Data analysis strategies for high dimensional social science data

M3 Conference May 2013

W. Holmes Finch, Maria Hernández Finch,

David E. McIntosh, & Lauren E. Moss

Ball State University

High dimensional data

- High dimensional data refers to the situation in which the number of variables of interest, *p*, is nearly as large as, or larger than the sample size, *N*

- Analysis of high dimensional data is common in genomics, and other areas of medical research

- High dimensional data may also occur in educational and psychological research

- Specialized groups that are of interest, such as individuals with autism, or children of migrant workers may be very rare in the population, and/or difficult to identify

High dimensional data

- High dimensional data causes a number of analytic problems with common statistical models, such as regression

- Such problems include overfitting of the data so that predictions with individuals not in the original sample are not accurate, as well as estimation bias in model coefficients and standard errors

- Given these problems, researchers faced with high dimensional data must find alternative modeling methods

Linear regression (OLS)

- Often researchers use linear regression models to relate one or more independent variables (IVs) to a single dependent variable (DV).

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_j x_{ji} + \varepsilon_i$

- Model parameter estimates are obtained by minimizing the least squares criterion:

$$E^2 = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_j x_{ji}))^2$$
$$E^2 = (y_i - \hat{y}_i)^2$$

OLS

- When *p* approaches *N*, several problems can result:

1. Overfitting of the model to the sample
2. Collinearity among the independent variables
3. Inability to obtain an admissible solution

Ridge Regression (RR)

- RR is designed to overcome the problems of overfitting and collinearity by imposing a penalty on estimates of β_j in a process known as shrinkage.

- Parameters are estimated by minimizing the penalized residual sum of squares function:

- $\left\{\sum_{i=1}^{N}(y_i-\widehat{y_i})^2+\lambda\sum_{j=1}^{P}\beta_j^2\right\}$

RR

- λ is a shrinkage parameter, where larger values are associated with greater shrinkage of the regression coefficients.

- The use of this shrinkage parameter ensures that the likelihood of extreme β_j estimates often associated with highly collinear data is minimized.

- This is an important property in high dimensional data, where collinearity can be a major problem.

RR

- The selection of λ involves the use of jackknife (leave one out) cross validation.

- Fitted values of y_i are obtained for each individual, when they are not included in the model estimation stage, and the squared difference between model predicted and actual values of y_i is calculated.

- The optimal value of λ is the one that results in the minimum mean square error value across all members of the sample.

Least Absolute Shrinkage and Selection Operator (LASSO)

- The LASSO is similar to RR in that it involves shrinkage of model parameter estimates.

- Model parameters are estimated in LASSO by minimizing:

- $\left\{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 + \lambda \sum_{j=1}^{P} |\beta_j|\right\}$

LASSO

- In practice, LASSO adjusts the OLS coefficient estimates by the constant factor λ and frequently behaves like a best subsets regression, selecting variables for inclusion or exclusion in the analysis

- Selection of values for λ is carried out in the same manner as for RR

Partial Least Squares (PLS)

- PLS is a data reduction technique, in which *p* IVs are reduced to *m* linear combinations.
- PLS involves the following steps:

1. Compute φ_{1j} for each IV, x_j. This value is equivalent to β_1 from $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
2. Compute $z_1 = \sum_{j=1}^{J} \varphi_{1j} x_j$
3. Fit model $y_i = \theta_1 z_1$
4. Orthogonalize all x_j with respect to z_1
5. Obtain φ_{2j} and compute $z_2 = \sum_{j=1}^{J} \varphi_{2j} x_j$ such that z_2 is orthogonal to z_1
6. Repeat to obtain z_1 to z_m, all of which are orthogonal

PLS

- In order to determine the number of linear combinations to estimate in PLS, we rely on jackknife cross validation, as described for RR.

- The minimum number of linear combinations, *m*, that reduces the jackknife mean square error across all members of the sample below a given threshold serves as the stopping point for PLS.

Supervised Principal Components (SPC)

- SPC is based on principal components regression, in which the *p* IVs are reduced to a small number of linear combinations that account for most of their variance

- However, unlike principal components regression, which only maximizes accounted for IV variance, SPC creates the linear combinations so as to maximize both IV variance and covariance of the IVs with the DV

