Re-fitting PCA, MPCA and PARAFAC Models to Incomplete Data Records

Barry M. Wise
EIGENVECTOR RESEARCH, INC.

Abstract

Many processes and analytical methods generate multivariate or multiway data sequentially. In calibration mode this is not generally a problem, one just waits until all the data is in, then sets about modeling. Often, particularly in process applications, it is desired to know how well the model represents the incoming data before the complete record is available. Several options have been proposed to deal with this problem. Some of the methods are based on in-filling the missing data so that the models may be applied in the usual way. This approach, however, suffers from all the problems associated with missing data. How does one fill in the record, particularly when the missing parts are systematic, not random? An alternative approach is considered here. Models are fit to partial data records by simply truncating the model loadings to coincide with the available data and fitting the partial factors using a classical least squares (CLS) approach. The estimated scores and residuals are found to converge to those of the entire record quite rapidly in the data sets considered. In fault detection applications the implies that it is often possible to detect a bad batch well before its completion. The partial refit method is compared to the method for in-filling missing data in PCA developed previously. The methods are found to be mathematically identical.

Why Partial Refitting?

- PCA/MPCA and CLS/PARAFAC models generally developed on complete data records
- Sometimes want to compare incoming data to model before complete record is available
- Batch process monitoring most common
- Where is process going to wind up?
- Can faults and off-normal batches be detected mid batch?

Approaches to Partial Records

- ~~Fill record with data mean~~
- ~~Propagate current deviation~~
- Solve for missing variables-complete the squares
- Solve for scores with truncated Classical Least Squares

Truncated CLS Approach

- Calculate scores by fitting truncated loadings to available data
- Normally done by projection, but loadings not orthogonal after truncation
- Use Classical Least Squares approach

CLS Approach to PCA Replacement

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_m & \mathbf{x}_g \end{bmatrix} \qquad \mathbf{P}_k = \begin{bmatrix} \mathbf{P}_m \\ \mathbf{P}_g \end{bmatrix}$$

$$\hat{\mathbf{t}}_i = \mathbf{x}_g \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \mathbf{P}_g$$

$$\hat{\mathbf{x}}_m = \hat{\mathbf{t}}_i \mathbf{P}_m^T = \mathbf{x}_g \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \mathbf{P}_g \mathbf{P}_m^T$$

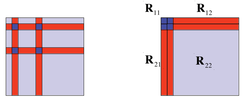
PP_{OLS}

P. Nomikos and J.F. MacGregor, "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, 1995, **37**(1), 41-58.
S. Garcia-Munoz, T. Kourti and J.F. MacGregor, "Model Predictive Monitoring for Batch Processes," *Ind. Eng. Chem. Res.*, 2004, 43, 5929-5941

Missing Variables in PCA

$$\mathbf{X} = \mathbf{T}_k \mathbf{P}_k^T + \mathbf{E}$$

$$Q_i = \mathbf{x}_i \left(\mathbf{I} - \mathbf{P}_k \mathbf{P}_k^T \right) \mathbf{x}_i^T = \mathbf{x}_i \mathbf{R} \mathbf{x}_i^T$$



Partition Residual Q

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_m & \mathbf{x}_g \end{bmatrix} \qquad \mathbf{P}_k = \begin{bmatrix} \mathbf{P}_m \\ \mathbf{P}_g \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$

$$Q_i = \begin{bmatrix} \mathbf{x}_m & \mathbf{x}_g \end{bmatrix} \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_m^T \\ \mathbf{x}_g^T \end{bmatrix}$$

