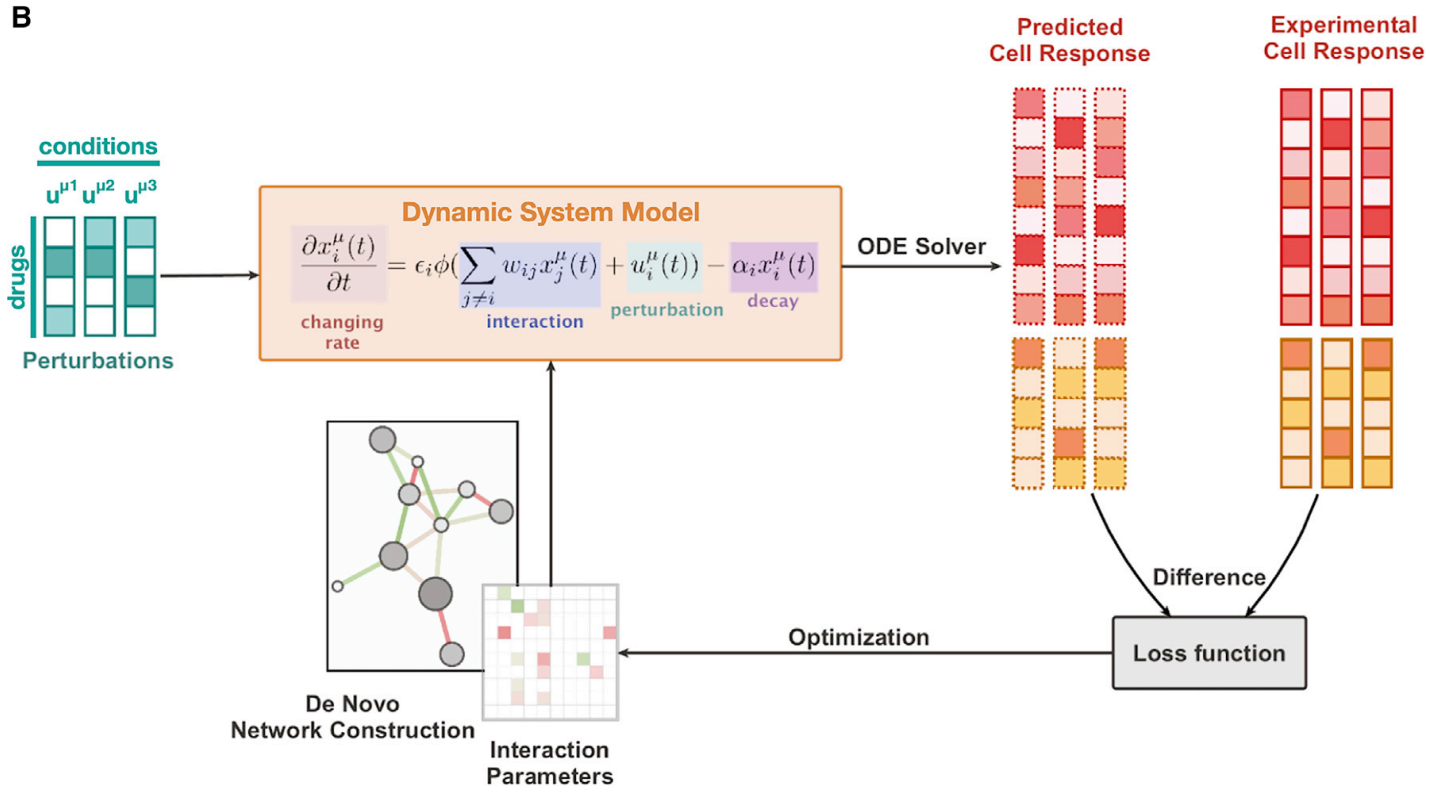
c) In Figure 2D, the authors use the VIP (variable importance of projection) scores to determine which variables are most important to predicting the outcome. Briefly, this score aims to summarize the influence of each PC on the output. Scores over 1 are typically taken to be significant, and the authors follow this advice. You want to calculate which scores would be consistently over 1 if the authors were to collect a series of entirely new datasets. How might they go about that? Describe the steps of this process for this dataset in detail.
Bootstrapping. You would repeatedly fit the model and calculate the VIP scores after resampling the dataset with replacement. This would provide you with a distribution of models, from which you can calculate what fraction of the time a certain VIP score is greater than 1.