SPC

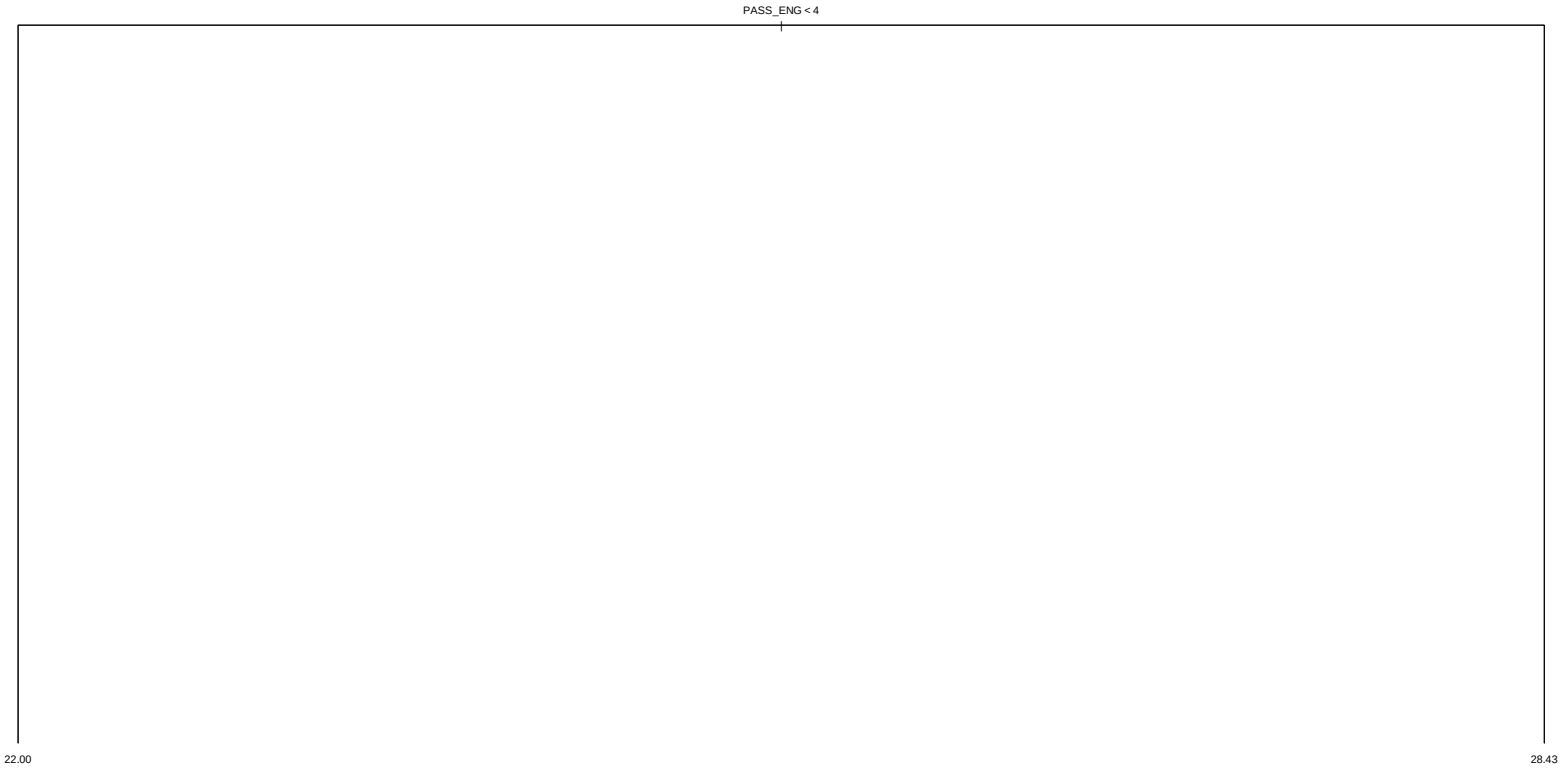
- The SPC algorithm involves the following steps:
  - Estimate $y_i = \beta_0 + \beta_j x_{ji}$ for each IV
  - Compare β_j to a predetermined threshold, *t*
  - Retain x_j only if β_j exceeds *t*
  - Compute first *m* principal components, $z_1, z_2, \ldots, z_m$, using only the retained IVs
  - Use Jackknife cross-validation to determine the optimal values of *t* and *m*
  - Estimate the model $y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \ldots + \beta_j z_{mi}$

Random Forest (RF)

- RF relies on the recursive partitioning algorithm used in classification and regression trees (CART)

- In CART, the data is repeatedly divided into ever smaller groups based on values of the IV, with each division minimizing, at that step, within node heterogeneity in terms of the DV

Sample CART tree



RF

- While CART has been shown to be a very effective tool for both prediction and explanation, it has a tendency to overfit the data.