Minimize Q

$$Q_i = \mathbf{x}_m \mathbf{R}_{11} \mathbf{x}_m^T + \mathbf{x}_g \mathbf{R}_{21} \mathbf{x}_m^T + \mathbf{x}_m \mathbf{R}_{21}^T \mathbf{x}_g^T + \mathbf{x}_g \mathbf{R}_{22} \mathbf{x}_g^T$$

$$\hat{\mathbf{x}}_m \forall \left\{ \mathbf{x}_m \mathbf{R}_{11} \mathbf{x}_m^T + \mathbf{x}_g \mathbf{R}_{21} \mathbf{x}_m^T + \mathbf{x}_m \mathbf{R}_{21}^T \mathbf{x}_g^T \right\} = \min$$

$$\hat{\mathbf{x}}_m = -\mathbf{x}_g \mathbf{R}_{21} \mathbf{R}_{11}^{-1} = \mathbf{x}_g \mathbf{P}_g \mathbf{P}_m^T \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)$$

$$\mathbf{R}_R = -\mathbf{R}_{21} \mathbf{R}_{11}^{-1}$$

Effect of Replacement

$$\mathbf{e} = \begin{bmatrix} \mathbf{e}_m & \mathbf{e}_g \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_m \mathbf{R}_{11} + \mathbf{x}_g \mathbf{R}_{21} & \mathbf{x}_m \mathbf{R}_{21}^T + \mathbf{x}_g \mathbf{R}_{22} \end{bmatrix}$$

Residuals on replaced variables are zero

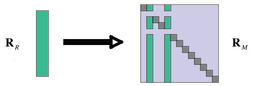
$$\mathbf{e}_m = \hat{\mathbf{x}}_m \mathbf{R}_{11} + \mathbf{x}_g \mathbf{R}_{21} = -\mathbf{x}_g \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{11} + \mathbf{x}_g \mathbf{R}_{21} = 0$$

$$Q_{\min} = -\mathbf{x}_g \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{21}^T \mathbf{x}_g^T + \mathbf{x}_g \mathbf{R}_{22} \mathbf{x}_g^T$$

$$\hat{\mathbf{t}}_i = \begin{bmatrix} \hat{\mathbf{x}}_m & \mathbf{x}_g \end{bmatrix} \begin{bmatrix} \mathbf{P}_m \\ \mathbf{P}_g \end{bmatrix} = \begin{bmatrix} \mathbf{x}_g \mathbf{R}_R & \mathbf{x}_g \end{bmatrix} \begin{bmatrix} \mathbf{P}_m \\ \mathbf{P}_g \end{bmatrix}$$

Replacement Matrix

- Map $\mathbf{R}_R$ into matrix $\mathbf{R}_M$ which replaces variables with estimate most consistent with PCA model
- Most challenging computation is inverse of $\mathbf{R}_{11}$



B. M. Wise and N. L. Ricker, "Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity," *IFAC Symposium on Advanced Control of Chemical Processes*, pps. 125-130, Toulouse, France, October 1991

Basis for PLS_Toolbox missing data functions and PCA cross-validation

Method of Wise and Ricker (1991) and Nomikos and MacGregor (1995) are equivalent!

Proof

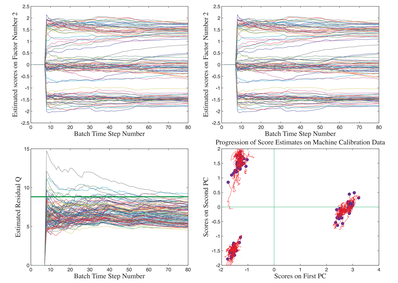
Note: $\mathbf{P}_m^T \mathbf{P}_m + \mathbf{P}_g^T \mathbf{P}_g = \mathbf{I} \rightarrow \mathbf{P}_m^T \mathbf{P}_m = \mathbf{I} - \mathbf{P}_g^T \mathbf{P}_g$

$$\begin{aligned}
\left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \mathbf{P}_g \mathbf{P}_m^T &= \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \mathbf{P}_g \mathbf{P}_m^T \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right) \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)^{-1} \\
&= \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \left[\mathbf{P}_g \mathbf{P}_m^T - \mathbf{P}_g \mathbf{P}_m^T \mathbf{P}_m \mathbf{P}_m^T \right] \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)^{-1} \\
&= \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \left[\mathbf{P}_g \mathbf{P}_m^T - \mathbf{P}_g \left(\mathbf{I} - \mathbf{P}_g^T \mathbf{P}_g \right) \mathbf{P}_m^T \right] \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)^{-1} \\
&= \left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \left[\mathbf{P}_g \mathbf{P}_m^T - \mathbf{P}_g \mathbf{P}_m^T + \mathbf{P}_g \mathbf{P}_g^T \mathbf{P}_g \mathbf{P}_m^T \right] \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)^{-1} \\
\left(\mathbf{P}_g \mathbf{P}_g^T \right)^{-1} \mathbf{P}_g \mathbf{P}_m^T &= \mathbf{P}_g \mathbf{P}_m^T \left(\mathbf{I} - \mathbf{P}_m \mathbf{P}_m^T \right)^{-1}
\end{aligned}$$

