BE 275 Midterm, Fall 2022

Question 1 (10 points)

a) What is cross-validation? Describe the process of performing it. What does it estimate, what does it mean for cross-validation to be an estimate, and how is cross-validation systematically biased?
Cross-validation is a strategy for estimating the prediction error by creating a validation set from your original dataset. During CV, you: (1) split your data into a training and test set; (2) fit your model, including any normalization; (3) quantify the error of the model with respect to the test dataset; (4) repeat the process with a different split. This is an estimate because in practice new predictions will not come from the same dataset—you are pretending you have an independent dataset. CV is systematically biased upwards in its error because we necessarily have to reduce the size of our starting dataset.

b) What are the hyperparameters of a model? How are these typically chosen?
A hyperparameter is a model parameter (quantity that changes the behavior of a model) that is not derived during the fitting process. It must be set before fitting, and is often chosen to optimize the prediction error of the model.

c) What is regularization? What aspects of a model does it improve and worsen?
Regularization is an extra penalty applied to the fitting of a model to decrease its variance. Regularization always worsens the fitting error of the model, but in many cases may improve the prediction error.

d) What are three reasons to use a partial least squares model over a LASSO model? What are two reasons you might choose to use a LASSO model instead?
Benefits of PLSR: Thinking in terms of patterns rather than individual measurements, ability to explore both observation variation and input-output relationships, may be more predictive if there are many correlated variables in the data.
Benefits of LASSO: Results in a sparse model, may be more predictive if only a small set of variables contribute to predictions.

e) What are three advantages and disadvantages to a Bayesian analysis over a frequentist one? What are two circumstances (however rare) under which the two approaches exactly agree?
(1) A Bayesian analysis allows for one to contribute existing knowledge about a problem, outside of the data itself. (2) A Bayesian analysis can seem arbitrary due to use of a prior distribution. (3) Frequentist analyses can fail, especially with little data or with extreme events.

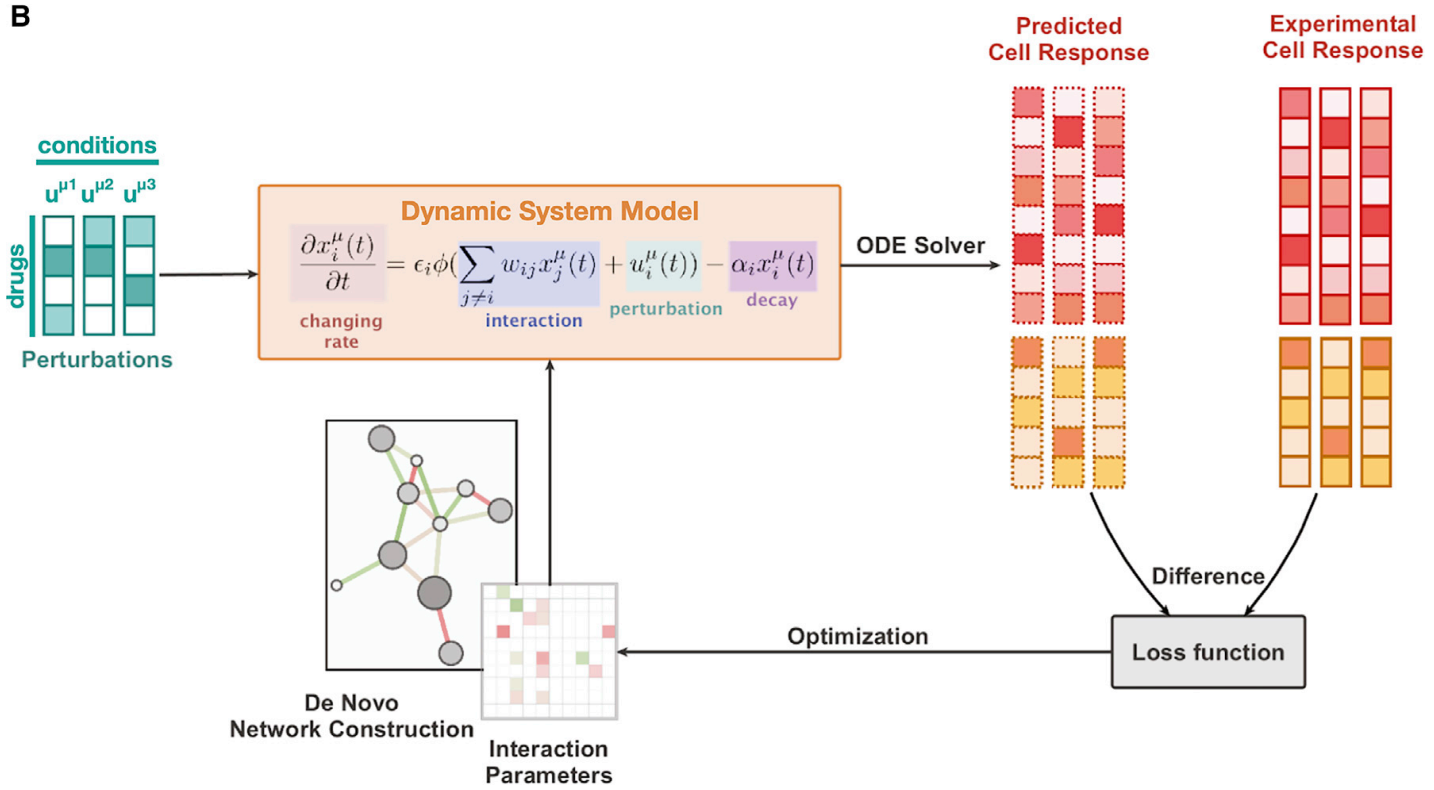
Both approaches can agree (1) when the prior is just right such that the posterior ends up the same or (2) at the limit of infinite data, which overwhelms the effect of the prior.

