

Partial least-squares regression with unlabeled data

Paman Gujral, Barry M. Wise, Michael Amrhein, and Dominique Bonvin

Laboratoire d'Automatique
Ecole Polytechnique Fédérale de Lausanne



- Introduction
 - Background, definitions and motivation

- Improved models using unlabeled data
 - PCR
 - PLSR

- Cautionary note!

Application domain: spectroscopy

Spectroscopic model: $\mathbf{x}_c^T = y_c \mathbf{s}^T + (\text{spectra from other species}) + \text{noise}_c$

Calibration: $\{\mathbf{X}_c, \mathbf{y}_c\} \rightarrow \hat{\mathbf{b}}$

Prediction: $\hat{\mathbf{y}}_p = \mathbf{X}_p \hat{\mathbf{b}}$

Application domain: spectroscopy

Spectroscopic model: $\mathbf{x}_c^T = y_c \mathbf{s}^T + (\text{spectra from other species}) + \text{noise}_c$

Calibration: $\{\mathbf{X}_c, \mathbf{y}_c\} \rightarrow \hat{\mathbf{b}}$

Prediction: $\hat{\mathbf{y}}_p = \mathbf{X}_p \hat{\mathbf{b}}$

Definitions

Labeled data: \mathbf{X} and \mathbf{y} are both available, e.g. $\{\mathbf{X}_c, \mathbf{y}_c\}$

Unlabeled data: \mathbf{X} is available but not \mathbf{y} , e.g. $\underbrace{\mathbf{X}_p, \text{extra measurements } \mathbf{X}_e}_{\mathbf{X}_u}$

Introduction

Application domain: spectroscopy

Spectroscopic model: $\mathbf{x}_c^T = y_c \mathbf{s}^T + (\text{spectra from other species}) + \text{noise}_c$

Calibration: $\{\mathbf{X}_c, \mathbf{y}_c\} \rightarrow \hat{\mathbf{b}}$

Prediction: $\hat{\mathbf{y}}_p = \mathbf{X}_p \hat{\mathbf{b}}$

Definitions

Labeled data: \mathbf{X} and \mathbf{y} are both available, e.g. $\{\mathbf{X}_c, \mathbf{y}_c\}$

Unlabeled data: \mathbf{X} is available but not \mathbf{y} , e.g. $\underbrace{\mathbf{X}_p, \text{extra measurements } \mathbf{X}_e}_{\mathbf{X}_u}$

unlabeled data \neq missing data

Constituents of prediction error

$$\hat{y}_p = \mathbf{x}_p^T \hat{\mathbf{b}} = (y_p \mathbf{s}^T + \text{spectra from other species}) \hat{\mathbf{b}} + (\text{noise})_p \hat{\mathbf{b}}$$
$$y_p - \hat{y}_p = \underbrace{y_p - (y_p \mathbf{s}^T + \text{spectra from other species}) \hat{\mathbf{b}}}_{\text{modeling error}} - \underbrace{(\text{noise})_p \hat{\mathbf{b}}}_{\text{disturbance}}$$

Definitions

Bias: $E[y_p - \hat{y}_p] \leftarrow$ only modeling error contributes to bias

Variance: $\text{var}[y_p - \hat{y}_p] \leftarrow$ modeling error & disturbance contribute to variance

RMSEP: $\sqrt{\text{Bias}^2 + \text{Variance}}$

Abbreviation: root-mean-square error in prediction (RMSEP)

Regression with unlabeled data

MOTIVATION

Use of \mathbf{X}_u reduces modeling error, thereby reducing both bias & variance of error.
Why?

Standard PCR

$$\text{STEP 1: } \mathbf{X}_c \stackrel{PCA}{=} \mathbf{T}_{PCR} \mathbf{P}_{PCR}^T + \mathbf{E}_{PCR}$$

$$\text{STEP 2: } \{\mathbf{T}_{PCR}, \mathbf{y}_c\} \xrightarrow{LS} \hat{\mathbf{b}}_{PCR}$$

Thomas' PCR with unlabeled data Th-PCR (1990, 1995)

$$\text{STEP 1: } \begin{bmatrix} \mathbf{X}_c \\ \mathbf{X}_u \end{bmatrix} \stackrel{PCA}{=} \begin{bmatrix} \mathbf{T}_{Th,c} \\ \mathbf{T}_{Th,u} \end{bmatrix} \mathbf{P}_{Th}^T + \mathbf{E}_{Th}$$

$$\text{STEP 2: } \{\mathbf{T}_{Th,c}, \mathbf{y}_c\} \xrightarrow{LS} \hat{\mathbf{b}}_{Th}$$

Abbreviations: principal component analysis (PCA); least squares (LS)

Regression with unlabeled data

MOTIVATION

Use of \mathbf{X}_u reduces modeling error, thereby reducing both bias & variance of error.
Why?