d) The authors go on to validate the importance of three different cytokines by using a neutralizing antibody against them, in essence setting the concentration to 0, and measuring the amount of adhesion. The results of this are shown in Figure 2E. Which results of testing these antibodies fit with the inferences of the model? Explain. ("Iso" is a negative control, MIP is MIP-1β.)
The model infers that all three of these cytokines are positively associated with adhesion. Adhesion is only blocked with the MIP blocking antibody, so that result is the only one of the three that is consistent with the model inference.

e) How are PCs necessarily related to one another with PCA? Do you see this relationship in Figure 2C? Based on this observation, what can you say about this relationship with PLSR? How would you expect Figure 2C would change if PCA were performed instead of PLSR?
In PCA the components are orthogonal to one another. That is not observed in Figure 2C—there is correlation between the X loadings components. PLSR does not have components that are orthogonal. Figure 2C would change quite a lot for PCA, because at very least the components would be orthogonal and there would be no Y loading because there would be no Y.

Question 5 (20 pts)

Yuan *et al*, *Cell Systems*, 2020 developed a computational model, CellBox, in which various perturbations (such as gene knockouts) are simulated by a mechanistic dynamical model. The model includes weights for every gene-gene interaction as its unknowns ($w_{i,j}$), which it uses to predict the effect of each perturbation. Predictions are made by inputting a static vector u that describes which genes have their expression inhibited or promoted. The model then simulates the change in expression of every gene until it reaches steady state, and then these steady states are compared to the experimental measurements, *minimizing the sum of squared error of the difference between the model and data*. The weights are then fit to cause the model to match the observed molecular changes. Through this process, the model is able to use perturbations to infer gene-gene interaction effects.



a) What type of statistical model that we discussed could be used to perform the fitting? Are there any caveats/ concerns with this model? What information must be provided besides the model and data to start the fitting process?
Non-linear least squares. One concern with this model is that the fitting process will improve the result, but there is no guarantee that you have reached a global optimum. Besides the model and data, you must provide a starting point for the fitting process.

b) Yuan *et al* includes roughly 200 perturbations (observations) and 100 genes in their fitting data. This means that their model has 10,000 gene-gene interaction coefficients (unknowns). Part of the goal of the model is to identify a small subset of these interaction coefficients that are important. Is there anything they need to do to augment their model to enable fitting? If so, describe how you would tune this additional component.
You will need to augment the model with regularization. Given the goal of having sparsity, l1 regularization would be best. This should likely have its strength tuned by cross-validating the model with different regularization strengths.

c) One potential purpose of the model could be to predict what would happen if you were to knockdown a gene. How could you check that the model can perform this task? Describe the steps to do so.
Because you are wanting to benchmark the model's ability to predict, you should cross-validate!

d) Another question the authors ask of the model is which interactions are present to a statistically significant degree given the data. (In other words, whether the interactions would consistently be found again if the dataset was collected anew.) How can you check this?
Bootstrapping.

e) You perform all the above steps, and are mostly satisfied with your model. However, when simulating the model, you find that it is always rapidly oscillating which does not seem biologically plausible. How could you adjust the fitting to prevent oscillatory behavior? Describe in detail. Can you/should you combine this with what you did in step (b)?
You could regularize the model to penalize against oscillatory behavior. Oscillatory behavior could be identified by examining the Jacobian of the ODE system at the steady state solution—oscillatory behavior would be indicated by imaginary eigenvalues within the Jacobian. As regularization, this could be combined with the regularization in step #2, and likely the strength of both should be tuned to optimize predictions. The regularization here will affect the strength of regularization needed in step (b) because both forms of regularization decrease model variance.