Knowing when to Quit:

Why do we waste so much time trying to get models out of bad data?

Barry M. Wise and Jeremy M. Shaver
Eigenvector Research, Inc.

EIGENVECTOR
RESEARCH INCORPORATED

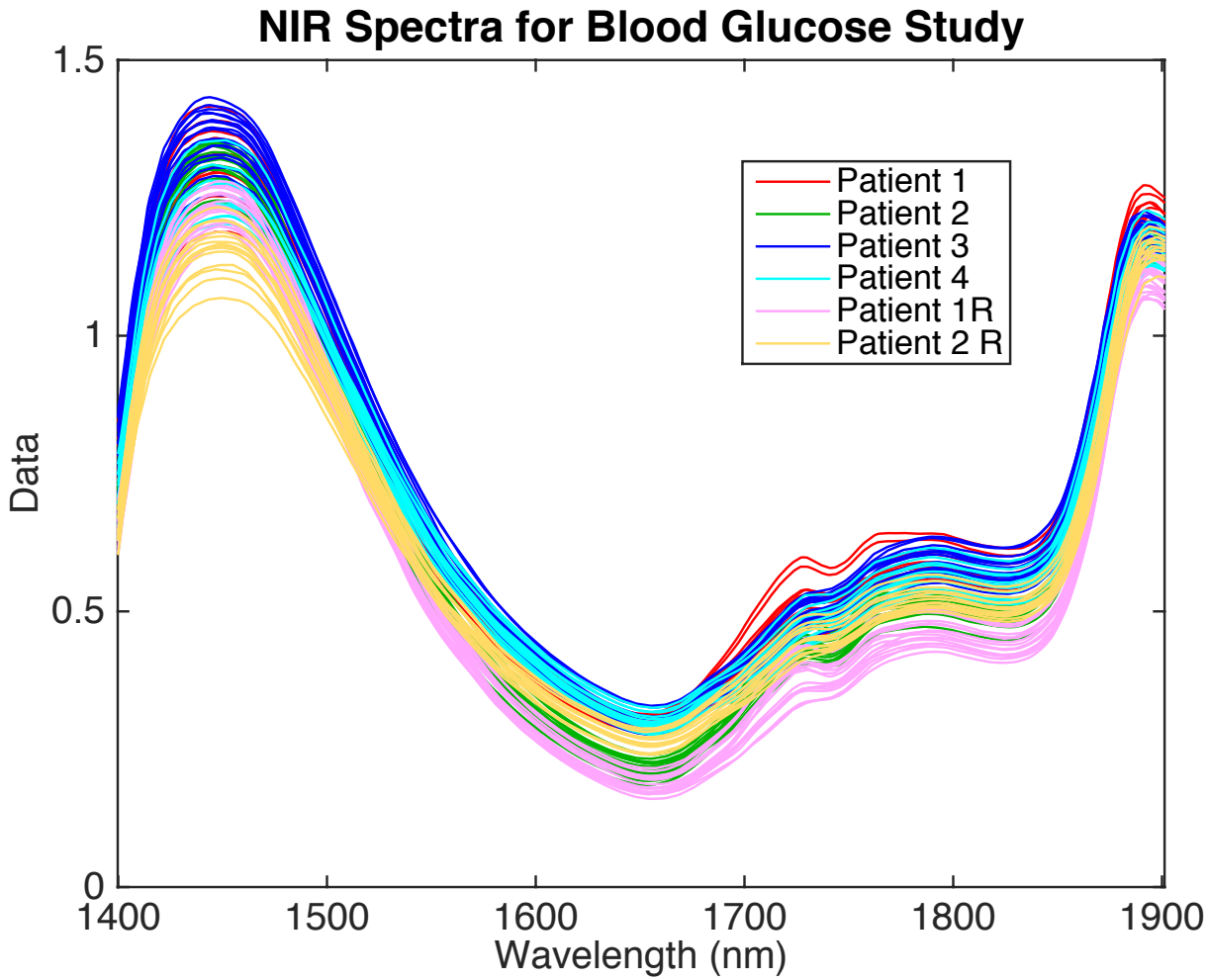
Outline

- The Problem
- An example
- Conclusion
- Discussion

The Problem

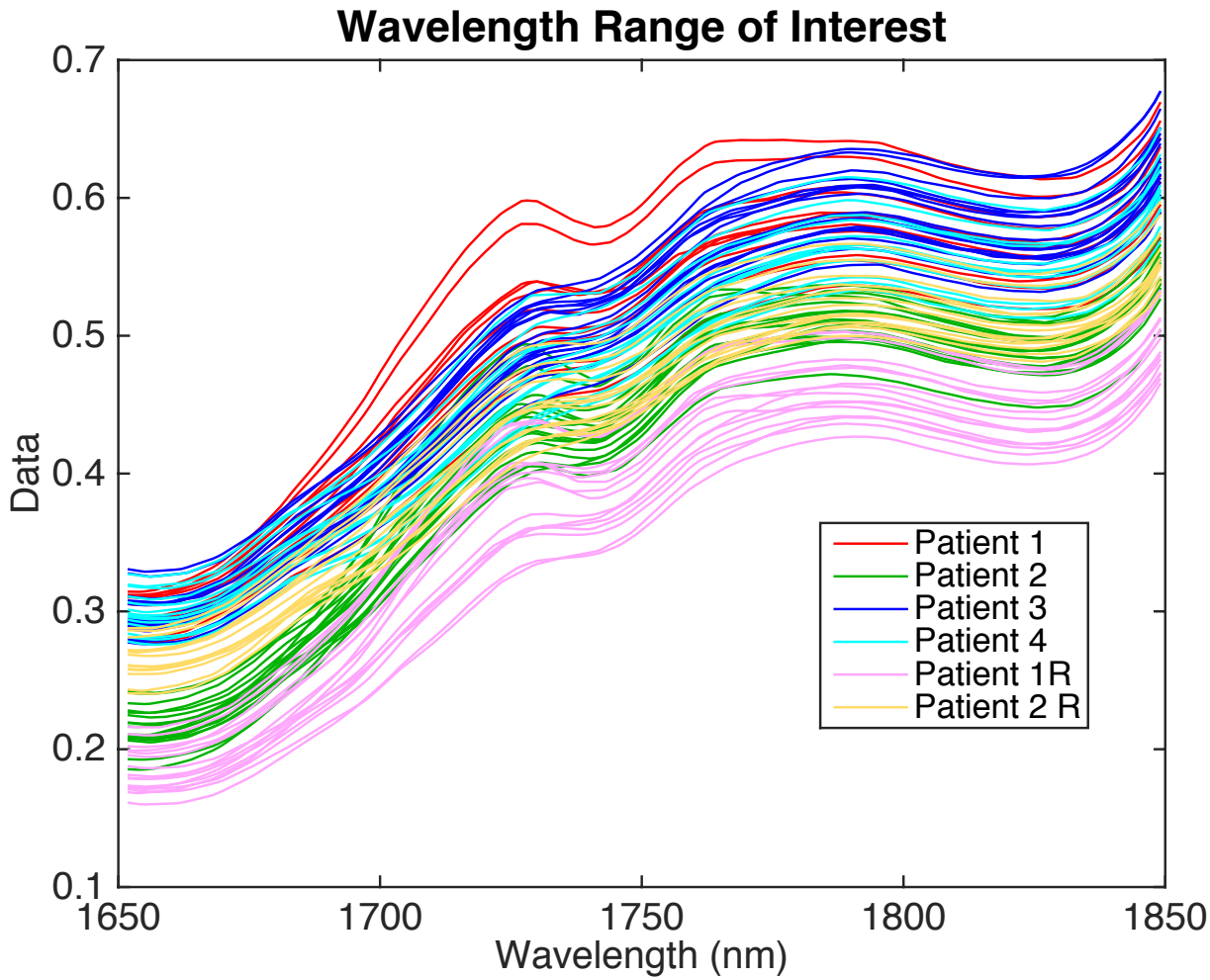
- Lots of bad data out there!
- Engineers and scientists developing methods based on more complex signals
  - Using spectroscopy on everything!
- Problems complicated by low signal-to-noise and/or low signal-to-clutter
  - Blood glucose
- May not be possible to get a reasonable model in some of these situations

Example Data Set
Blood glucose by NIR (again!)