Standard PCR

$$\text{STEP 1: } \mathbf{X}_c \stackrel{PCA}{=} \mathbf{T}_{PCR} \mathbf{P}_{PCR}^T + \mathbf{E}_{PCR}$$

$$\text{STEP 2: } \{\mathbf{T}_{PCR}, \mathbf{y}_c\} \xrightarrow{LS} \hat{\mathbf{b}}_{PCR}$$

Thomas' PCR with unlabeled data Th-PCR (1990, 1995)

$$\text{STEP 1: } \begin{bmatrix} \mathbf{X}_c \\ \mathbf{X}_u \end{bmatrix} \stackrel{PCA}{=} \begin{bmatrix} \mathbf{T}_{Th,c} \\ \mathbf{T}_{Th,u} \end{bmatrix} \mathbf{P}_{Th}^T + \mathbf{E}_{Th}$$

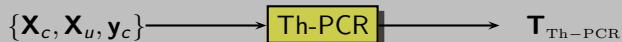
$$\text{STEP 2: } \{\mathbf{T}_{Th,c}, \mathbf{y}_c\} \xrightarrow{LS} \hat{\mathbf{b}}_{Th}$$

Because scores, loadings, and error are estimated from more samples.
Hence, better noise averaging

Abbreviations: principal component analysis (PCA); least squares (LS)

Improved PCR model with unlabeled data

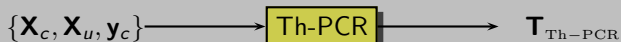
Solution 1: Thomas' PCR



Thomas E.V. *Journal of Chemometrics*, 9, 471-481.

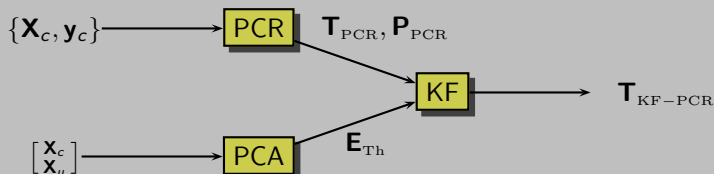
Improved PCR model with unlabeled data

Solution 1: Thomas' PCR



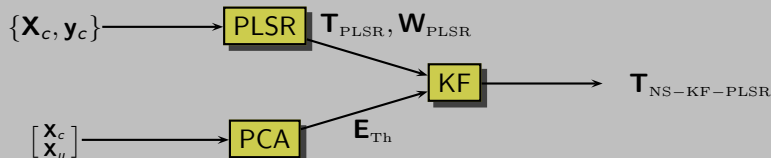
Thomas E.V. *Journal of Chemometrics*, 9, 471-481.

Solution 2: Using a static Kalman filter (KF)



Improved PLSR model with unlabeled data

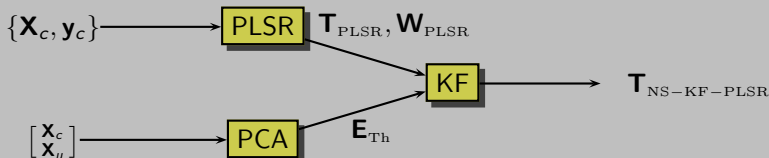
Solution 1: Using a static **non-sequential** KF



Ergon R. *et al.* *Journal of Chemometrics*, 16, 401-407.

Improved PLSR model with unlabeled data

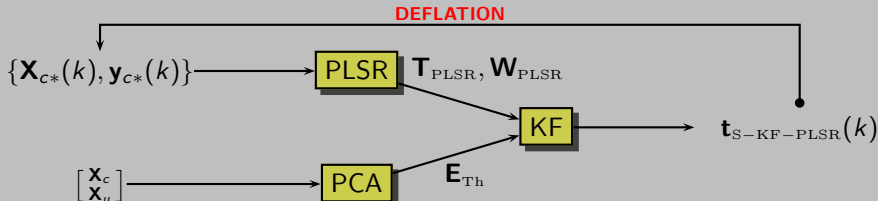
Solution 1: Using a static **non-sequential** KF



Ergon R. *et al.* *Journal of Chemometrics*, 16, 401-407.

