- Chose this dataset because it is well packaged.

Future Tasks:

- Compare SelAC's inference with permutations

4 ROC SEMPPR

5 C. elegans

Project State:

- Same as previous entry, but a better understanding.

Current Tasks:

- It appears that the algorithm is partitioning all of the noise in the data to ROC.
- Run with constraint $\beta > 0$ may not have stabilized.
- Rerun using estimates of error in ϕ_{ROC} from ROC-SEMPPR model fits.
  - Did so with estimates of sd in ϕ ROC on the natural, rather than log scale.
  - Fits are quite different b/w the simplex and vector results. However, should be skeptical since they were run with single, rather than multiple chains.
  - Cedric is re-running ROC to get ϕ ROC sd on log ϕ scale.

6 LSA

Kevin Dunn and Mehmet

Project State:

- Up until now Kevin had been taking a 'machine learning' approach rather than regression.
- I finished processing the pre- and post-course exam respones.
  - We now have individual answers on a per student basis.
  - Have provided Kevin with anonymized data and mapping between post-course test questions (which included other questions) and pre-course test questions.

- Have directed Kevin to use 'stan' as Cedric has been doing in R. For the differences in the pre-post score (response variable), we can quantify our uncertainty using the binomial. Assuming n questions on both pre- and post- exams,

$$\mathrm{sd}_{\Delta\mathrm{Score}} = \sqrt{nk_1/n(1-k_1/n)} + \sqrt{nk_2/n(1-k_2/n)}$$

  where k_1 and k_2 are the number of correct answers in the pre- and post-course exams.
- An alternative approach would be to try and predict k_2 using everything, including k_1. This might actually make more sense.
- In addition, Kevin should be able use the SD in the component scores to describe our uncertainty in each of them.

Current Tasks: Write up of model for Kevin to fit is below.

Let $\vec{X_{i,j}}$ represent the set of $n_{i,j}$ completed responses submitted by student i that are used to calculate metric j which can be verbal score, total words, etc. Thus, the student's mean response is $\bar{X}_{i,j} = 1/n_{i,j} \sum_k^{n_{i,j}} X_{i,j,k}$. Similarly, the standard deviation in the individual student's responses is $S_{i,j} = \sqrt{\mathrm{Var}\left(\vec{X_{i,j}}\right)}$. The above predictors are measures of quality of student responses. In order to include the quantity of student reponses as a predictor, I propose we use $X_{i,q}$ = Total number of text questions completed/Maximum total number of questions that could have been completed.
Rather than use the difference in pre- and post-course exam scores, I propose we use the pre-exam score as another predictor where $X_{i,\mathrm{pre}} = k$ the number of questions the student answered correctly in the pre-course exam. Under the bionomial model. our uncertainty in $X_{i,\mathrm{pre}}$ is $S_{textpre} = \sqrt{m(1-k/n)}$ where n here is the number of questions in the pre-course exam. Similarly, the response variable we are trying to predict would be $Y_{i,\mathrm{post}} = m$ the number of questions answered correctly and our uncertainty in this imprecise metric is $S_{textpost} = \sqrt{m(1-m/n)}$.

29 April 2016

1 LSA

Kevin Dunn and Mehmet

TODO list

Kevin Dunn

- Document functions and pipeline
- Clearly comment current code
- Create unit tests and overall pipeline
- Determine required subcomponents of NLTK that need to be installed to run code. (Full NLTK install is 10GB!)

Mehmet Learn more about running stan

Me Refine model for stan

Current model

$$\begin{aligned} Y &= \text{Student's post course GCA score} \\ &\sim N(\mu = y, \sigma = \sqrt{k(1-k/m)}) \end{aligned}$$

where y is the students 'true' post-course knowledge and σ is our uncertainty in this derived from binomial distribution where m is the number of GCA questions and k is the number the student answered correctly

$$\begin{aligned} X_0 &= \text{Student's pre course GCA score} \\ &\sim N(\mu = x_0, \sigma = \sqrt{k(1-k/m)}) \end{aligned}$$

where x_0 is the students 'true' pre course knowledge

$$\begin{aligned} \vec{X}_i &= \text{Student's individual measurements of metric } i \\ \bar{X}_i &= \text{Mean metric } i \text{ of student's LSA responses} \\ &\sim N(x_i, \sigma = s_{X_i}/\sqrt{n_i}) \end{aligned}$$

where the σ is the standard error of our estimator $\bar{X}_i$ for x_i

$$\begin{aligned} n_i &= |\vec{X}_i| \\ &= \text{Number of observations of metric } X_i \text{ used to calculate } \bar{X}_i. \end{aligned}$$