- One solution to the overfitting problem is to grow a large number of trees from the same sample using the bootstrap technique

RF

- In RF a large number of *B* bootstrap samples with replacement of size *N* is obtained, as is a subset of the predictor variables

- For each bootstrap sample a CART tree is grown, and predicted values obtained for each individual in the sample

- Prediction results for each individual are then averaged across the *B* trees to obtain a single predicted value for the entire RF.

RF

- Variable importance in RF can be measured using a permutation statistic that is calculated as follows:

  1. For each time x_j is used in a split, calculate the mean square error for y_i
  2. Permute the values of x_j randomly so that any relationships with y_i are removed
  3. Calculate the mean square error for y_i for the permuted x_j based on the occasions when it is used
  4. Repeat steps 2 and 3 a large number (e.g. 1000) of times

- Variables with larger values of the permutation statistic are deemed to be more important, as they have a greater impact on prediction accuracy

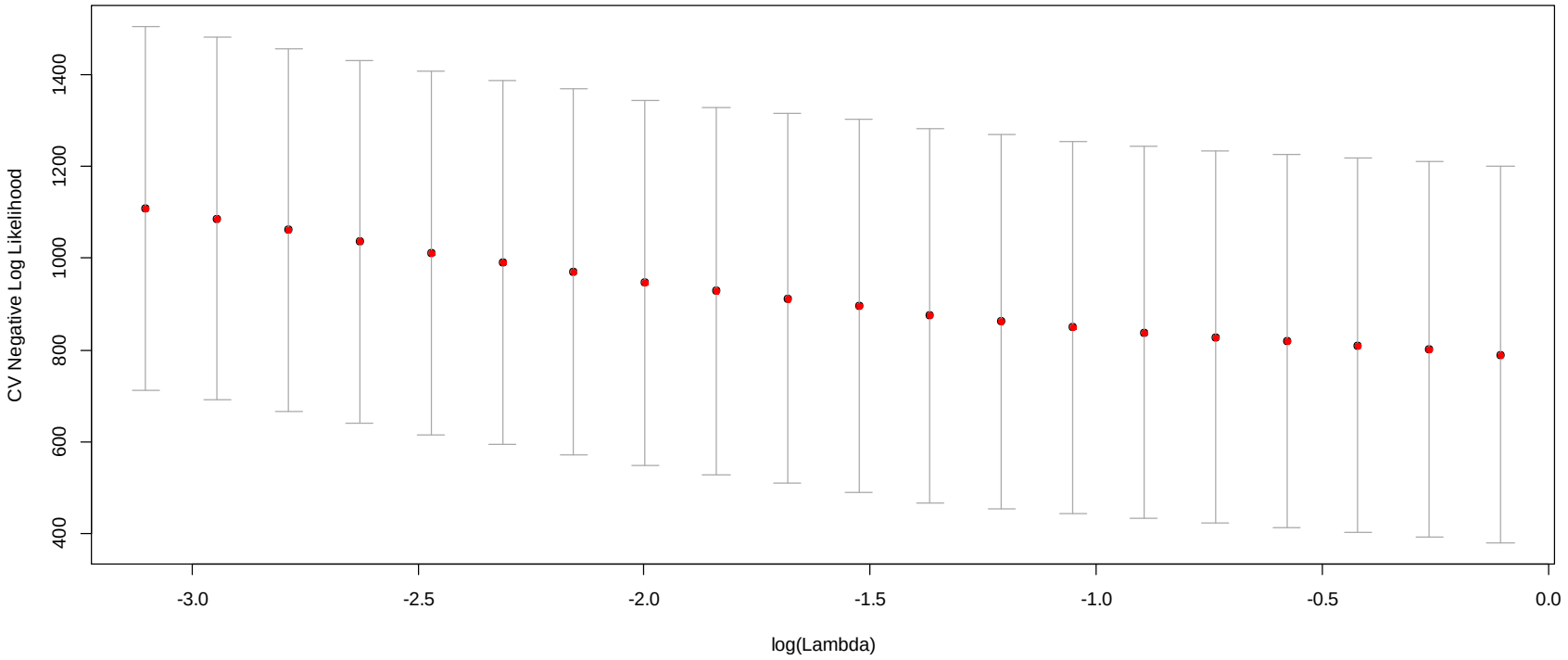
Current data

- Data from 2 studies were analyzed using OLS, and the alternative methods
  - Study 1: 14 Latino migrant students
    - Study focused on relating 6 measures of cognitive functioning, level of English language acquisition, and engagement in learning with an outcome measuring metacognition
  - Study 2: 19 children with Autism
    - Study focused on relating 13 measures of language and reading fluency with a standardized measure of reading aptitude