EIGENVECTOR RESEARCH INCORPORATED

Region of Interest

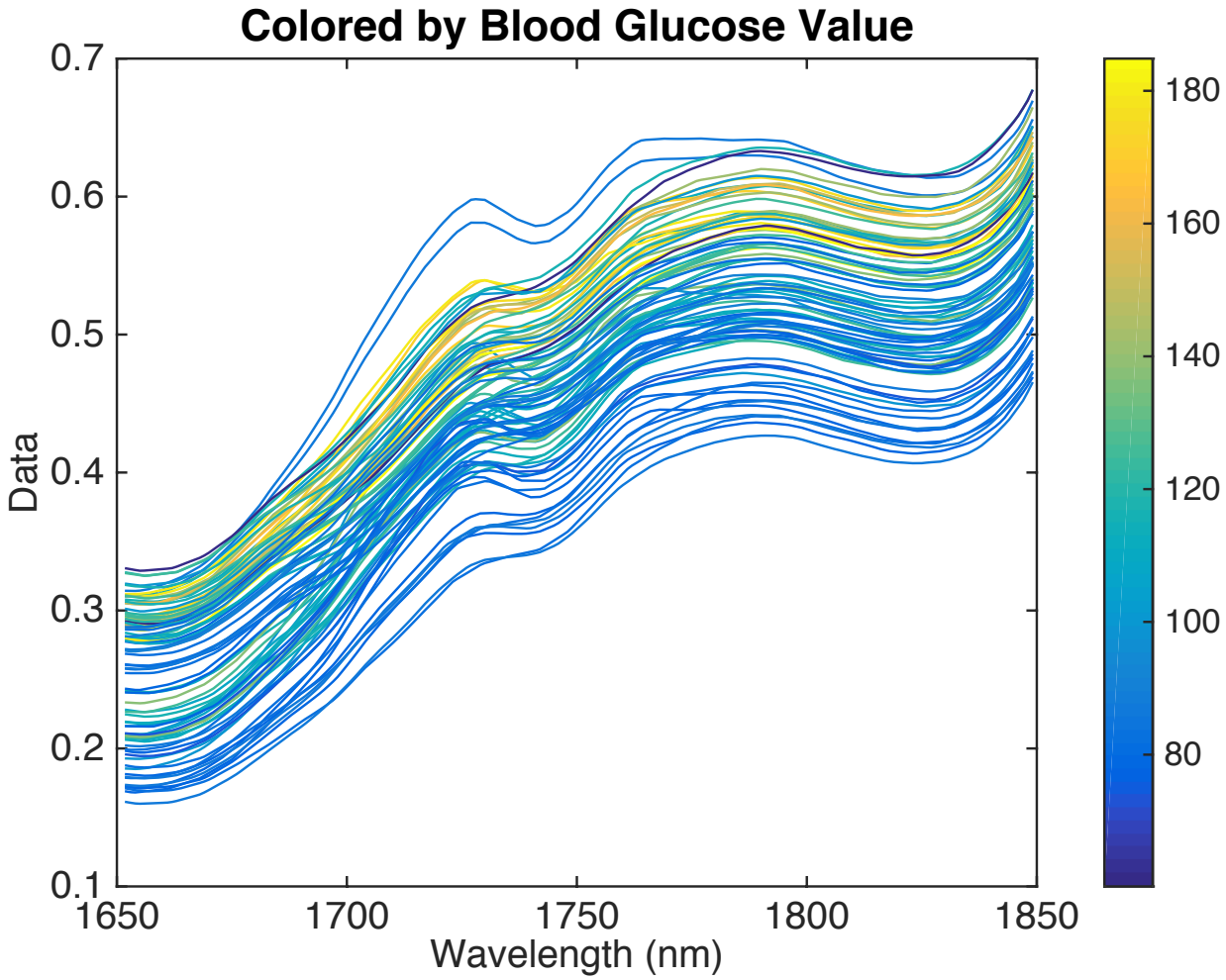


EIGENVECTOR RESEARCH INCORPORATED

What to try

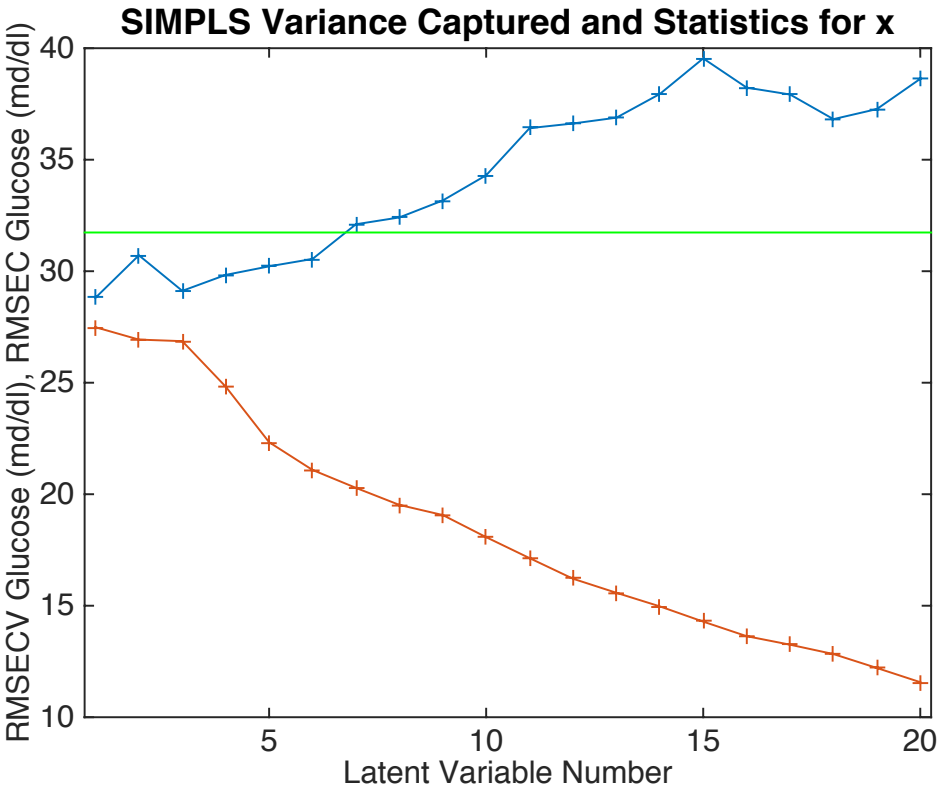
- Basic “plain vanilla” models
- Look at preprocessed data and color-by
- Nearly exhaustive search—Model Optimizer
  - Try to stick with “reasonable” preprocessing
- Interrogate best models
  - Over-fit?
  - Any reason model be over-optimistic?
  - Squint test

