Principal Components Regression with Implicit Cross-Validation

Neal B. Gallagher
Eigenvector Research, Inc.
nealg@eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Ridge and Principal Components Regression
- PCR with Implicit Cross-Validation
- Results for Caustic
- Results for Liquid Fed Ceramic Melter
- Conclusions

EIGENVECTOR RESEARCH INCORPORATED

Why Implicit Cross-Validation?

- The objective is to find a way to automate model identification for principal components regression (PCR) and other factor-based regression models
  - Allow multiple models to be tested near their optimum
  - Automate window-based approaches that rely on regression models (e.g. piece-wise standardization)
  - Give novices a good first choice
- How to automatically choose the number of factors in PCR?

EIGENVECTOR RESEARCH INCORPORATED

Ridge Regression

$$\mathbf{Xb} = \mathbf{y}$$

- $\mathbf{X}$ MxN predictor block
- $\mathbf{y}$ Mx1 predicted vector
- $\mathbf{b}$ Nx1 regression vector
- θ scalar ridge parameter

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X} + \theta\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- singular value decomposition of $\mathbf{X}$

$$\mathbf{X} = \mathbf{USV}^T$$

$$\hat{\mathbf{b}} = \mathbf{V}\left(\mathbf{S}^T\mathbf{S} + \theta\mathbf{I}\right)^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y}$$

EIGENVECTOR RESEARCH INCORPORATED

Modify RR

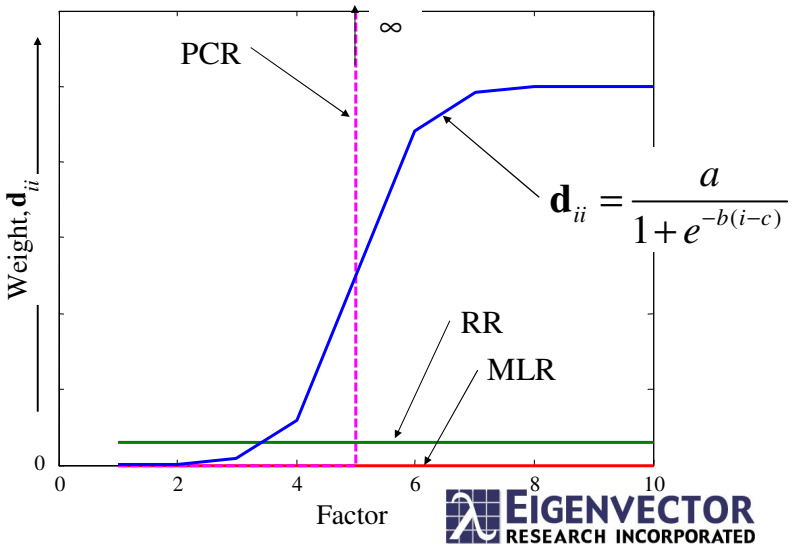
- substitute a diagonal matrix $\mathbf{D}^2$ for $\theta\mathbf{I}$
- gives PCR for
  - $\mathbf{d}_{ii} = 0$ for $i = 1,...,K$
  - $\mathbf{d}_{ii} = \infty$ for $i = K+1,...,N$
- propose that $\mathbf{d}_{ii}$ is from a sigmoid

$$\hat{\mathbf{b}} = \mathbf{V}\left(\mathbf{S}^T\mathbf{S} + \mathbf{D}^2\right)^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y}$$

$$\mathbf{d}_{ii} = \frac{a}{1+e^{-b(i-c)}}$$

EIGENVECTOR RESEARCH INCORPORATED

Regression Summary



EIGENVECTOR RESEARCH INCORPORATED

Estimation of a, b, and c

- Cross-validation
  - divide data into j=1,...,J subsets
  - $\mathbf{I}_{c,j}$: 1 for samples included in calibration, 0 for left for test
  - $\mathbf{I}_{p,j}$=**I**- $\mathbf{I}_{c,j}$ selects test samples
  - minimize $O(a,b,c)$ squared cross-validation error

$$\hat{\mathbf{b}}_j = \mathbf{V}\left(\left(\mathbf{S}+\mathbf{D}\right)^T \mathbf{U}^T\mathbf{I}_{c,j}\mathbf{U}\left(\mathbf{S}+\mathbf{D}\right)\right)^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{I}_{c,j}\mathbf{y}$$

$$O(a,b,c) = \sum_{j=1}^{J}\left(\mathbf{X}\hat{\mathbf{b}}_j - \mathbf{y}\right)^T \mathbf{I}_{p,j}\left(\mathbf{X}\hat{\mathbf{b}}_j - \mathbf{y}\right)$$