Migrant worker students

STUDY 1

Results: LASSO

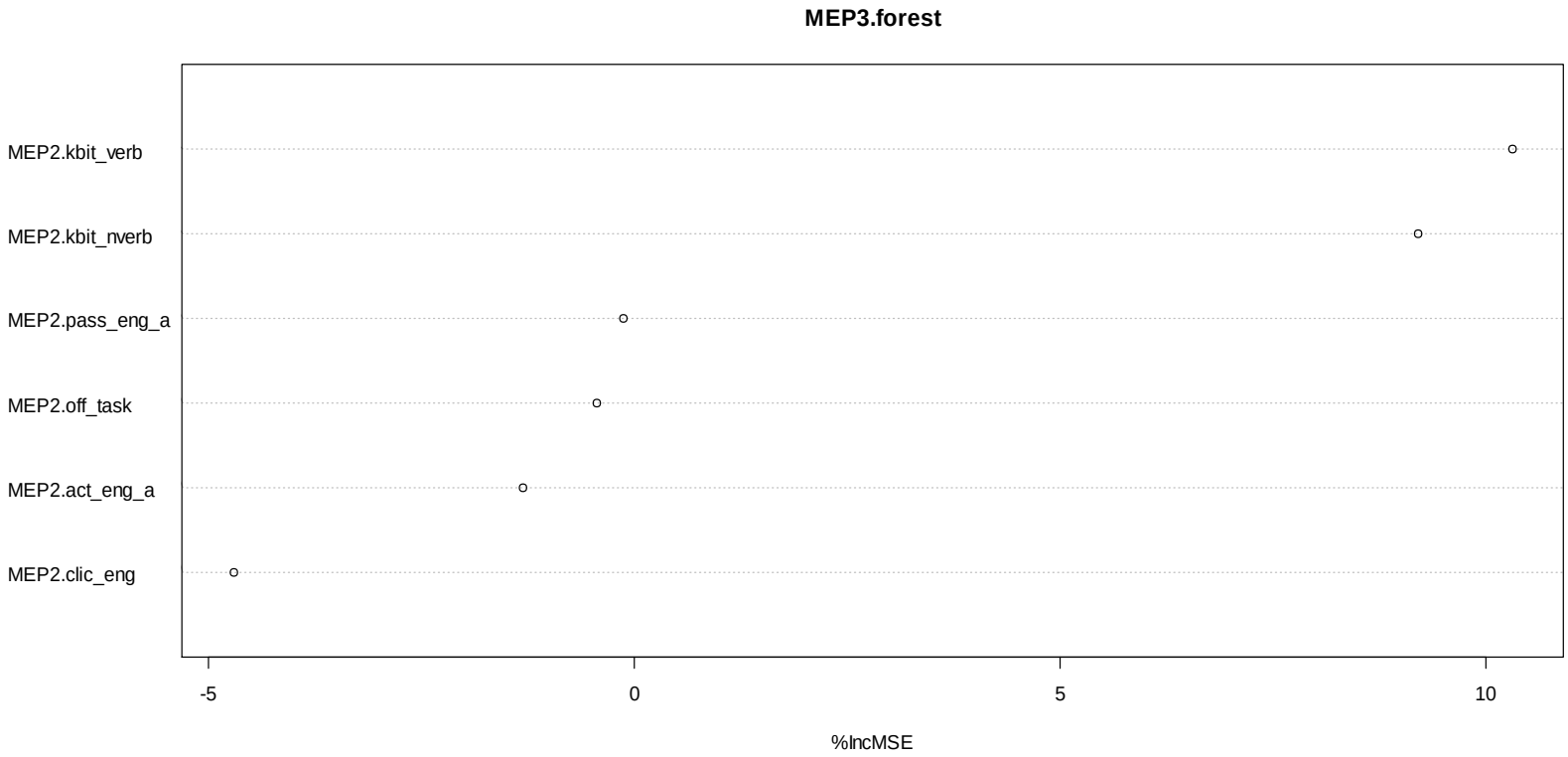


Coefficients

	OLS (*=VIF>5) R^2=0.36		RR R^2=0.36		LASSO
Variable	b / SE	P	b / SE	P	b
KBIT-2 verbal	0.16 / 0.26	0.564	0.09 / 0.15	0.177	0.82
KBIT-2 nonverbal	0.09 / 0.20	0.669	0.08 / 0.11	0.211	0.06
Active Engagement	-0.02 / 0.18	0.935	0.01 / 0.09	0.802	0.01
Passive Engagement	0.02 / 0.09	0.873	0.00 / 0.03	0.929	0
CLIC English	0.23 / 0.26	0.399	**0.15 / 0.08**	**0.024**	0.17
Time off task	-0.01 / 0.06	0.850	-0.01 / 0.09	0.854	-0.01