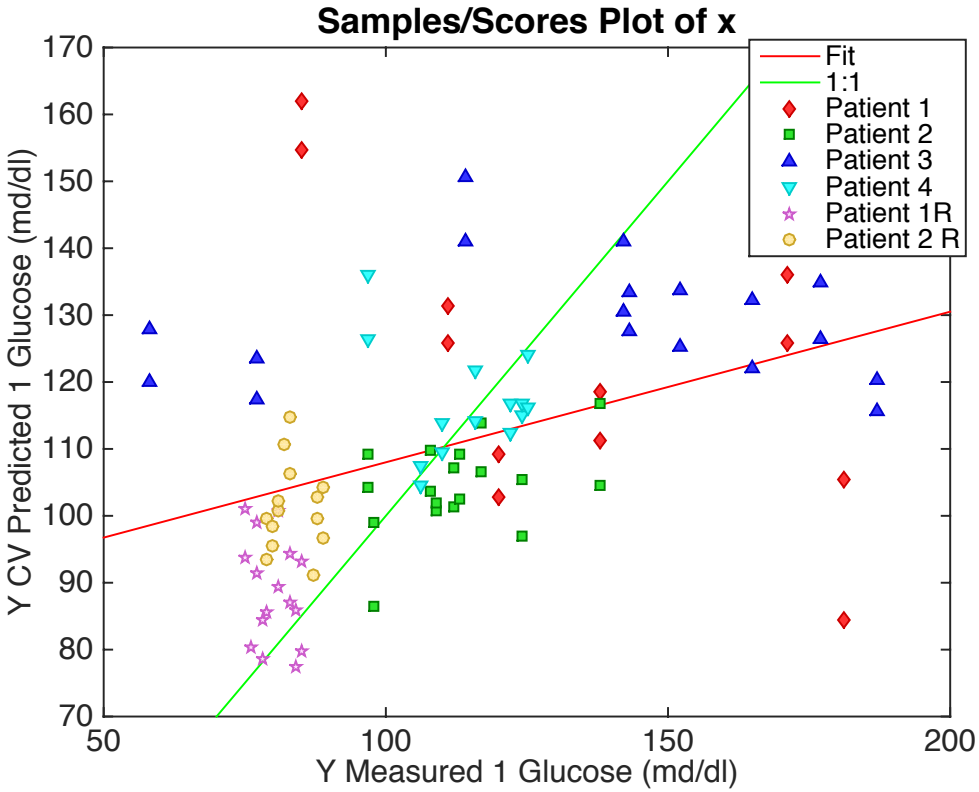
No Obvious Trend



EIGENVECTOR RESEARCH INCORPORATED

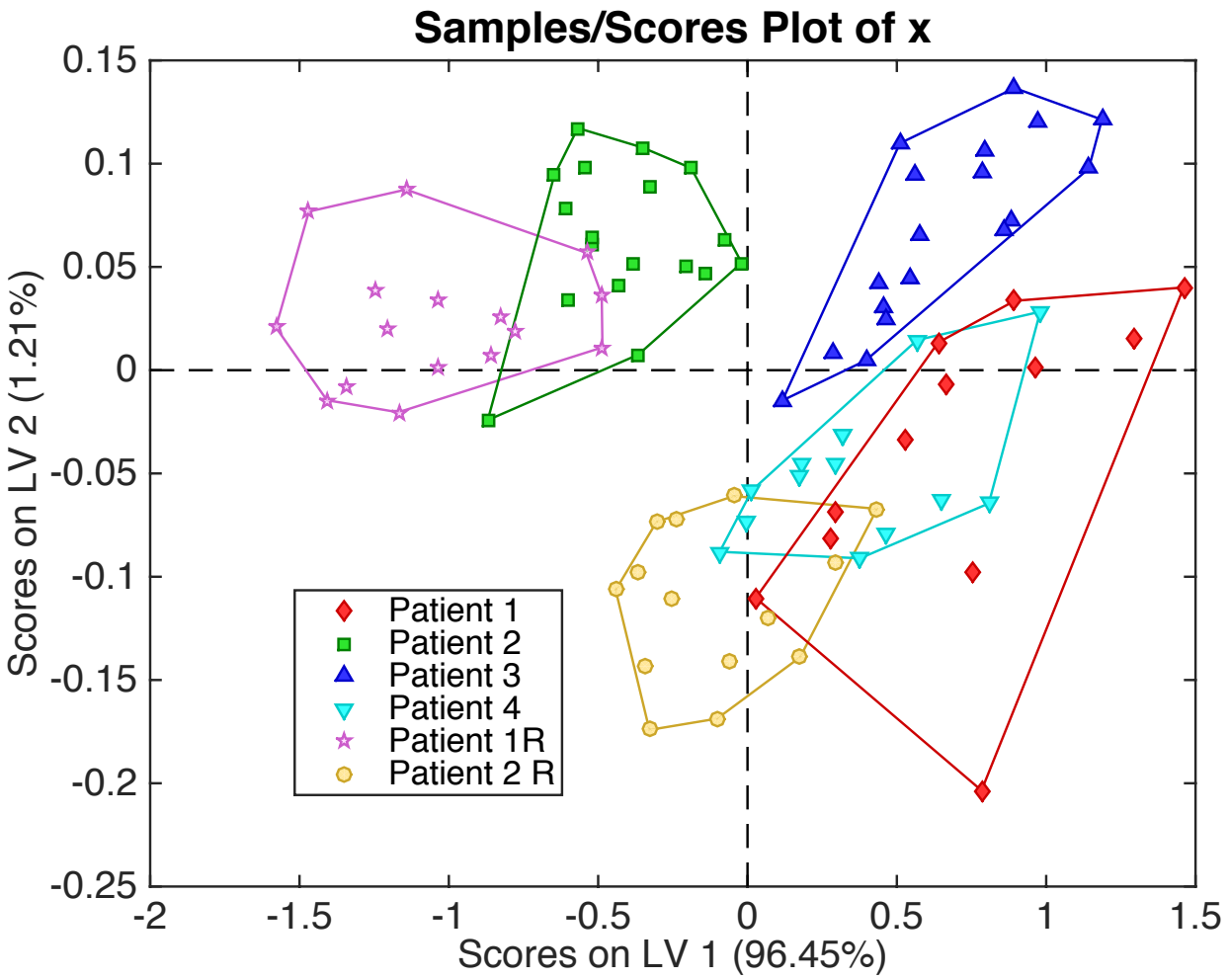
Initial Model





