## the data that was used and the models that were developed to predict the effect of stimulation parameters and pre-stim state on the changes in the biomarker

## DATA

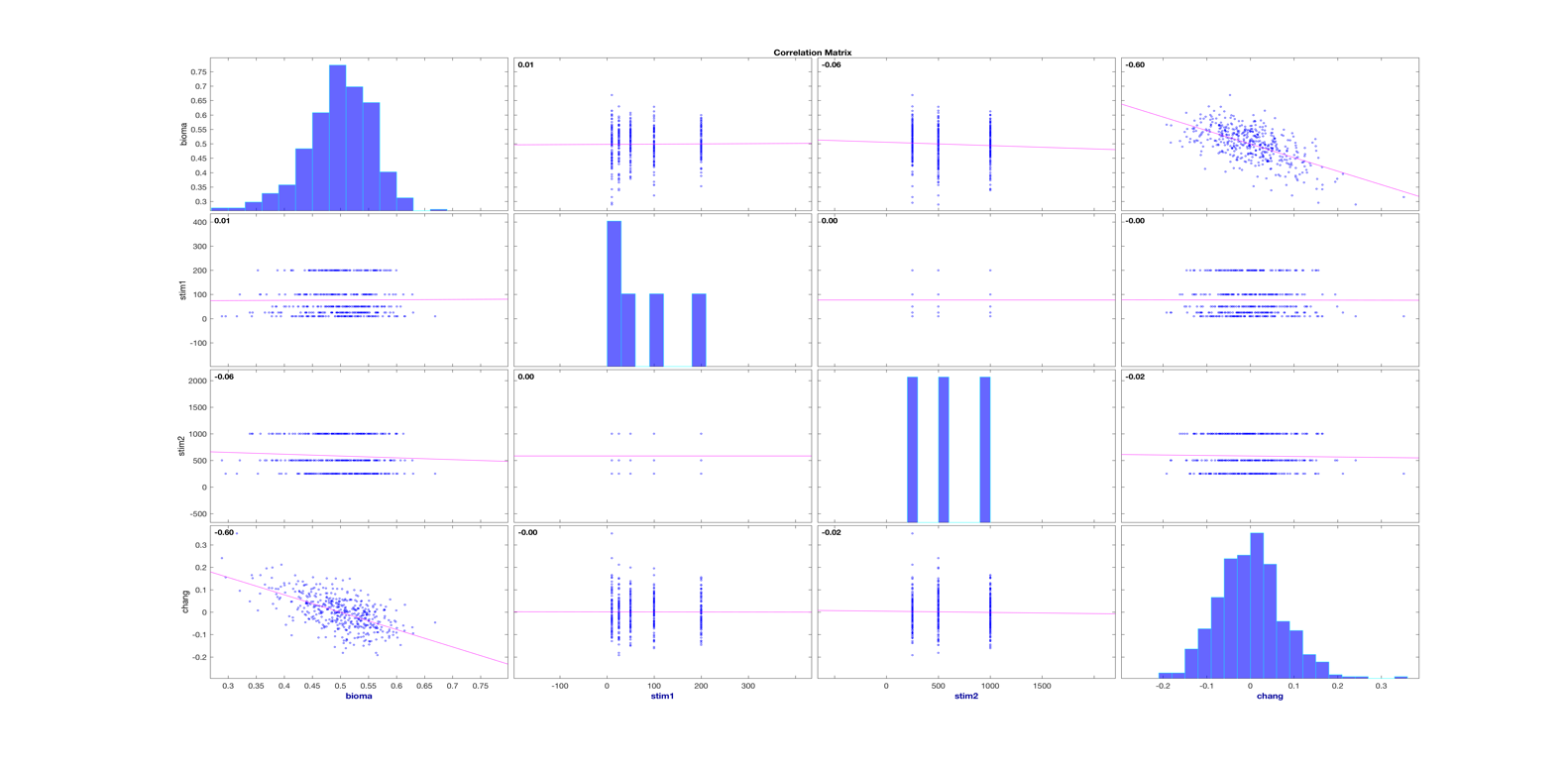
For 59 locations, the following data was used to predict Y from X

X = [scalar biomarker, stimulation frequency, stimulation duration]

Amplitude was always constant = 1

Y = [change in biomarker]

The pairwise scatterplot is below:



Description of figure:

- this is the output of corrplot (see [documentation](http://www.mathworks.com/help/econ/corrplot.html))

- there are four vectors we are examining: three columns of X (x1,x2,x3), one column of Y (defined above)

- diagonal is the histogram for each of those four vectors

- off-diagonal subplots are pairwise scatter plots, ie. row1-col2 is x1 against x2, row1-col4 is x1 against Y

Conclusions from scatter plots:

- no apparent nonlinear pattern in the point spreads

- biomarker has a negative correlation with the change

- stim1 has no real correlation with the change

- stim2 has a slight negative correlation with the change

## MODELS

All were regression models

* Linear least-squares fit as a baseline
* Lasso regularization
* SVM with kernels: Gaussian, RBF, and linear (dot)
  1. Linear as expected achieved results similar to linear least-squares
* Regression trees
  1. Like decision trees but for regressions
* Neural network
  1. Inputs, one hidden node layer, output
  2. Hidden nodes ranged from 2 – 20
  3. Solvers tried:
     1. Levenberg-Marquardt – the default
     2. Bayes regularized, this trims hidden nodes to avoid over-fitting

## 2- the results that you got

* Focused most of analysis and experimentation on the first location in the cell array, hoping to find something to generalize to other locations
* Cross validation:
  1. Hold one observation out, train on remaining, test against that one held out
  2. Compute the r^2 between the true Y and the regression prediction Y\_hat
  3. Compute the linear correlation coefficient (corr) between Y,Y\_hat
  4. Note that within each training, NNet and lasso internally separated to make test/train sets
     1. NNet separates the data into train/test/validate = 75/15/15%
     2. Lasso uses KFolds=10
* **Nothing beat linear least-squares fit**
* SVMs with linear kernel matched or *slightly* outperformed
* NNet with Bayes regularization was the next closest
* SVM with nonlinear kernels (gauss,rbf) performed worse
* Running linear across all locations in the cell array:

R^2        mean 0.309486      std 0.0848791     min 0.146309       max 0.67754

rho       mean 0.551539      std 0.0739574     min 0.382749       max 0.823137

* Also computed the correlation (rho) of biomarker with change in biomarker across all locations to get a sense of the inherent linear correlation. The hope is that one of these regression models might pick up an underlying relationship that is stronger than this linear correlation.

rho   mean -0.5545    std  0.0721    min -0.8265    max -0.3940

* Back to looking only at sample location #1 (first cell), I tried running with/without the stimulation parameters, and there's no real difference.

X = [biomarker  stim1 stim2]

y\_linear       r2=0.3558   rho   0.5966

y\_svm          r2=0.3612   rho   0.6013

X = [biomarker]

y\_linear       r2=0.3581   rho   0.5985

y\_svm          r2=0.3616   rho   0.6017

## 3. your conclusion

* linear least-squares regression performed the best, and even that was poor
* all nonlinear methods overfit, and efforts at regularization barely started approaching that of linear least-squares
* while there are of course many more methods for regression, the methods tried represent a fair sampling of available techniques and efforts at exploring various parameters/options. Since none of them showed promise, it suggests that **this biomarker with/without stimulation parameters is not (easily) predictive of the biomarker change**.

## suggested next steps

* add more features
  1. could this scalar biomarker be expanded?
  2. Additional biomarkers?