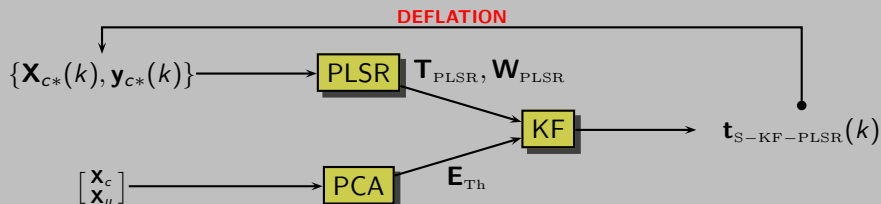
Solution 2: Using a static **sequential** KF

Initialize: $\{\mathbf{X}_{c^*}(0), \mathbf{y}_{c^*}(0)\} = \{\mathbf{X}_c, \mathbf{y}_c\}$



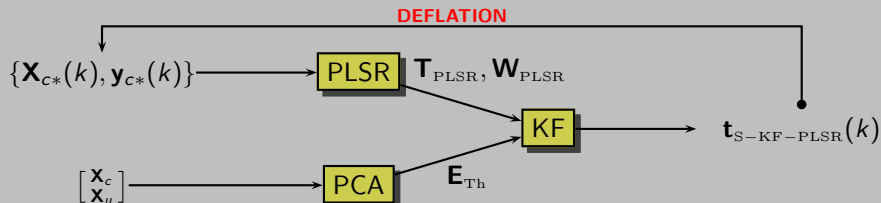
Interpretation of improved PLSR model

Solution 2 repeated ...

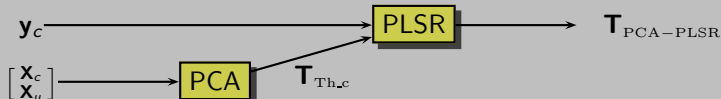


Interpretation of improved PLSR model

Solution 2 repeated ...



Solution 3:



Analytical results: $\mathbf{T}_{\text{S-KF-PLSR}} = \mathbf{T}_{\text{PCA-PLSR}}$!

Cautionary note

So far,

$$\begin{aligned}\mathbf{X}_c &= \mathbf{Y}_c \mathbf{S} + \mathbf{E}_c \\ \mathbf{X}_u &= \mathbf{Y}_u \mathbf{S} + \mathbf{E}_u\end{aligned}$$

Now consider,

$$\begin{aligned}\mathbf{X}_c &= \mathbf{Y}_c \mathbf{S} + \mathbf{E}_c \\ \mathbf{X}_u &= \mathbf{Y}_u \mathbf{S} + \mathbf{1} \mathbf{d}^T + \mathbf{E}_u\end{aligned}$$

Interpretation of \mathbf{d} :

- \mathbf{d} = drift due to difference in background, temperature or pH, extra component due to new analyte etc.
- $\mathbf{1} \mathbf{d}^T + \mathbf{E}_u$ = non-zero mean noise

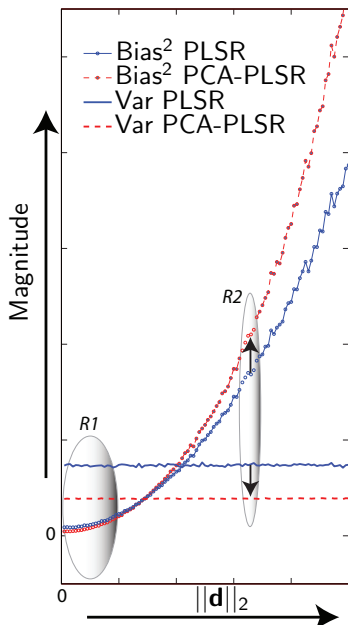
Constituents of prediction error in presence of drift

$$\hat{y}_p = \mathbf{x}_p^T \hat{\mathbf{b}} = (y_p \mathbf{s}^T + \text{spectra from other species}) \hat{\mathbf{b}} + \mathbf{d}^T \hat{\mathbf{b}} + (\text{noise})_p \hat{\mathbf{b}}$$
$$y_p - \hat{y}_p = \underbrace{y_p - (y_p \mathbf{s}^T + \text{spectra from other species}) \hat{\mathbf{b}}}_1 - \underbrace{\mathbf{d}^T \hat{\mathbf{b}}}_2 - \underbrace{(\text{noise})_p \hat{\mathbf{b}}}_3$$