Question 2 (10 points)

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, which it uses to predict the effect of each perturbation. Predictions are made by running the model 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. Perturbations are performed by setting certain genes to an abundance of zero. 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?
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. 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 inhibit or knockout other genes. 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.

Question 3 (10 points)

When there is a constant risk of an event happening, such as one's disease progressing, the time to that event can be modeled as an exponential distribution:

$$p(t) = \lambda e^{-\lambda t}$$

This distribution forms the basis for the analysis of clinical trials, wherein each patient is assumed to be at constant risk of having some event of interest. A study is often looking for a significant difference in the time to an event across the group.

a) A process called censoring happens when an event does not happen by the end of the study for one of the patients. With censoring, the exact time to an event is not known, but it is known that it happens after the experiment ends. This means that we know the time to an event (t) is larger than the experiment time T, but we do not know t's exact value. Given an experiment runs for T amount of time, and the event has not yet happened, what is an expression for this probability?

$$p(t > T) = \int_T^\infty \lambda e^{-\lambda t} \delta t = e^{-\lambda T}$$

b) You run a clinical trial arm with three patients who each have an event at 1 year, 2 years, and no event before the trial ends at 3 years. What is the expression for the probability of observing this result (leave in terms of λ)? Does it depend on the length of the trial?
This is just a matter of plugging in the data, and recognizing that independent points multiply.
$p(\text{data}) = (\lambda e^{-\lambda})(\lambda e^{-2\lambda})(e^{-3\lambda}) = \lambda^2 e^{-6\lambda}$

c) Often clinical trials are designed such that a certain fraction of patients progress in their disease (i.e. have an event) before the trial is stopped. Say a clinical trial has precisely half of its patients progress in their disease by the end of 1 year. What is λ?
$p(t > T) = 0.5 = 1 - e^{-\lambda T}$
$0.5 = e^{-\lambda}$
$\lambda = 0.69$ years

d) Does this distribution show a positive or negative skew? What does this tell us about where we will tend to find events? (You can analyze this qualitatively but should show an expression.)
Positive skew. This can be shown a variety of ways. If you sketch it out there is a long tail to the right indicating that it has a positive skew. The integral is very nasty to analytically calculate the answer, but it comes out to 2 regardless of λ.

e) What can you say about the distribution of the average time to events across a trial arm (N=10), compared to the average time to an event distribution for individuals within the trial?
The average time to events is a sampling distribution, so: (1) It will tend to be more normal-like. (2) It will tend to have the same average as the distribution of individuals. (3) The variance of the distribution will be less.