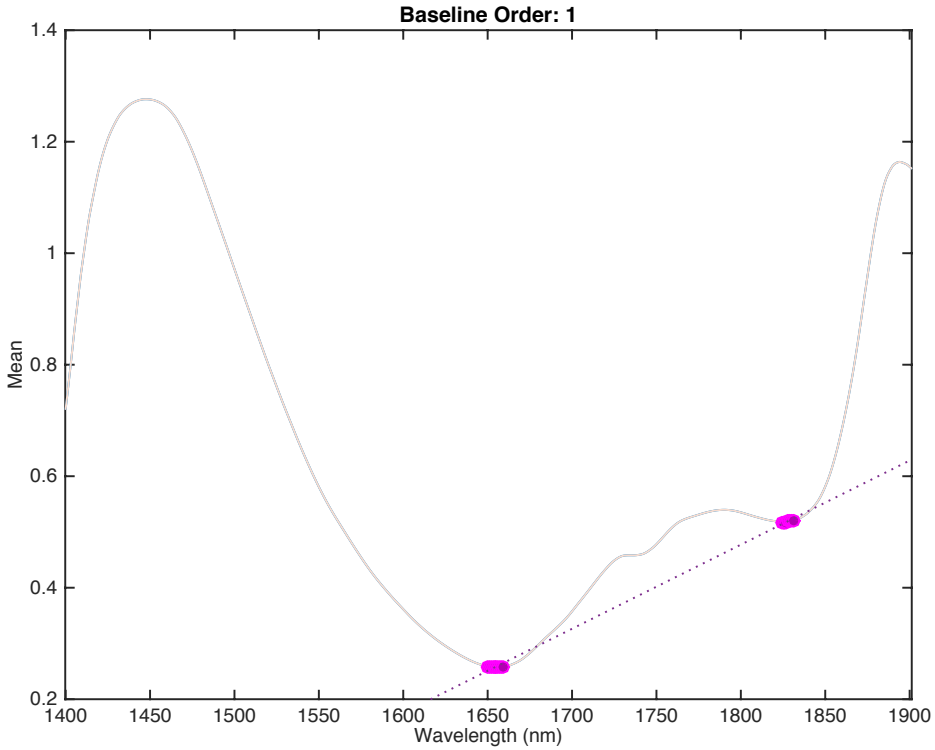
EIGENVECTOR RESEARCH INCORPORATED

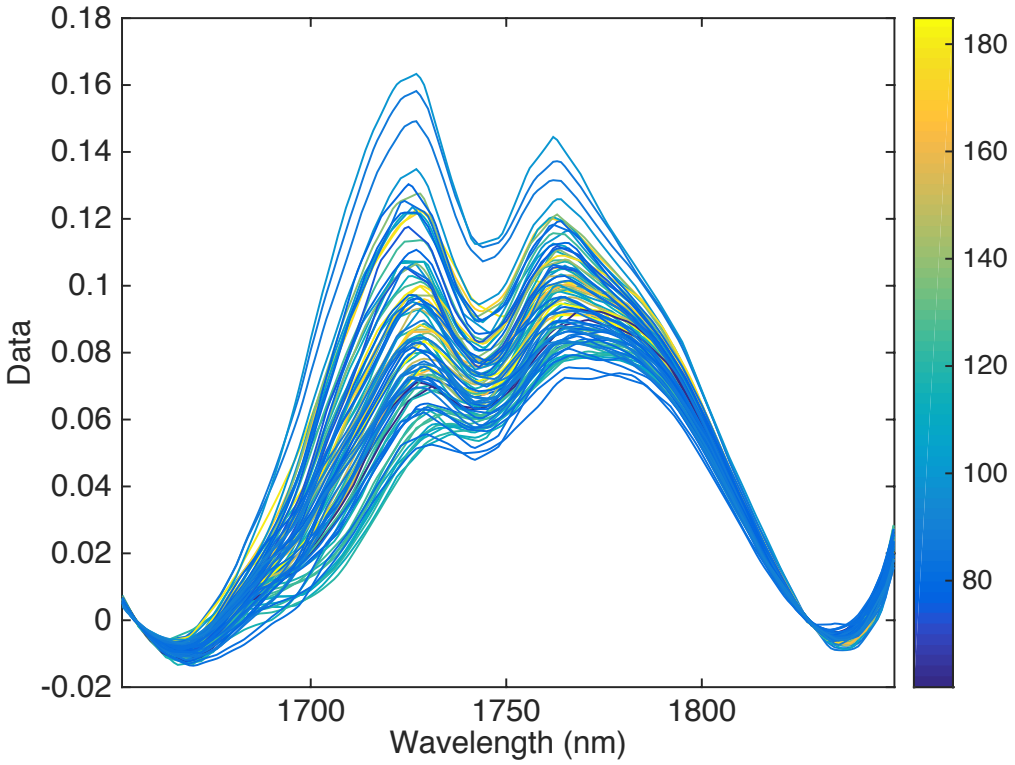
Scores from MC Model



EIGENVECTOR RESEARCH INCORPORATED