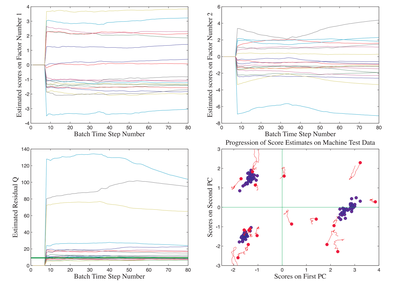
Thanks to Tamara G. Kolda!

EIGENVECTOR RESEARCH, INC.
3905 West Eaglerock Drive
Wenatchee, WA 98801
(509)662-9213
bmw@eigenvector.com
www.eigenvector.com

MPCA on Etch Calibration



MPCA on Etch Test



Refitting PARAFAC Models

- Refitting PARAFAC model with fixed loadings in all but one mode is a single CLS step
- Loadings of fixed factors multiplied out and unfolded
- Unfolded loadings fit to data unfolded in sample mode

Residuals in a CLS Fit

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{c}} \mathbf{S}^T$$

$$\mathbf{e} = \mathbf{x} - \mathbf{x} \mathbf{S} \left(\mathbf{S}^T \mathbf{S} \right)^{-1} \mathbf{S}^T$$

$$\mathbf{e} = \mathbf{x} \left(\mathbf{I} - \mathbf{S} \left(\mathbf{S}^T \mathbf{S} \right)^{-1} \mathbf{S}^T \right)$$

$$Q = \mathbf{x} \left(\mathbf{I} - \mathbf{S} \left(\mathbf{S}^T \mathbf{S} \right)^{-1} \mathbf{S}^T \right) \mathbf{x}^T = \mathbf{x} \mathbf{R} \mathbf{x}^T$$

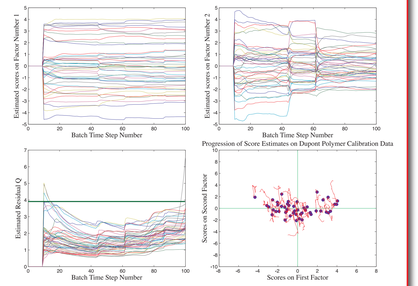
Comment on CLS with Missing Data

- Solution based on complete the squares works for CLS, as does solution based on truncated CLS
- Solutions the same, as before

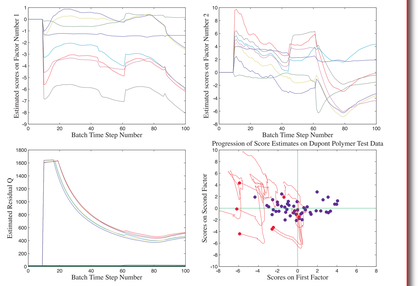
Data Sets

- Semiconductor etch process
  - 80 time steps by 12 variables by 107 batches
  - 20 test batches
- Dupont batch polymer process
  - 100 time steps by 10 variables by 47 batches
  - 8 test batches
- EEM of sugar (see MATLAB demo)
  - 7 excitation by 44 emission by 268 samples

MPCA on Dupont Polymer Calibration



MPCA on Dupont Polymer Test



Conclusions

- Methods based on minimizing residual and truncated CLS equivalent for refitting PCA/MPCA models to incomplete data records
- Same is true for CLS/PARAFAC models
- Solutions converge to final results very early for some processes (semiconductor etch, EEM of sugar) not so well for others (Dupont polymer)

EIGENVECTOR RESEARCH INCORPORATED