EIGENVECTOR RESEARCH INCORPORATED

Caustic Data

- IR of NaCl, NaOH, varying T
  - designed experiment w/ 95 samples
    - 71 for calibration and cross-validation, 24 for test set
  - 7382-9696 cm^{-1}, 2nd derivative, mean-center
  - estimate NaCl wt% from 2nd derivative spectra
  - Caustic data courtesy M.B. Seasholtz, The DOW Chemical Company

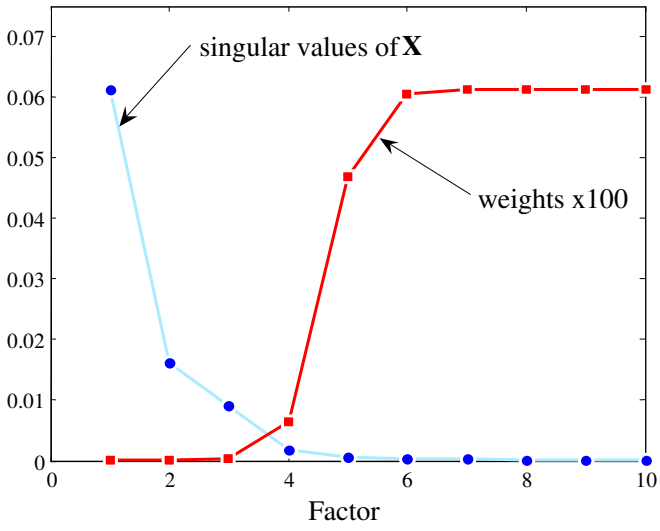
EIGENVECTOR RESEARCH INCORPORATED

Results for Caustic Data (NaCl)

	RMSEC	RMSECV	RMSEP	#Factors
PCR	0.058	0.060	0.072	4 PCs
PCR	0.064	0.066	0.073	3 PCs
PLS	0.057	0.060	0.072	4 LVs
PLS	0.063	0.065	0.072	3 LVs
PCR	0.056	0.058	0.070	ICV*

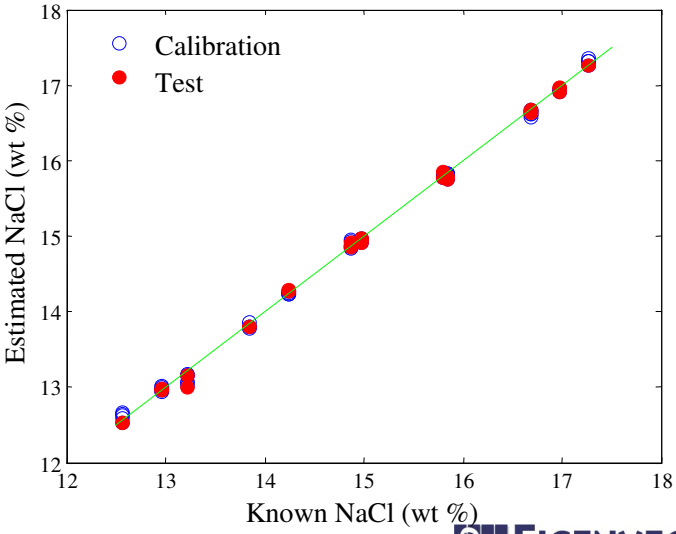
*a = 0.00061; b = 3.33; c = 4.64

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

results for the caustic data set using PCR-ICV



EIGENVECTOR RESEARCH INCORPORATED

LFCM Process Data

- Liquid Fed Ceramic Melter Data
  - 20 temperatures in 2 thermocouple wells and 1 level
    - 295 for calibration and cross-validation, 200 for test set
  - mean-center
  - estimate level from temperatures
  - data in PLS_Toolbox [plsdata]