Random Forest

- Average R^2=0.288



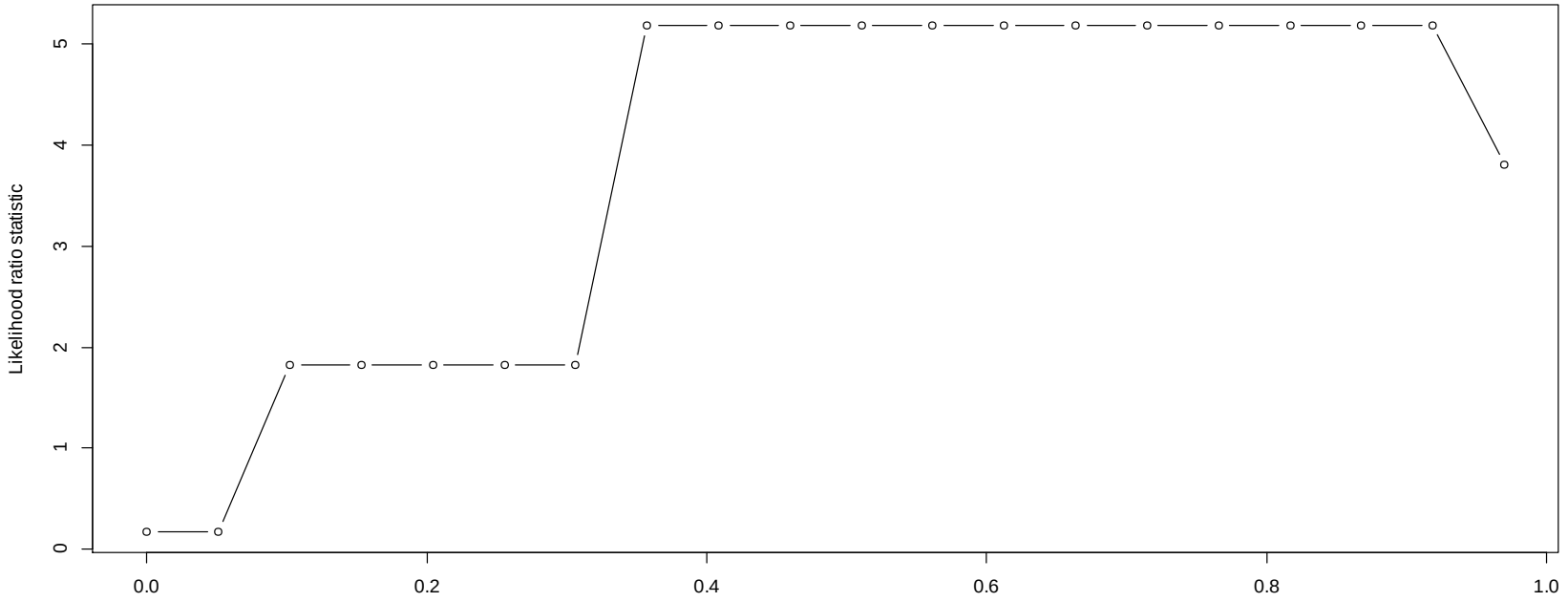
PLS

- One statistically significant linear combination was found (β=3.34, p=0.03)

Variable	PLS Coefficient (φ)
KBIT-2 verbal	1.85
KBIT-2 nonverbal	2.27
Active Engagement	0.28
Passive Engagement	-0.01
CLIC English	2.24
Time off task	-0.35

SPC

- Optimal threshold is near 0.35
- Selected variables must have coefficients > 0.35



SPC

- First principal component was significantly (p=0.055) related to metacognition score (KMA reading)

