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?

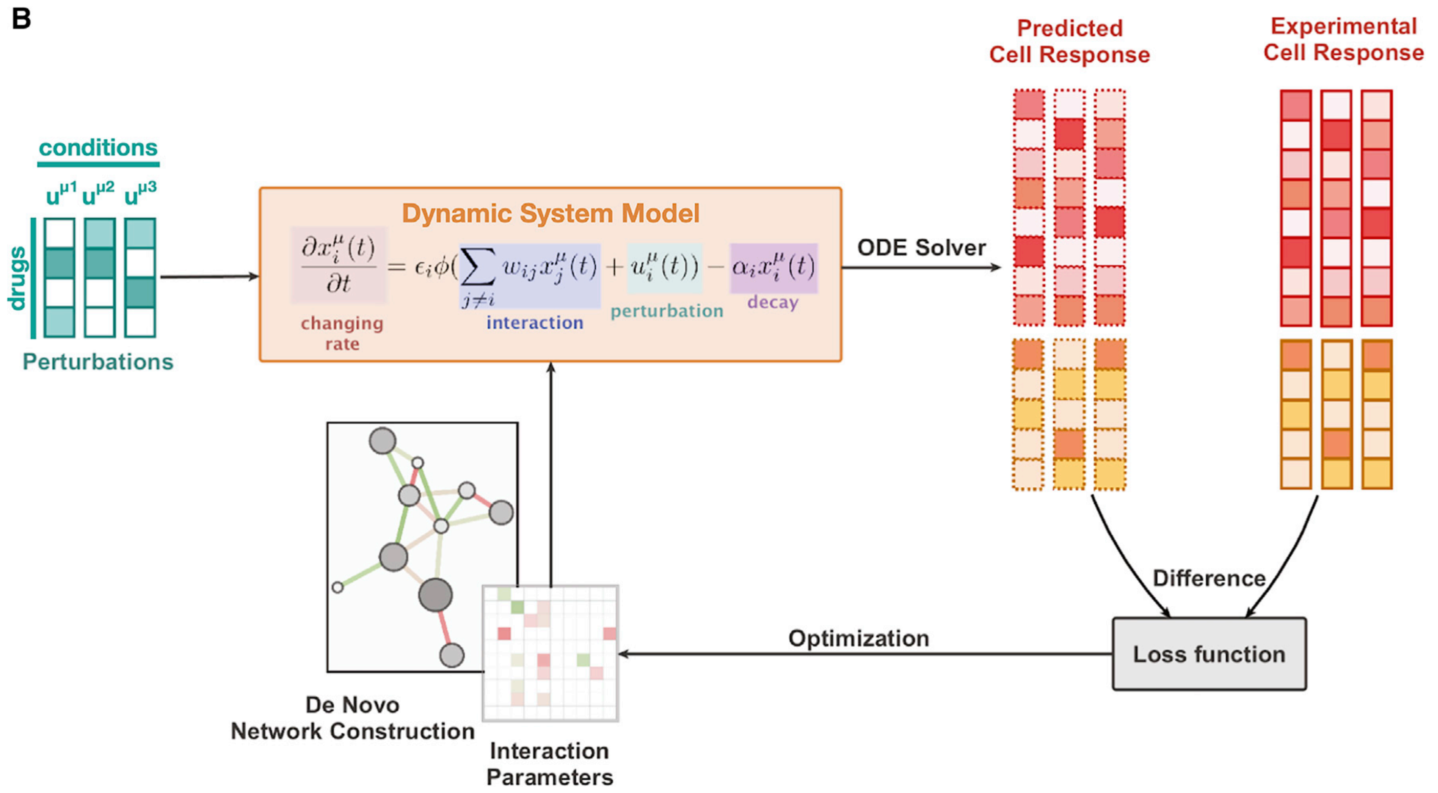
b) What are the hyperparameters of a model? How are these typically chosen?

c) What is regularization? What aspects of a model does it improve and worsen?

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?

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?

Question 2 (10 points)


B
conditions
$u^{\mu1}$ $u^{\mu2}$ $u^{\mu3}$
drugs
Perturbations
Dynamic System Model
$\frac{\partial x_i^\mu(t)}{\partial t} = \epsilon_i \phi(\sum_{j \neq i} w_{ij} x_j^\mu(t) + u_i^\mu(t)) - \alpha_i x_i^\mu(t)$
changing rate
interaction
perturbation
decay
ODE Solver
Predicted Cell Response
Experimental Cell Response
Difference
Loss function
Optimization
De Novo Network Construction
Interaction Parameters

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?