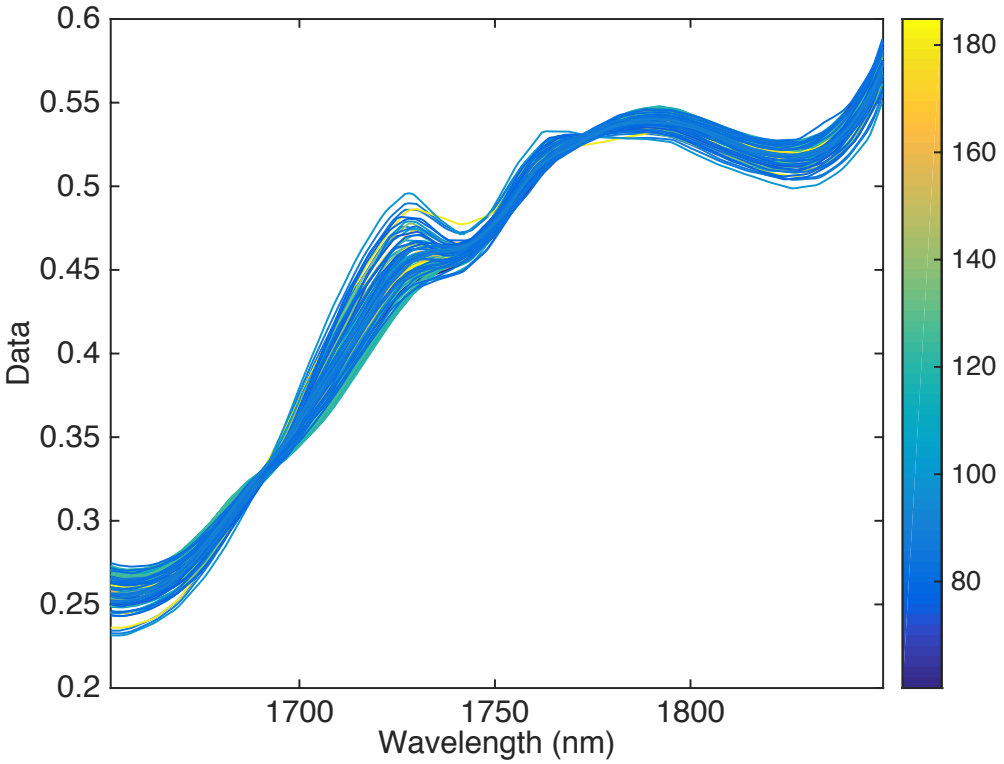
Try Different Preprocessing

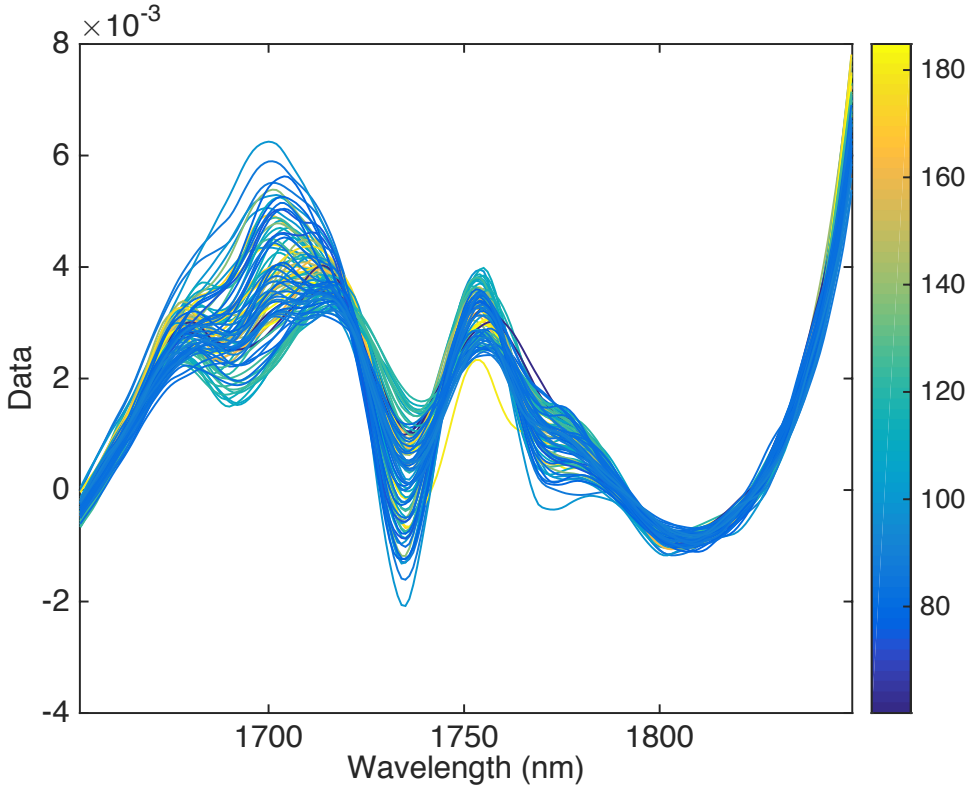




EIGENVECTOR RESEARCH INCORPORATED

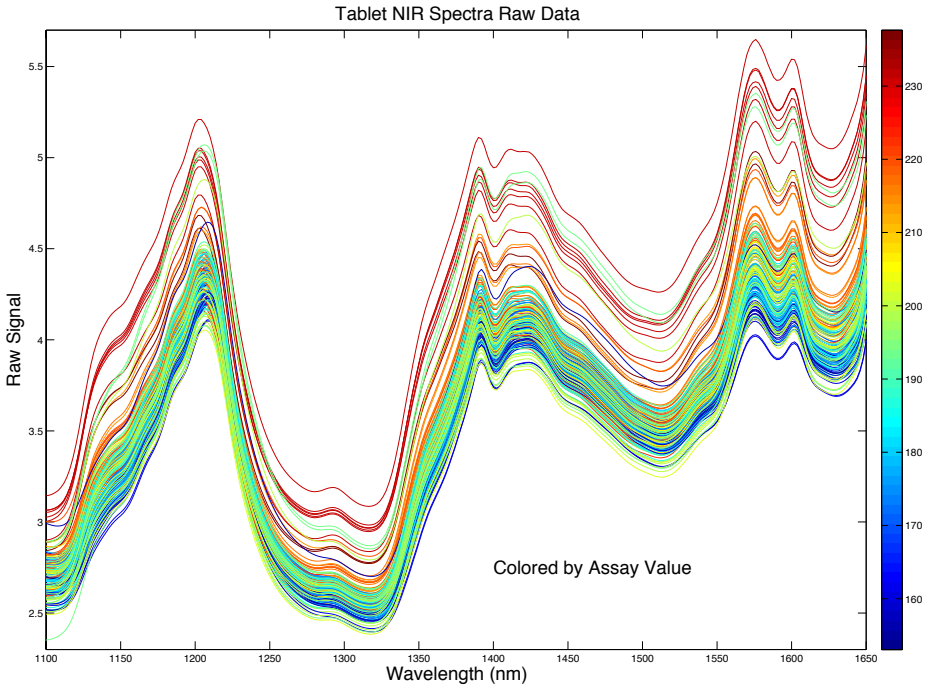