Question 4 (10 points)

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.

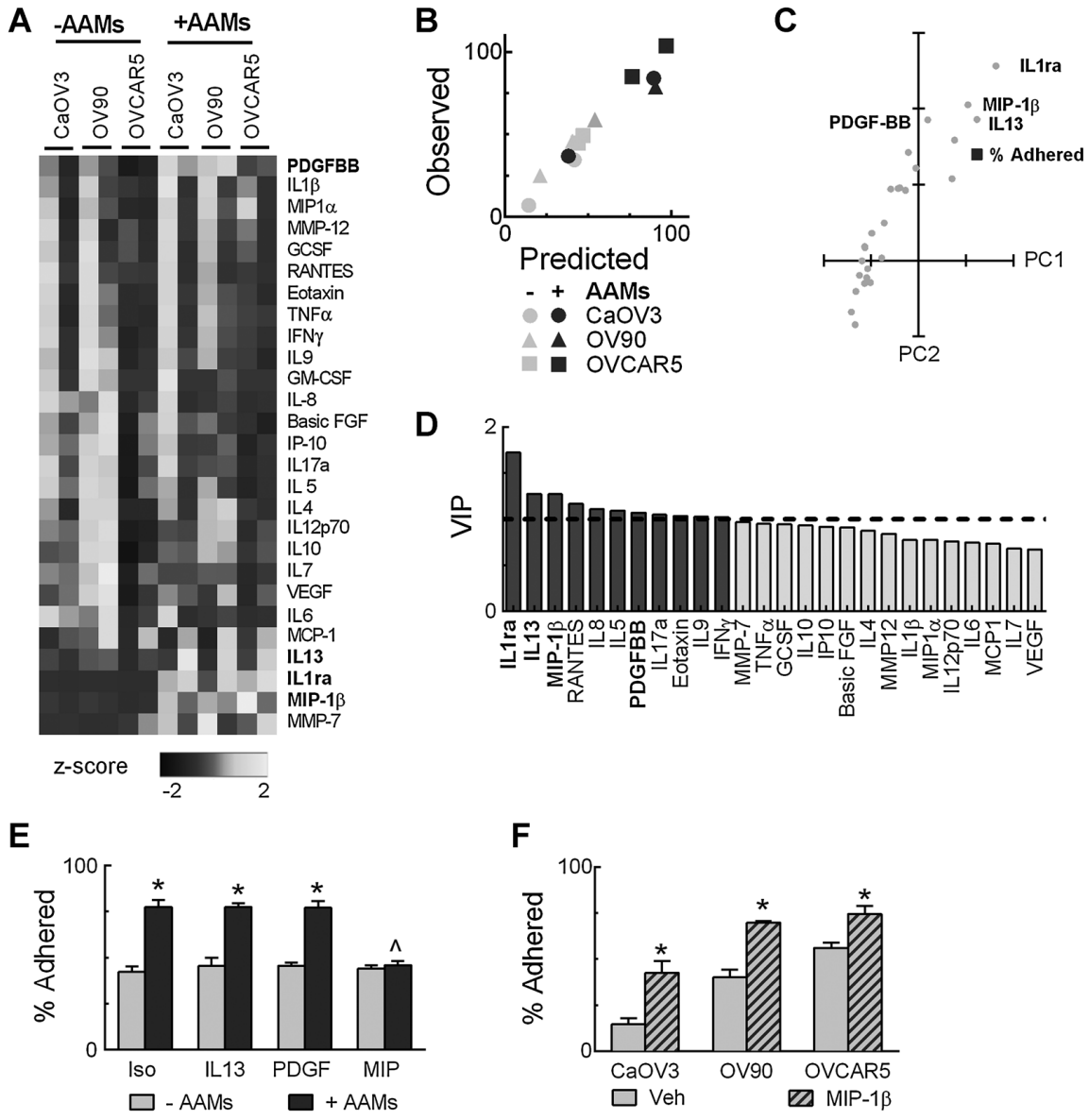


Figure 2.