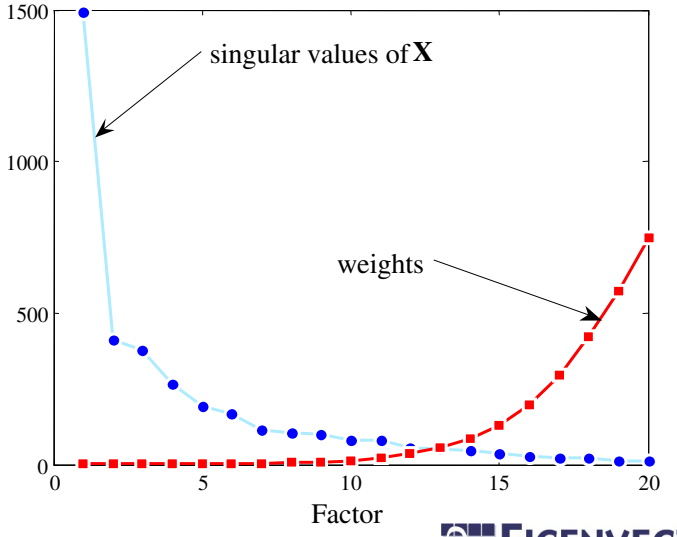
EIGENVECTOR RESEARCH INCORPORATED

Results for LFCM Data

	RMSEC	RMSECV	RMSEP	#Factors
MLR	0.099	0.112	0.150	20 (all)
RR	0.100	0.112	0.147	θ=0.0074
PCR	0.106	0.111	0.137	6 PCs
PLS	0.103	0.110	0.140	3 LVs
PCR	0.101	0.108	0.147	ICV*

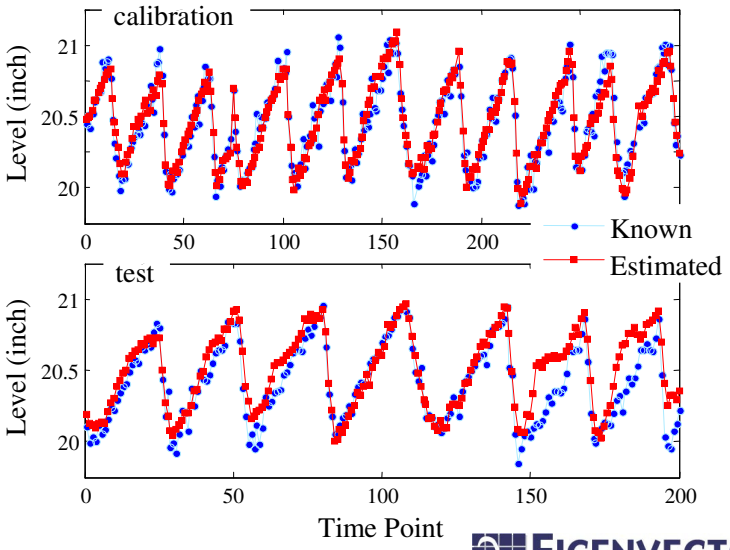
*a = 1493; b = 0.62; c = 18

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

results for LFCM data set using PCR-ICV



EIGENVECTOR RESEARCH INCORPORATED

Conclusions

- Implicit cross-validation was used to select parameters of sigmoid weighting function
- Automates a PCR-like regression
  - extend to other regressions?
- Lot's of bookkeeping for cross-validation
- Additional work
  - identify globally useful bounds on sigmoid parameters
  - penalty on "number of factors" parameter c

EIGENVECTOR RESEARCH INCORPORATED

Extra Slides

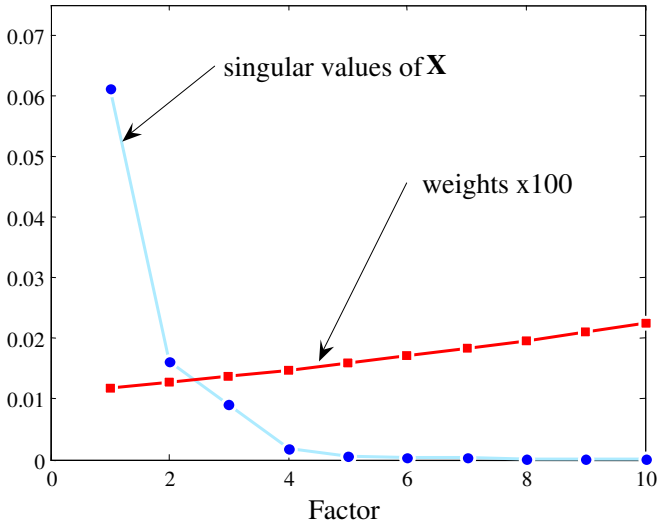
EIGENVECTOR RESEARCH INCORPORATED

Results for Caustic Data (NaOH)

	RMSEC	RMSECV	RMSEP	#Factors
PCR	0.036	0.040	0.042	4 PCs
PCR	0.070	0.073	0.087	3 PCs
PLS	0.035	0.039	0.041	4 LVs
PLS	0.067	0.070	0.084	3 LVs
PCR	0.032	0.045	0.034	ICV*