- Coefficient = 0.30

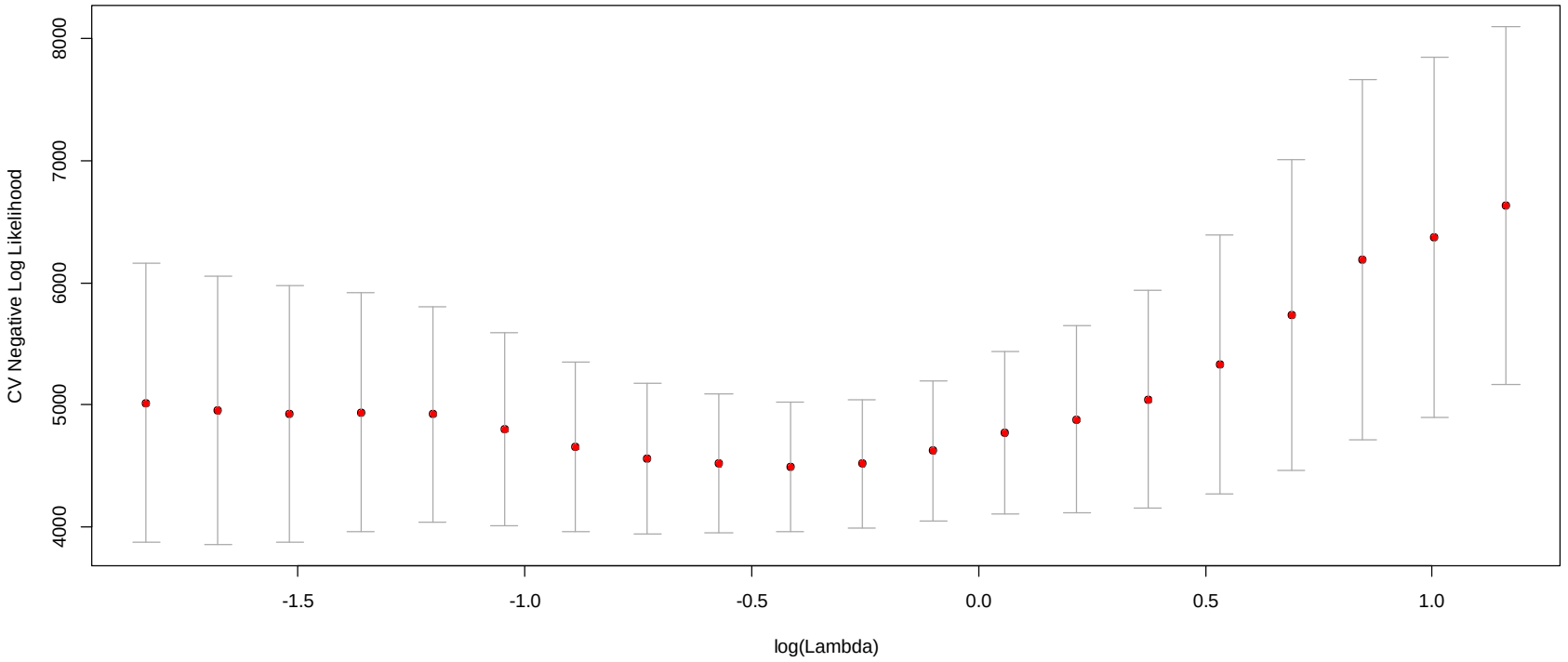
SPC feature scores and importance

Variable	Feature score (Coefficient)	Importance	PLS Coefficient (φ)
KBIT-2 verbal	**0.97**	**2957.64**	1.85
KBIT-2 nonverbal	**0.93**	**2705.19**	2.27
Active Engagement	**1.00**	**604.55**	0.28
Passive Engagement	0.33	NA	-0.01
CLIC English	0.06	NA	2.24
Time off task	0.06	NA	-0.35