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 procedure was used to calculate the results in Figure 2B?
Cross-validation.

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 instead want to calculate which scores would be consistently over 1 if the authors were to collect an entirely new dataset. 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 are PLSR PCs geometrically defined with respect to one another? 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, because it explains orthogonal variation in the covariance matrix. (Thus, PLSR components represent orthogonal patterns of variation in covariance space.) 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 (10 points)

Ford *et al*, *Clin Infect Dis*, 2021, report that a SARS-CoV-2 rapid test has a sensitivity of roughly 80% in symptomatic people and 40% in asymptomatic people. The specificity was determined to be more than 99.5% in both cases.

a) Write out Bayes' law, and then rewrite the equations to reflect the probability of an individual actually being SARS-CoV-2 *negative*, given they had a *negative* test result.

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

A stands for actually. T stands for test.

$$p(A\_ \mid T\_) = \frac{p(T\_ \mid A\_)p(A\_)}{p(T\_)}$$

b) The incidence of SARS-CoV-2 in Los Angeles on this day overall is about 1 in 10,000. 5% of those with related symptoms are turning out to be positive for SARS-CoV-2. Calculate the probability of both a symptomatic and asymptomatic person actually being negative, given they test negative on a rapid test. Is a tested symptomatic, or untested asymptomatic, individual more likely to be negative?

One estimate of the chance an untested, asymptomatic individual might have SARS-CoV-2 is the overall incidence rate. So, the chance of them not having it would be 99.99%.

For a symptomatic individual:

$$p(A_- \mid T_-) = \frac{p(T_- \mid A_-)p(A_-)}{p(T_-)} = \frac{p(T_- \mid A_-)p(A_-)}{p(T_- \mid A_-)p(A_-) + p(T_- \mid A_+)p(A_+)}$$

$$p(A_- \mid T_-) = \frac{(0.995)(0.95)}{(0.995)(0.95) + (0.2)(0.05)} = \frac{18.905}{18.905 + 0.01} = 99.95\,\%$$

Therefore, the tested symptomatic individual is still less likely to be negative.

You are working on deploying a new medical device in hospitals and want to ensure there are sufficient backups in place in case one fails. To understand this, you want to model the amount of time it takes a device to fail. You expect that failures are at constant risk over time, and so you model the time to failure as an exponential distribution:

$$p(t) = \lambda e^{-\lambda t}$$

c) You want to use $p(\lambda) = 1/\lambda$ as your prior expectation of the failure rate (in units of years). So far, one device failed at 1 year, and another at 2 years. Derive an expression for the posterior distribution of the failure rate.

First, use Bayes' law.

$$p(\lambda \mid t) = \frac{p(t \mid \lambda)\, p(\lambda)}{p(t)} = p(t \mid \lambda)p(\lambda)$$

$$p(\lambda \mid t) = \lambda e^{-\lambda} \lambda e^{-2\lambda} \frac{1}{\lambda} = \lambda e^{-3\lambda}$$

$$\int_0^\infty \lambda e^{-3\lambda} \delta\lambda = 1/9$$

$$p(\lambda \mid t) = 9\lambda e^{-3\lambda}$$

d) For a certain failure rate, the Binomial distribution gives the probability of k devices out of n failing within a single year:

$$p(n,k) = \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$$

Derive an expression for the chance of seeing 2 devices out of 4 fail in a given year, given your observations in (c).

$$p(4,2) = \binom{2}{4} \lambda^2 (1-\lambda)^2 = 6\lambda^2 (1-\lambda)^2$$

$$\int_0^\infty (9\lambda e^{-3\lambda})(6\lambda^2 (1-\lambda)^2) \delta\lambda$$