Other Preprocessing

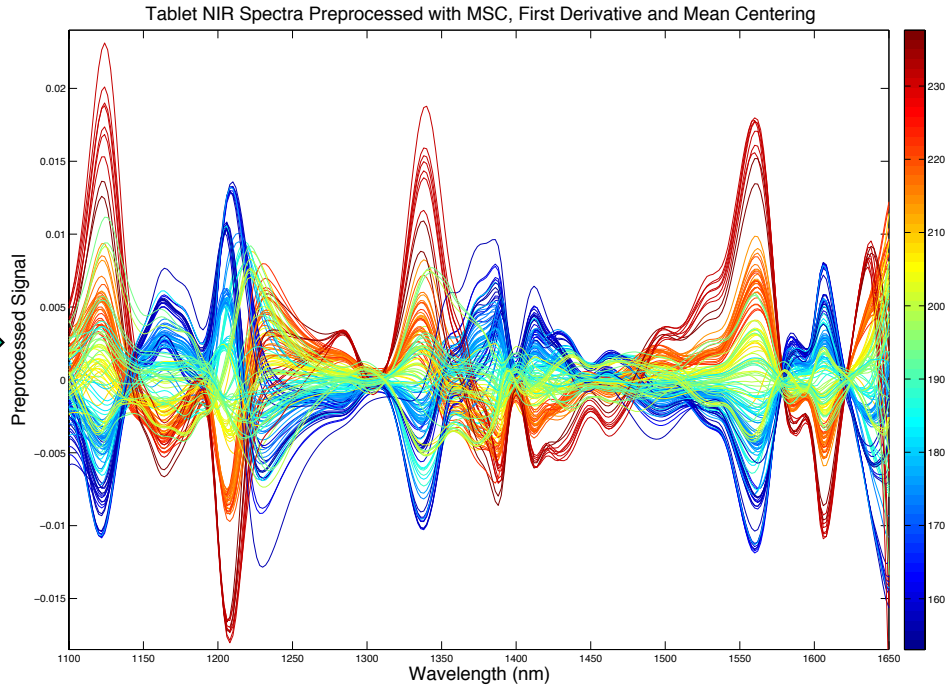




EIGENVECTOR RESEARCH INCORPORATED

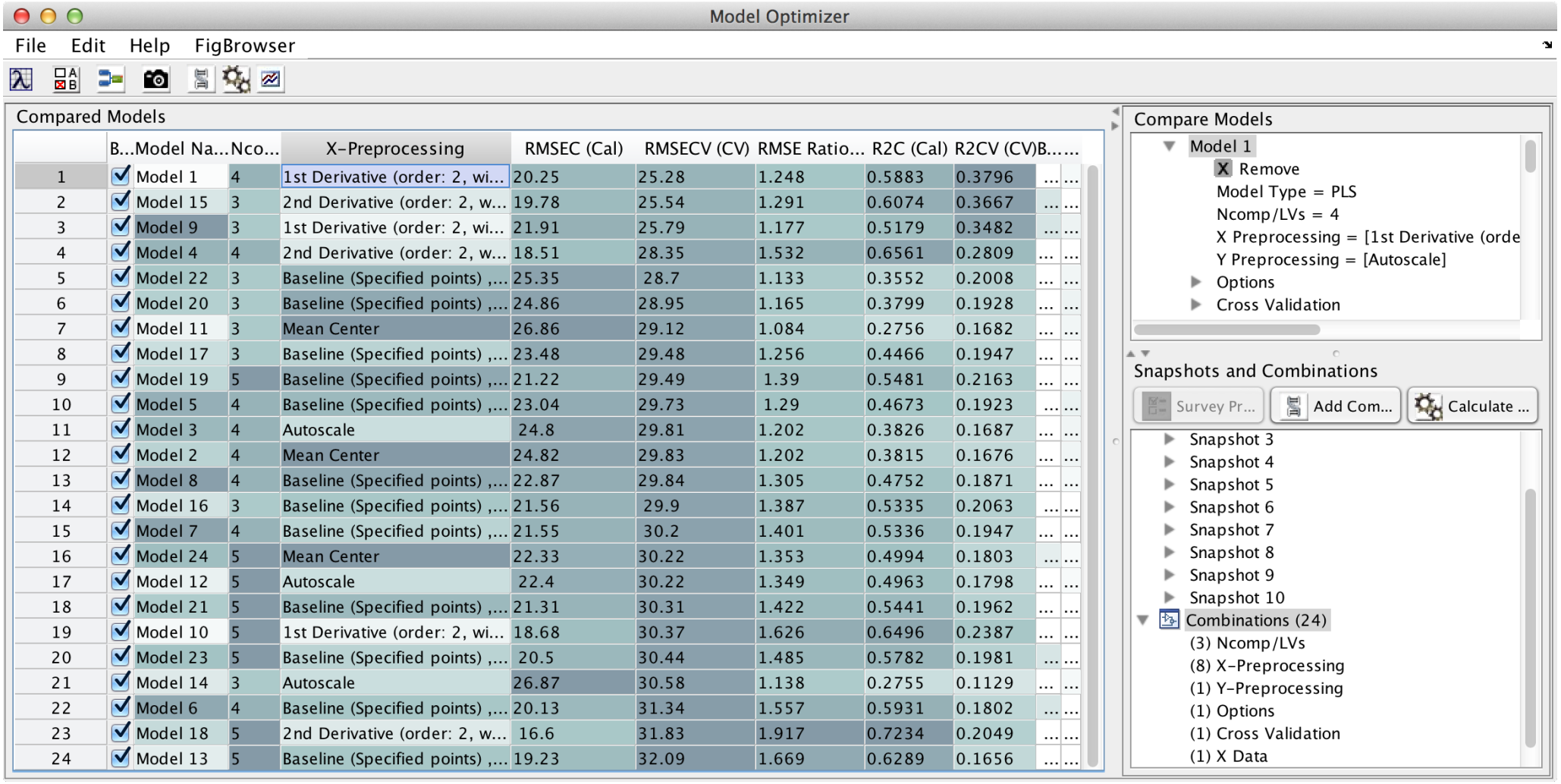
When Preprocessing Works!