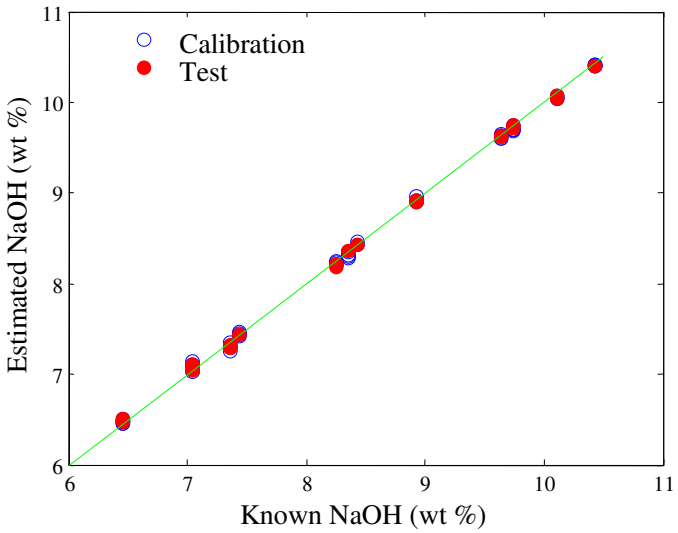
*a = 0.00061; b = 0.10; c = 15.4

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

results for the caustic data set using PCR-ICV



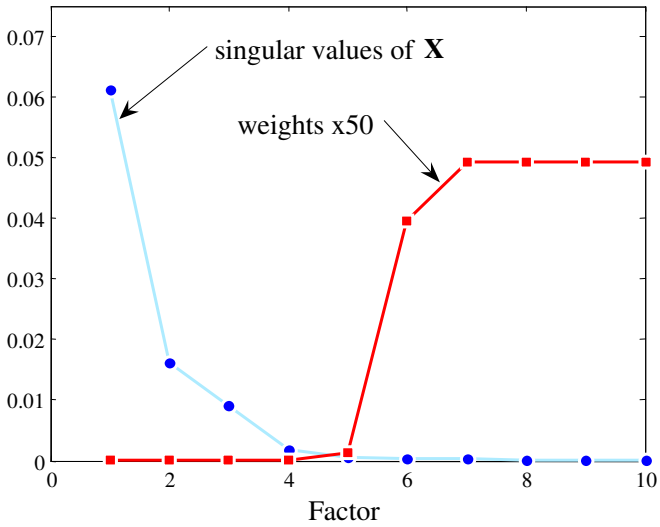
EIGENVECTOR RESEARCH INCORPORATED

Results for Caustic Data (Temp)

	RMSEC	RMSECV	RMSEP	#Factors
PCR	3.75	3.98	2.75	4 PCs
PCR	4.08	4.24	3.60	3 PCs
PLS	3.73	3.97	2.71	4 LVs
PLS	4.05	4.22	3.55	3 LVs
PCR	3.68	3.89	2.61	ICV*

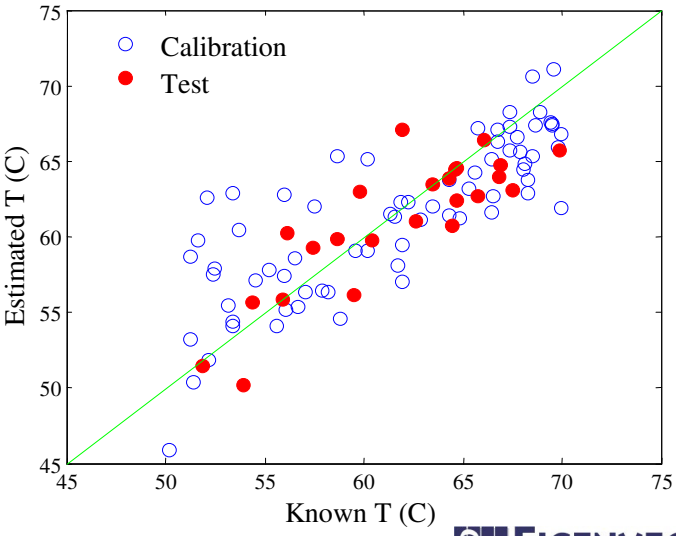
*a = 0.00099; b = 5; c = 5.72

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

results for the caustic data set using PCR-ICV



EIGENVECTOR RESEARCH INCORPORATED