Children with autism

STUDY 2

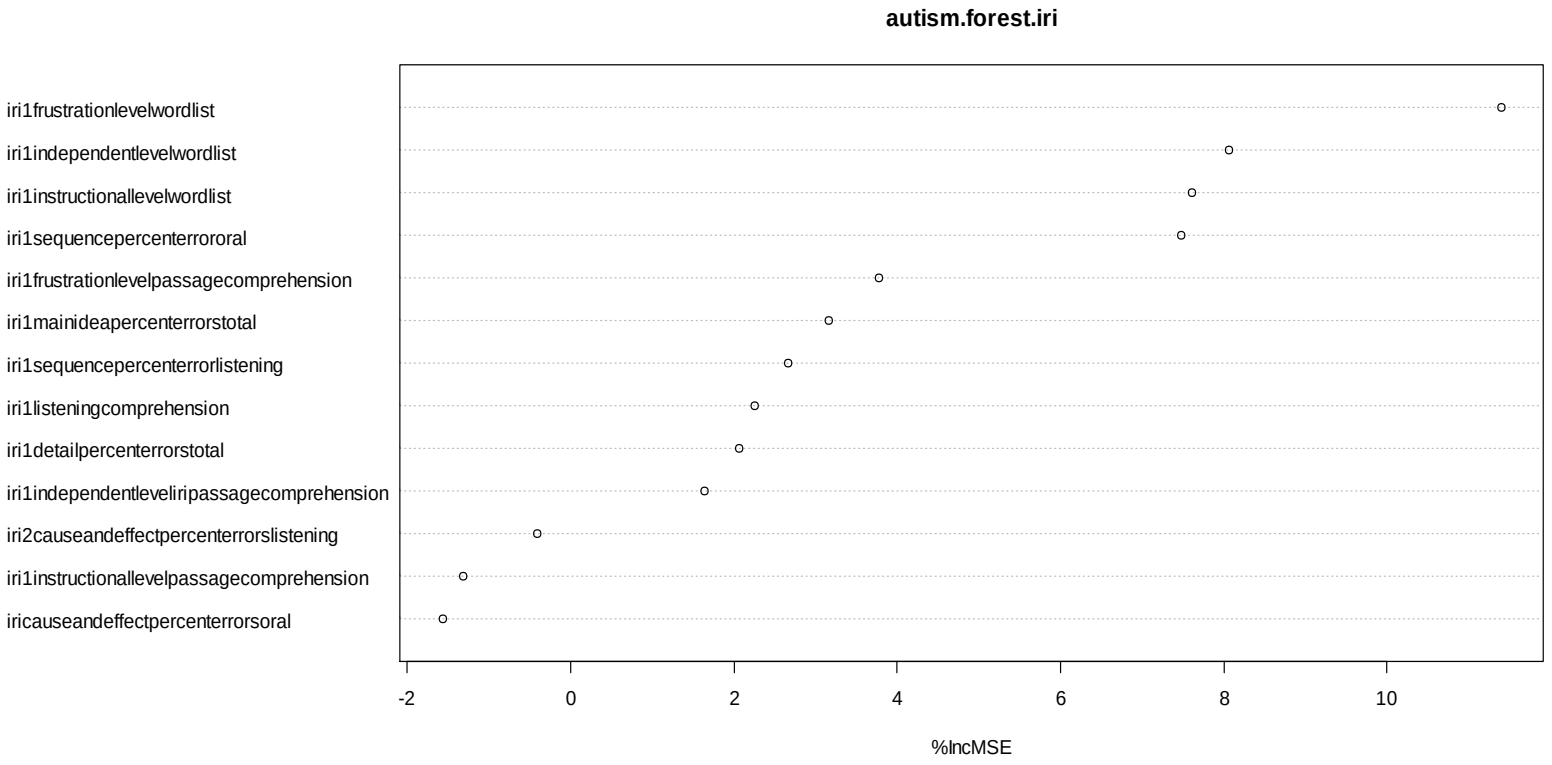
Results: LASSO



	OLS (*=VIF>5) R^2=0.86		RR R^2=0.69		LASSO
Variable	b / SE	P	b / SE	P	b
Word independent	5.9* / 7.9	0.487	**21.7 / 11.9**	**0.068**	15.1
Word instructional	-6.1* / 5.3	0.301	18.2 / 11.8	0.122	9.1
Word frustration	9.5* / 6.3	0.193	**22.0 / 12.1**	**0.068**	28.6
Passage independent	-9.7* / 21.6	0.673	-10.5 / 10.1	0.298	0
Passage instructional	2.1* / 29.1	0.673	-11.0 / 8.4	0.189	0
Passage frustration	40.0* / 31.9	0.266	14.1 / 9.8	0.150	1.4
Listening	-34.7* / 35.2	0.370	3.0 / 8.9	0.739	0
Main idea	0.0* / 0.2	0.993	-2.7 / 13.8	0.846	0
Details	0.1* / 0.4	0.826	-11.8 / 13.1	0.365	0
Sequence oral	0.1* / 0.2	0.674	-9.6 / 13.6	0.481	-4.6
Sequence listening	0.3 / 0.2	0.250	**24.6 / 13.4**	**0.065**	11.3
Cause/effect oral	0.2 / 0.2	0.329	8.7 / 13.2	0.509	0
Cause/effect listening	0.0* / 0.3	0.991	-9.1 / 13.2	0.491	0