EIGENVECTOR RESEARCH INCORPORATED

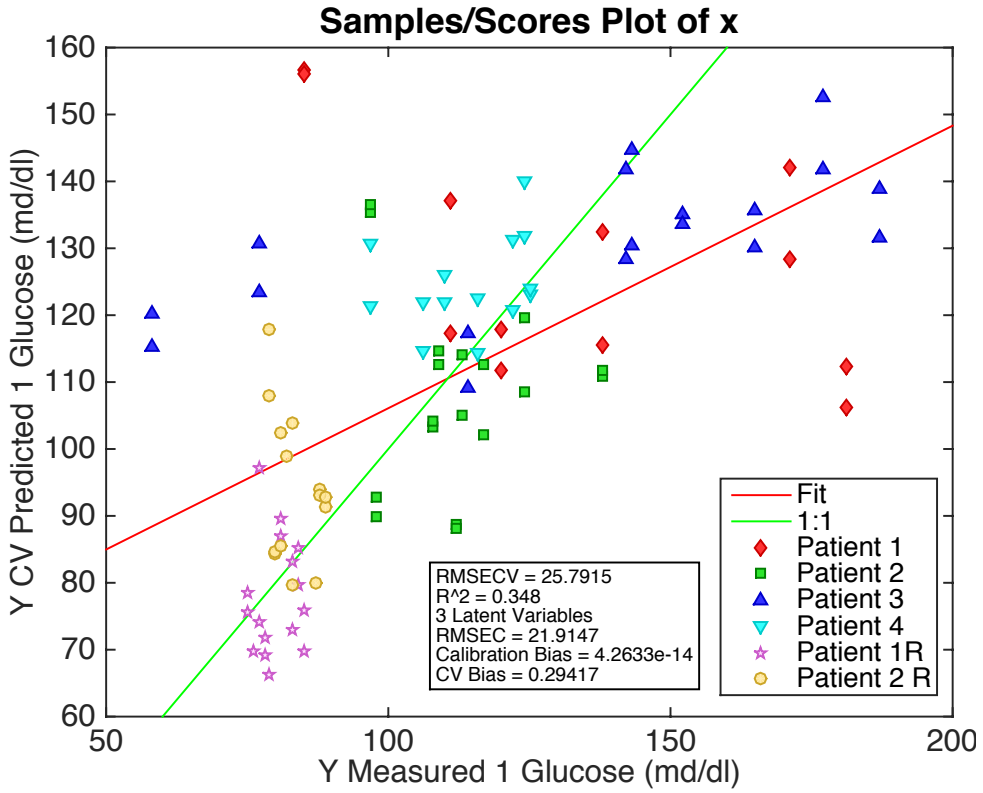
Model Optimizer

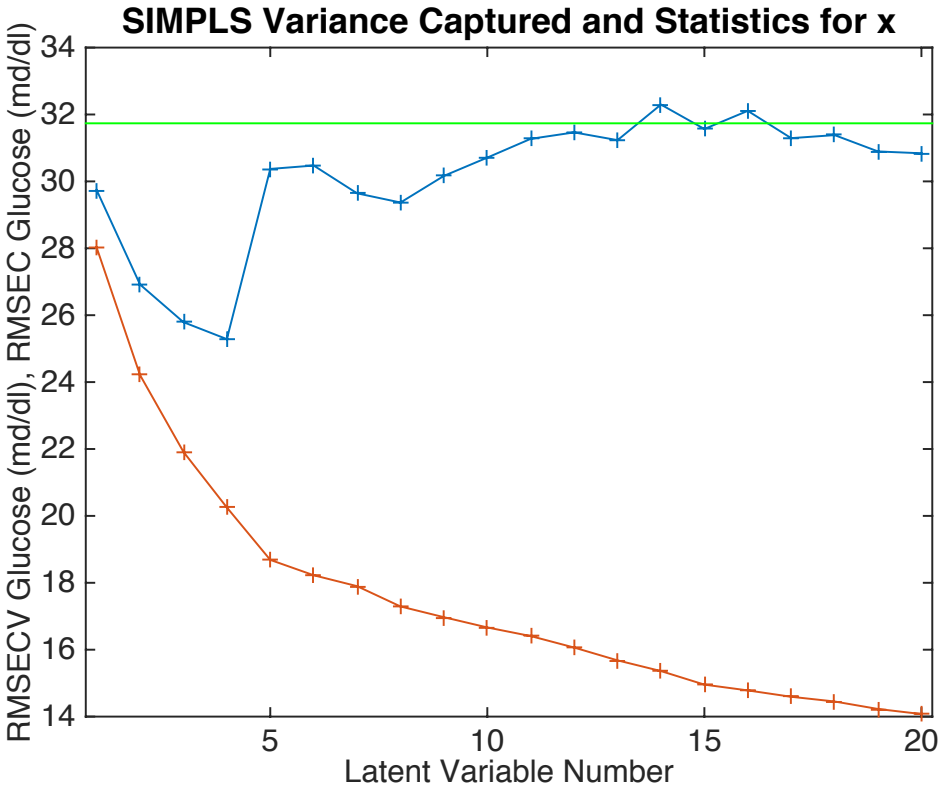


	B...	Model Na...	Nco...	X-Preprocessing	RMSEC (Cal)	RMSECV (CV)	RMSE Ratio...	R2C (Cal)	R2CV (CV)	B.....
1	☑	Model 1	4	1st Derivative (order: 2, wi...	20.25	25.28	1.248	0.5883	0.3796	... ...
2	☑	Model 15	3	2nd Derivative (order: 2, w...	19.78	25.54	1.291	0.6074	0.3667	... ...
3	☑	Model 9	3	1st Derivative (order: 2, wi...	21.91	25.79	1.177	0.5179	0.3482	... ...
4	☑	Model 4	4	2nd Derivative (order: 2, w...	18.51	28.35	1.532	0.6561	0.2809	... ...
5	☑	Model 22	3	Baseline (Specified points) ,...	25.35	28.7	1.133	0.3552	0.2008	... ...
6	☑	Model 20	3	Baseline (Specified points) ,...	24.86	28.95	1.165	0.3799	0.1928	... ...
7	☑	Model 11	3	Mean Center	26.86	29.12	1.084	0.2756	0.1682	... ...
8	☑	Model 17	3	Baseline (Specified points) ,...	23.48	29.48	1.256	0.4466	0.1947	... ...
9	☑	Model 19	5	Baseline (Specified points) ,...	21.22	29.49	1.39	0.5481	0.2163	... ...
10	☑	Model 5	4	Baseline (Specified points) ,...	23.04	29.73	1.29	0.4673	0.1923	... ...
11	☑	Model 3	4	Autoscale	24.8	29.81	1.202	0.3826	0.1687	... ...
12	☑	Model 2	4	Mean Center	24.82	29.83	1.202	0.3815	0.1676	... ...
13	☑	Model 8	4	Baseline (Specified points) ,...	22.87	29.84	1.305	0.4752	0.1871	... ...
14	☑	Model 16	3	Baseline (Specified points) ,...	21.56	29.9	1.387	0.5335	0.2063	... ...
15	☑	Model 7	4	Baseline (Specified points) ,...	21.55	30.2	1.401	0.5336	0.1947	... ...
16	☑	Model 24	5	Mean Center	22.33	30.22	1.353	0.4994	0.1803	... ...
17	☑	Model 12	5	Autoscale	22.4	30.22	1.349	0.4963	0.1798	... ...
18	☑	Model 21	5	Baseline (Specified points) ,...	21.31	30.31	1.422	0.5441	0.1962	... ...
19	☑	Model 10	5	1st Derivative (order: 2, wi...	18.68	30.37	1.626	0.6496	0.2387	... ...
20	☑	Model 23	5	Baseline (Specified points) ,...	20.5	30.44	1.485	0.5782	0.1981	... ...
21	☑	Model 14	3	Autoscale	26.87	30.58	1.138	0.2755	0.1129	... ...
22	☑	Model 6	4	Baseline (Specified points) ,...	20.13	31.34	1.557	0.5931	0.1802	... ...
23	☑	Model 18	5	2nd Derivative (order: 2, w...	16.6	31.83	1.917	0.7234	0.2049	... ...
24	☑	Model 13	5	Baseline (Specified points) ,...	19.23	32.09	1.669	0.6289	0.1656	... ...