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.

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.

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?

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 #2?

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?

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?

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?

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.)

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?

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.

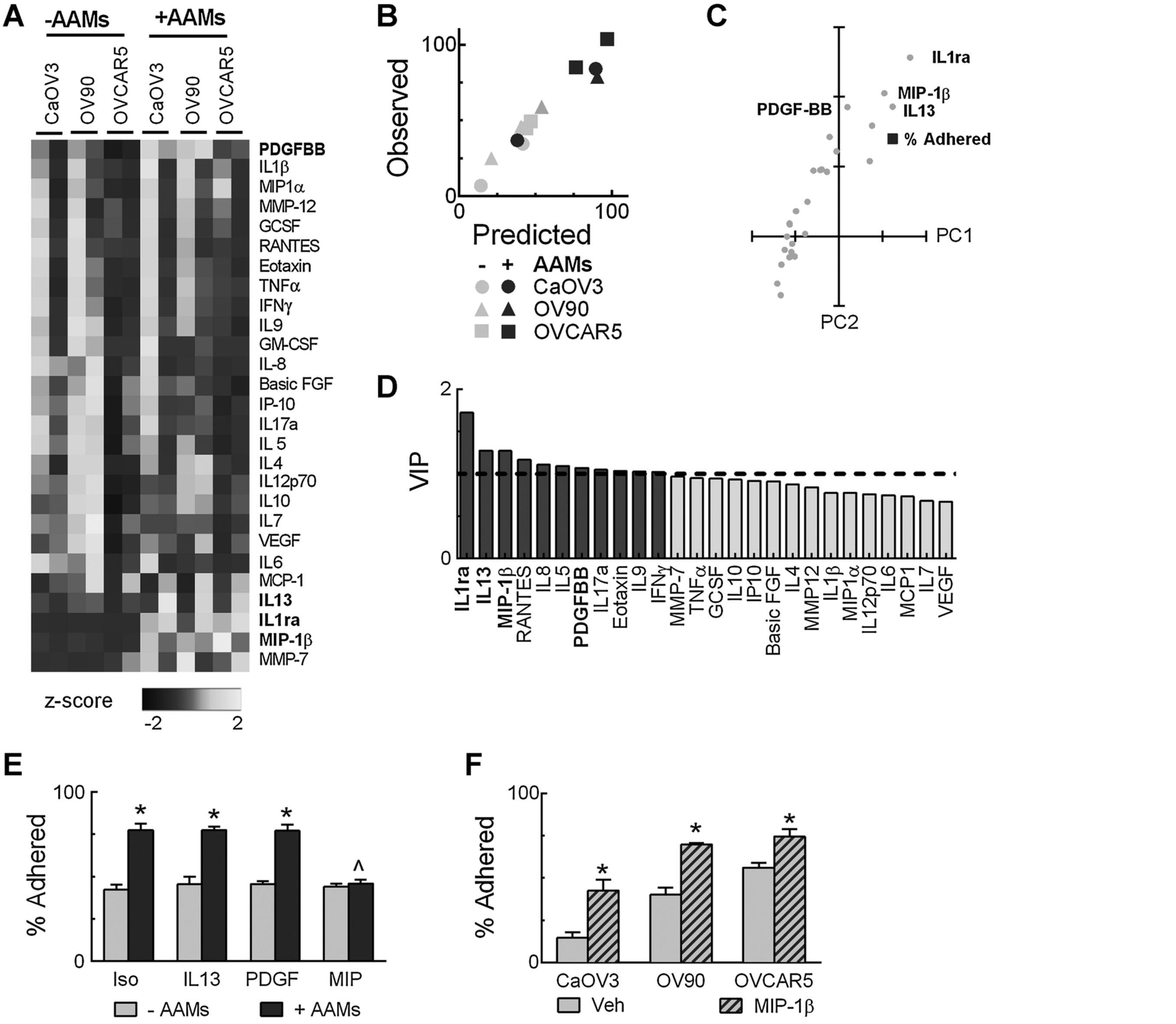


Figure 2.

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?

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?

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.

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.)

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?

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.

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?

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.

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).