On the bias-variance trade-off in principal component regression with unlabeled data

Paman Gujral[1], Barry M. Wise[2], Michael Amrhein[1] and Dominique Bonvin[1]

[1] Laboratoire d'Automatique, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

[2] Eigenvector Research Inc., 3905 West Eaglerock Drive, WA–98801 Wenatchee, USA

Abstract

It has been shown that the prediction error from PCR can be reduced by using both the labeled and unlabeled data for stabilizing the principal component subspace, while using only the labeled calibration data in the regression step (T. V. Edward, Journal of Chemometrics, 1995, 9(6), pp. 471-481). When the unlabeled data represents the labeled data well, this leads to a reduction in both the bias and the variance components of RMSEP. However, in many practical problems, the unlabeled data may represent the labeled data only approximatively. One such case is analyzed where the two data sets have a slightly different background.

Background and motivation

Multivariate spectroscopic calibration: instrumental measurements $\mathbf{X}_c$ are related to the corresponding reference analyte concentrations $\mathbf{y}_c$ by the inverse regression model $\mathbf{y}_c = \mathbf{X}_c\mathbf{b} + \mathbf{e}$

Labeled data $\{\mathbf{X}_c, \mathbf{y}_c\}$, unlabeled data $\{\mathbf{X}_u, --\}$

- Unlabeled data might encompass the additional measurements available during calibration or the prediction data available off-line
- Usually, spectral data from a sample are easy and inexpensive to obtain - the reference analysis is the resource-demanding step

Standard PCR:

1. Compute the PCA factorization of $\mathbf{X}_c = \mathbf{T}_{1c}\mathbf{P}_1 + \mathbf{E}_1$
2. Estimate $\mathbf{b}$ via a least-squares regression between $\mathbf{X}_c\mathbf{P}_1$ and $\mathbf{y}_c$

$\mathbf{X}_u$ can be used during calibration

Edwards' PCR with unlabeled data:

1. Compute the PCA factorization of $\begin{bmatrix}\mathbf{X}_c\\ \mathbf{X}_u\end{bmatrix} = \begin{bmatrix}\mathbf{T}_{2c}\\ \mathbf{T}_{2u}\end{bmatrix}\mathbf{P}_2 + \mathbf{E}_2$
2. Estimate $\mathbf{b}$ via a least-squares regression between $\mathbf{X}_c\mathbf{P}_2$ and $\mathbf{y}_c$

Case A: $\mathbf{X}_c$ and $\mathbf{X}_u$ are from the same measurement model $\mathbf{X} = \mathbf{Y}\,\mathbf{S} + \mathbf{E}$

Monte Carlo simulation study to compute percentage reduction in bias and variance components of RMSEP with Edwards' PCR for following examples where the y-ranges for labeled and unlabeled data are varied:

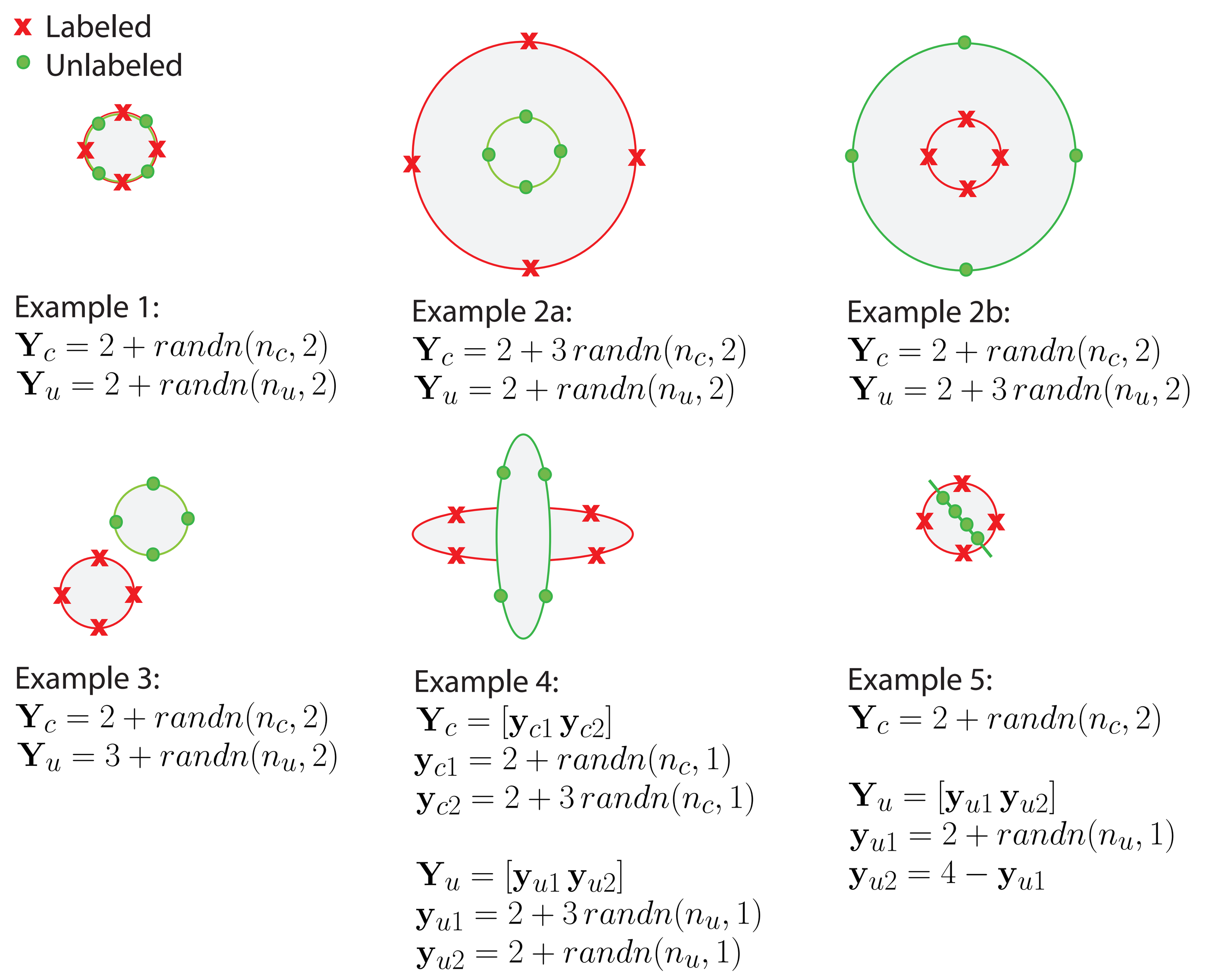

Fig. 1: Schematic diagram of $\mathbf{Y}_c$ and $\mathbf{Y}_u$ using MATLAB command $randn()$ that draws samples from a normal distribution.

Example	%ΔRMSEP
1	12
2a	0
2b	38
3	12
4	8
5	19

Edwards' PCR leads to lower RMSEP for all examples except 2a, where unlabeled data have low leverage

$$RMSEP^2 = bias^2 + variance$$

Both bias and variance components of RMSEP are reduced due to better latent space estimation

Case B: $\mathbf{X}_u$ has drift, i.e. $\mathbf{X}_u$ comes from the measurement model $\mathbf{X}_u = \mathbf{Y}_u\,\mathbf{S} + \mathbf{1}\,\mathbf{d}^{\mathrm{T}} + \mathbf{E}_u$