Random Forest

- Mean R^2=0.227



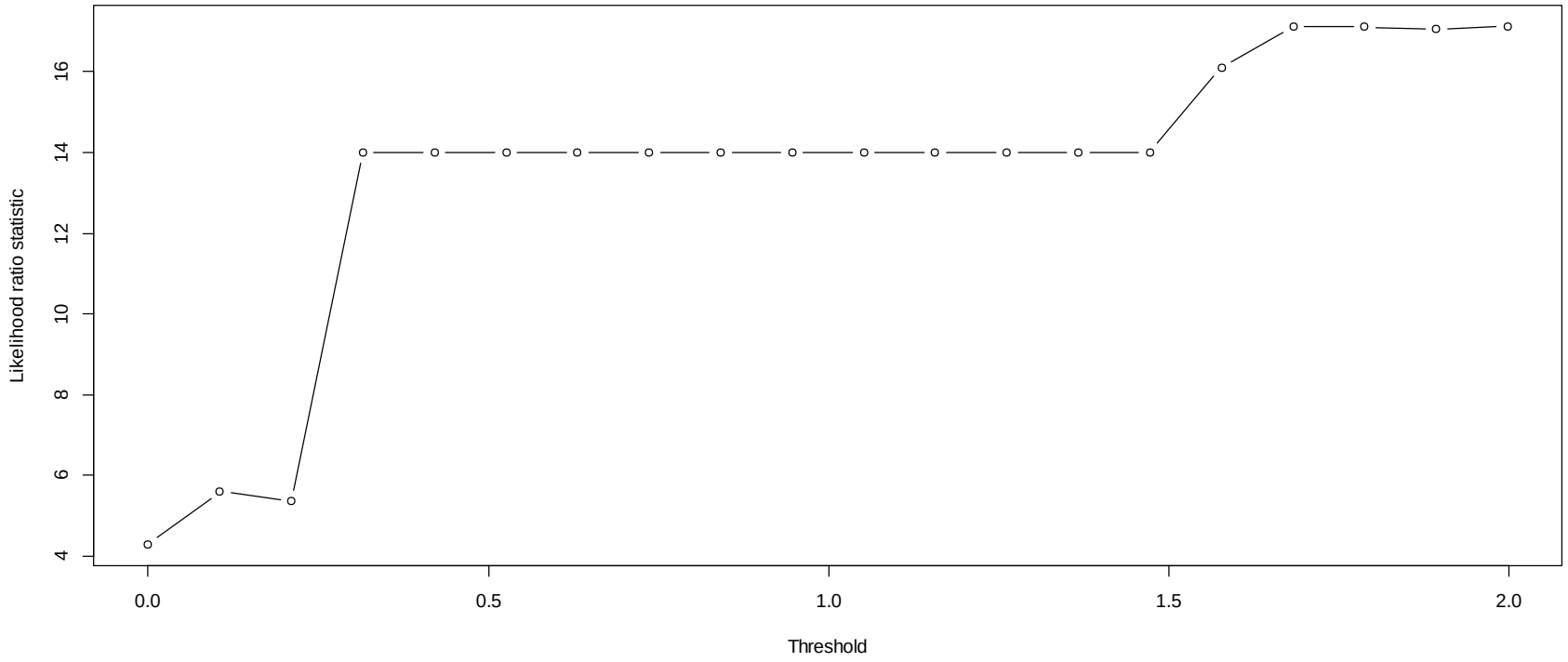
PLS

One statistically significant linear combination was found (b=5.53, *p*=0.002)

Variable	PLS Coefficient (φ)
Word independent	6.82
Word instructional	6.91
Word frustration	8.96
Passage independent	-3.75
Passage instructional	-5.74
Passage frustration	0.96
Listening	-1.65
Main idea	-1.44
Details	-1.41
Sequence oral	-3.42
Sequence listening	3.74
Cause/effect oral	-0.00
Cause/effect listening	-2.65

SPC

- Optimal threshold is near 0.2
- Selected variables must have coefficients > 0.2



SPC

- First principal component was significantly (p=0.033) related to reading aptitude

- Coefficient = 0.40

SPC feature scores and importance

Variable	Feature score (Coefficient)	Importance	PLS Coefficient (φ)
Word independent	**2.17**	**829.41**	6.82
Word instructional	**1.95**	**663.81**	6.91
Word frustration	**2.00**	**1202.84**	8.96
Passage independent	**1.54**	**451.44**	-3.75
Passage instructional	**1.57**	**598.77**	-5.74
Passage frustration	**1.85**	**811.85**	0.96
Listening	**1.62**	**885.11**	-1.65
Main idea	**-0.21**	**188.94**	-1.44
Details	**-0.23**	**158.74**	-1.41
Sequence oral	**-0.26**	**109.92**	-3.42
Sequence listening	0.07	NA	3.74
Cause/effect oral	-0.08	NA	-0.00
Cause/effect listening	-0.19	NA	-2.65

Conclusions

- There are several alternatives available for researchers using high dimensional data
- Some approaches are variants of standard OLS (RR, LASSO) that employ shrinkage parameters
- Other methods involve data reduction (PLS, SPC)
- Yet another is nonparametric in nature (RF)

Conclusions

- Results provided by the various methods may differ substantially
- Researchers should select the method that best addresses their research problems
- When the predictors represent different facets of a common construct, data reduction methods such as SPC and PLS may be optimal
- When collinearity is a particular problem, the researcher may select RR or LASSO
- If it is believed that the relationships between predictors and the outcome are nonlinear, RF may be the best choice