Prediction error ($y_p - \hat{y}_p$) has three constituents:

- 1 due to the PCR/PLSR modeling error
- 2 **due to drift**
- 3 due to random disturbance

Cautionary note



- R1: Bias and variance lower in PCA-PLSR when drift is negligible
- R2: Reduction in variance is offset by the increase in bias due to drift

Cautionary note

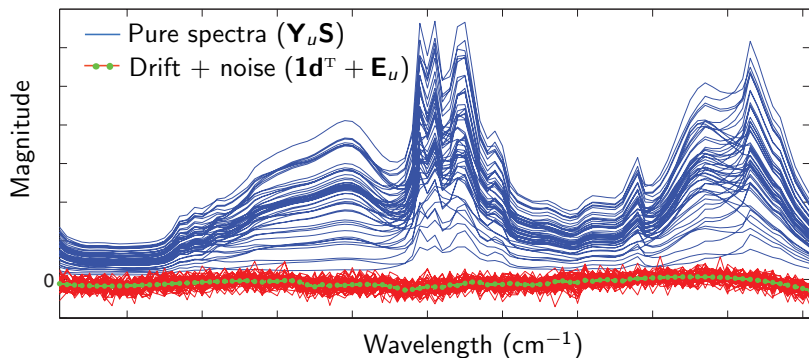


Figure: One realization of data with $\|\mathbf{d}\|_2$ in region R_2

Even very small amounts of drift can offset the benefits

- Improved models using unlabeled data
 - PCR \rightarrow Thomas' PCR
 - PLSR \rightarrow Ergon's PLSR based on Kalman filtering framework

- Improved models using unlabeled data
 - PCR \rightarrow Thomas' PCR
 - PLSR \rightarrow Ergon's PLSR based on Kalman filtering framework
 - Interpretation: equivalence of S-KF-PLSR and PCA-PLSR

- Improved models using unlabeled data
 - PCR \rightarrow Thomas' PCR
 - PLSR \rightarrow Ergon's PLSR based on Kalman filtering framework
 - Interpretation: equivalence of S-KF-PLSR and PCA-PLSR
- Cautionary note: avoid using prediction data as unlabeled data

Extra 1

WITH LABELED DATA ONLY

Wold's PLSR

orthogonal \mathbf{T}_w

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

$$\mathbf{y}_c = \mathbf{T}_w \mathbf{q}_w + \mathbf{f}$$

$$\mathbf{E}_w \mathbf{P} \neq \mathbf{0}$$

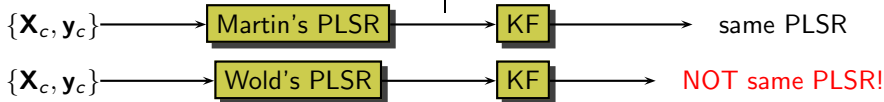
Martin's PLSR

non-orthogonal \mathbf{T}_M

$$\mathbf{X}_c = \mathbf{T}_M \mathbf{W}^T + \mathbf{E}_M$$

$$\mathbf{y}_c = \mathbf{T}_M \mathbf{q}_M + \mathbf{f}$$

$$\mathbf{E}_M \mathbf{W} = \mathbf{0}$$



The error term in Wold's PLSR is not consistent. Modification proposed by Ergon: *Re-interpretation of NIPALS results solves PLSR inconsistency problem*, DOI: 10.1002/cem.1180

Why is the bias due to drift larger in Th-PCR than in PCR?

$$\begin{aligned}\mathbf{b}_{\text{PCR}}^T \mathbf{d} &= \|\mathbf{b}_{\text{PCR}}\|_2 \|\mathbf{d}\|_2 \cos(\theta_{\text{PCR}}) \\ \mathbf{b}_{\text{Th}}^T \mathbf{d} &= \|\mathbf{b}_{\text{Th}}\|_2 \|\mathbf{d}\|_2 \cos(\theta_{\text{Th}})\end{aligned}$$

Use of unlabeled data makes loading subspace include \mathbf{d} , hence

$$\begin{aligned}\theta_{\text{Th}} &< \theta_{\text{PCR}} \\ \mathbf{b}_{\text{Th}}^T \mathbf{d} &> \mathbf{b}_{\text{PCR}}^T \mathbf{d}\end{aligned}$$

The same holds also for PCA-PLSR.