EIGENVECTOR RESEARCH INCORPORATED

Best Model Replicate CV

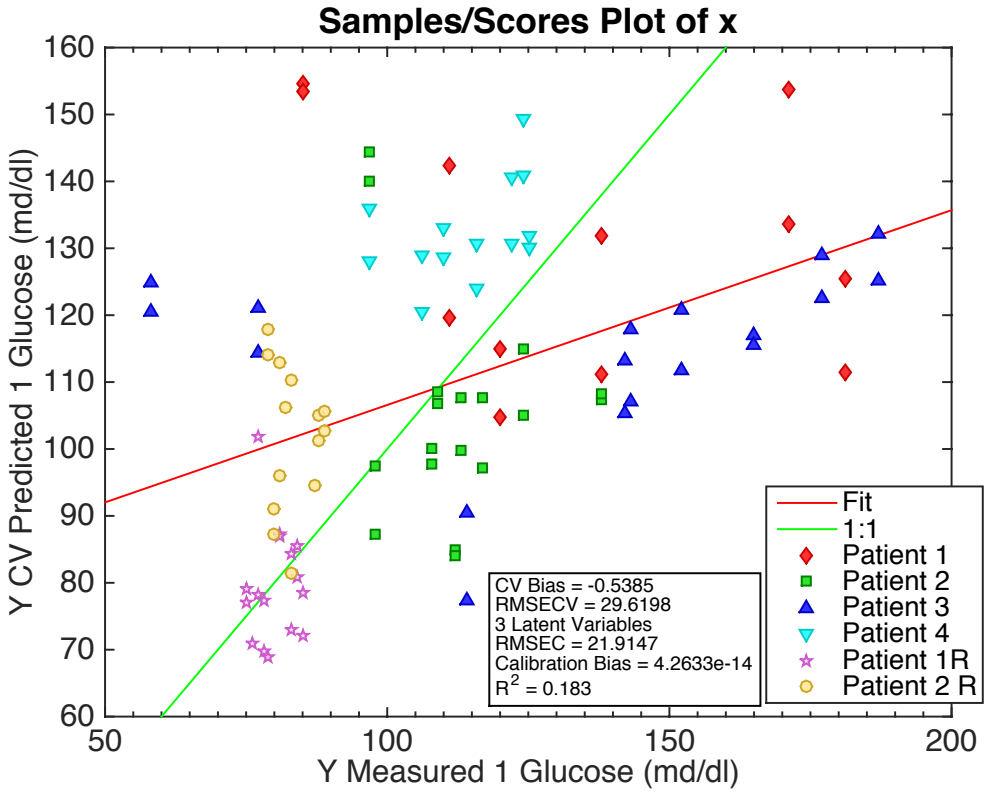


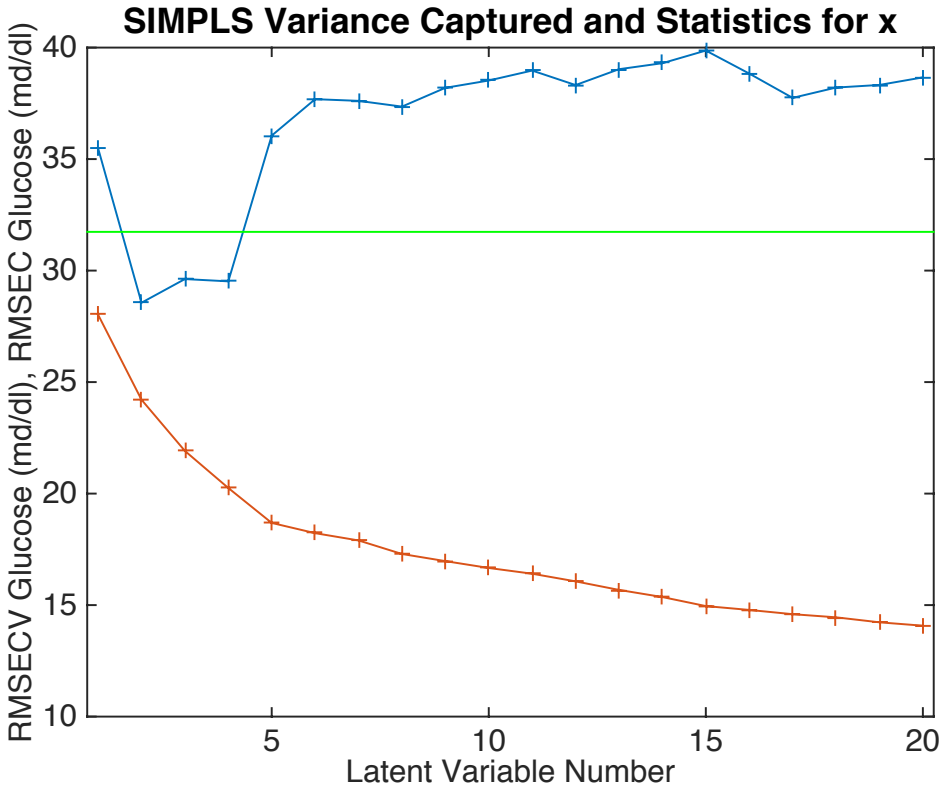


Cross-validation by replicate (46 splits)

EIGENVECTOR RESEARCH INCORPORATED

Best Model Patient CV

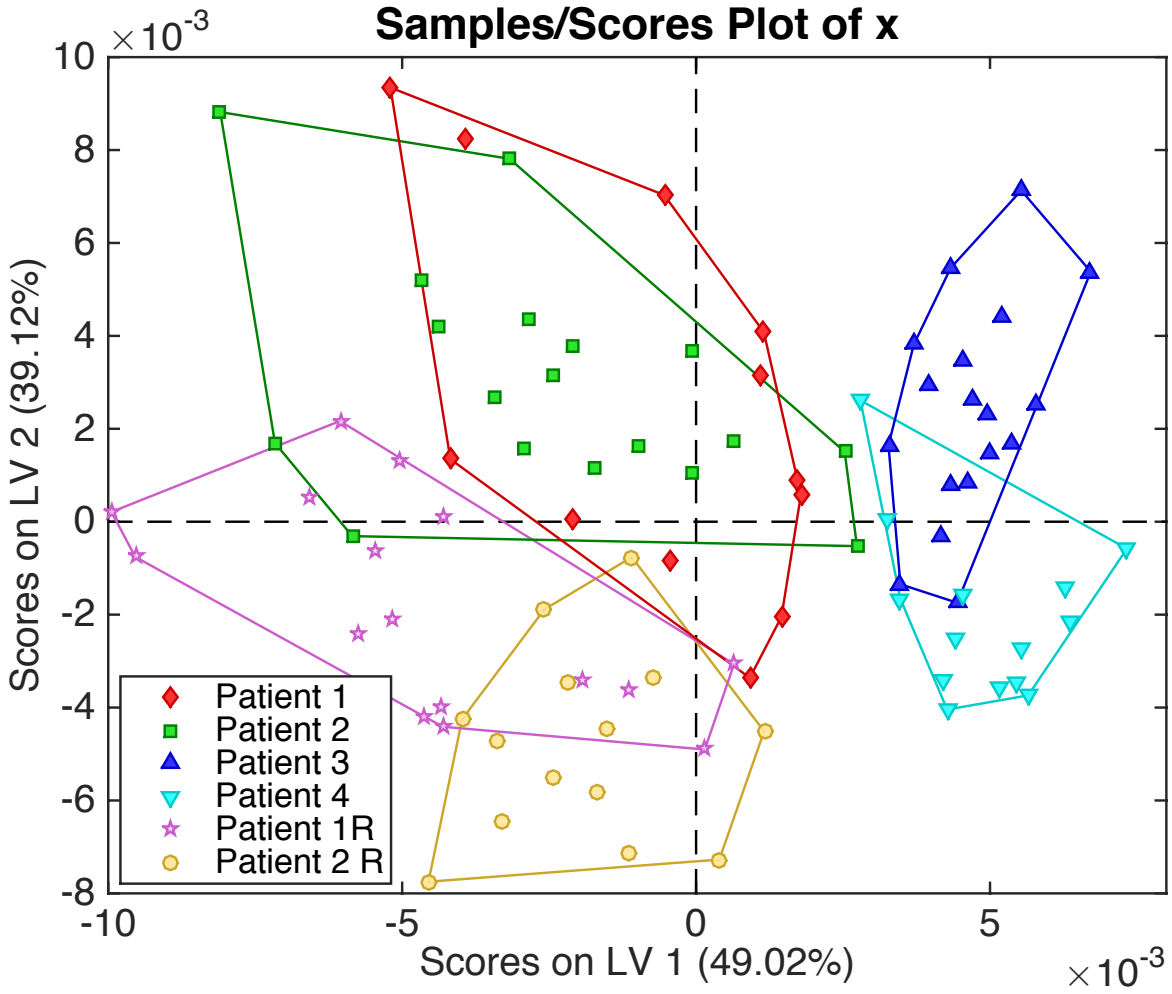




Cross-validation by patient (6 splits)

EIGENVECTOR RESEARCH INCORPORATED

Scores in 1st Derivative Model



EIGENVECTOR RESEARCH INCORPORATED

Conclusions (IMHO)

- Preprocessing and variable selection
  - Can often make a good model better
  - Seldom make a bad model good
- Should be able to see signal in preprocessed data
  - At least in spectroscopy
  - Discrete variables maybe not
- Exhaustive search increases chance of false discovery – don't try too hard?
- Can be difficult to deal with clients
  - At what point do you tell them to give up?
- Still worry about missing something that works!