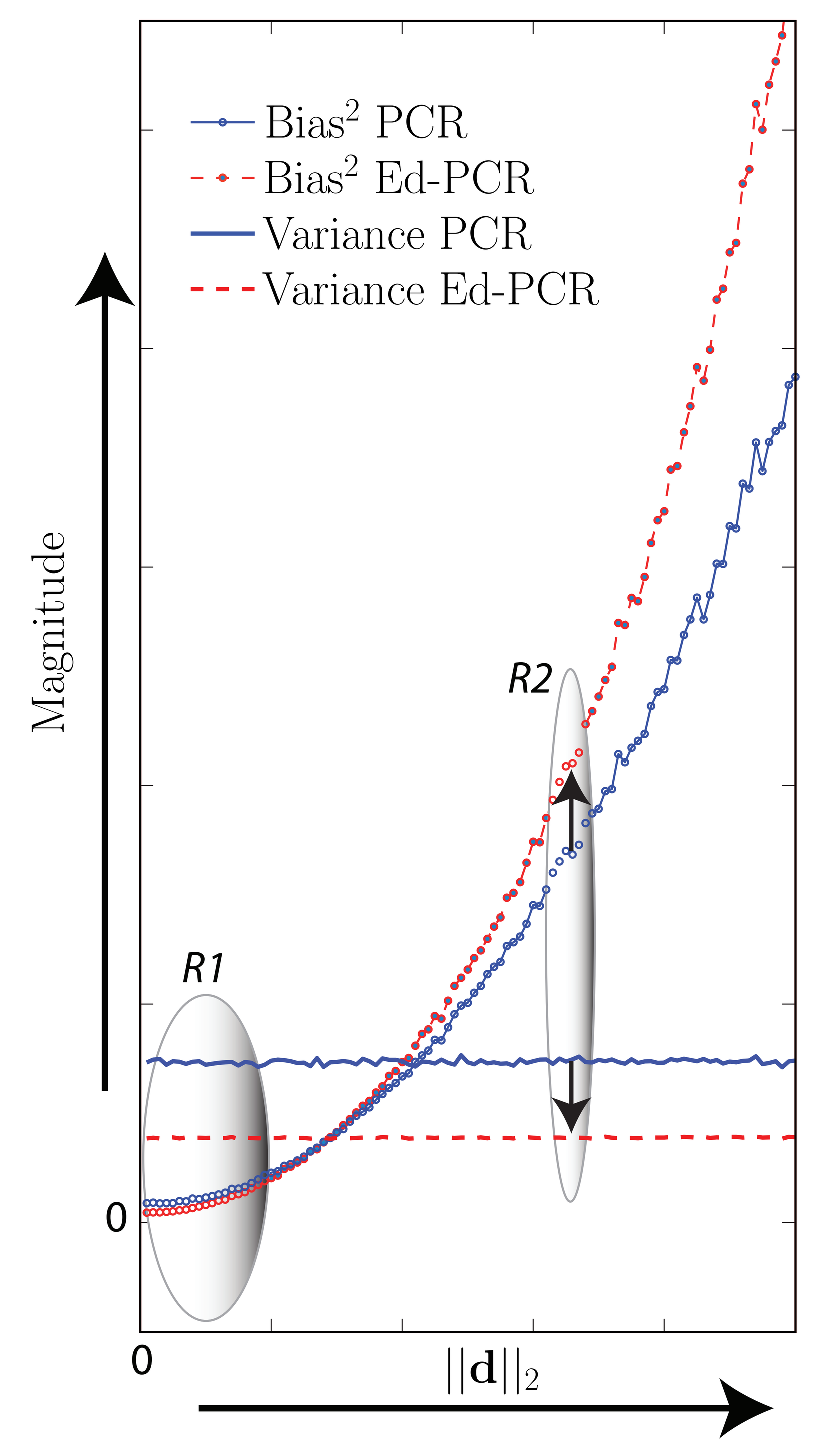

Fig. 2: Monte Carlo simulation to compute the bias and variance components of RMSEP with PCR and with Edwards' PCR, plotted as a function of $||\mathbf{d}||_2$.

$$\mathbf{X}_c = \mathbf{Y}_c\,\mathbf{S} + \mathbf{E}_c$$
$$\mathbf{X}_u = \mathbf{Y}_u\,\mathbf{S} + \mathbf{1}\,\mathbf{d}^{\mathrm{T}} + \mathbf{E}_u$$

Interpretation of $\mathbf{d}$:

- non-zero mean noise
- difference in background (baseline)
- extra component due to new analyte, change in temperature, pH or probe alignment etc.

Note the regions

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

Why is the bias due to drift more in Edwards' PCR?

$$\mathbf{b}_{pcr}^{\mathrm{T}}\mathbf{d} = ||\mathbf{b}_{pcr}||_2\,||\mathbf{d}||_2\,cos(\theta_{pcr})$$
$$\mathbf{b}_{Ed}^{\mathrm{T}}\mathbf{d} = ||\mathbf{b}_{Ed}||_2\,||\mathbf{d}||_2\,cos(\theta_{Ed})$$

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

$$\theta_{Ed} < \theta_{pcr}$$
$$\mathbf{b}_{Ed}^{\mathrm{T}}\mathbf{d} > \mathbf{b}_{pcr}^{\mathrm{T}}\mathbf{d}$$

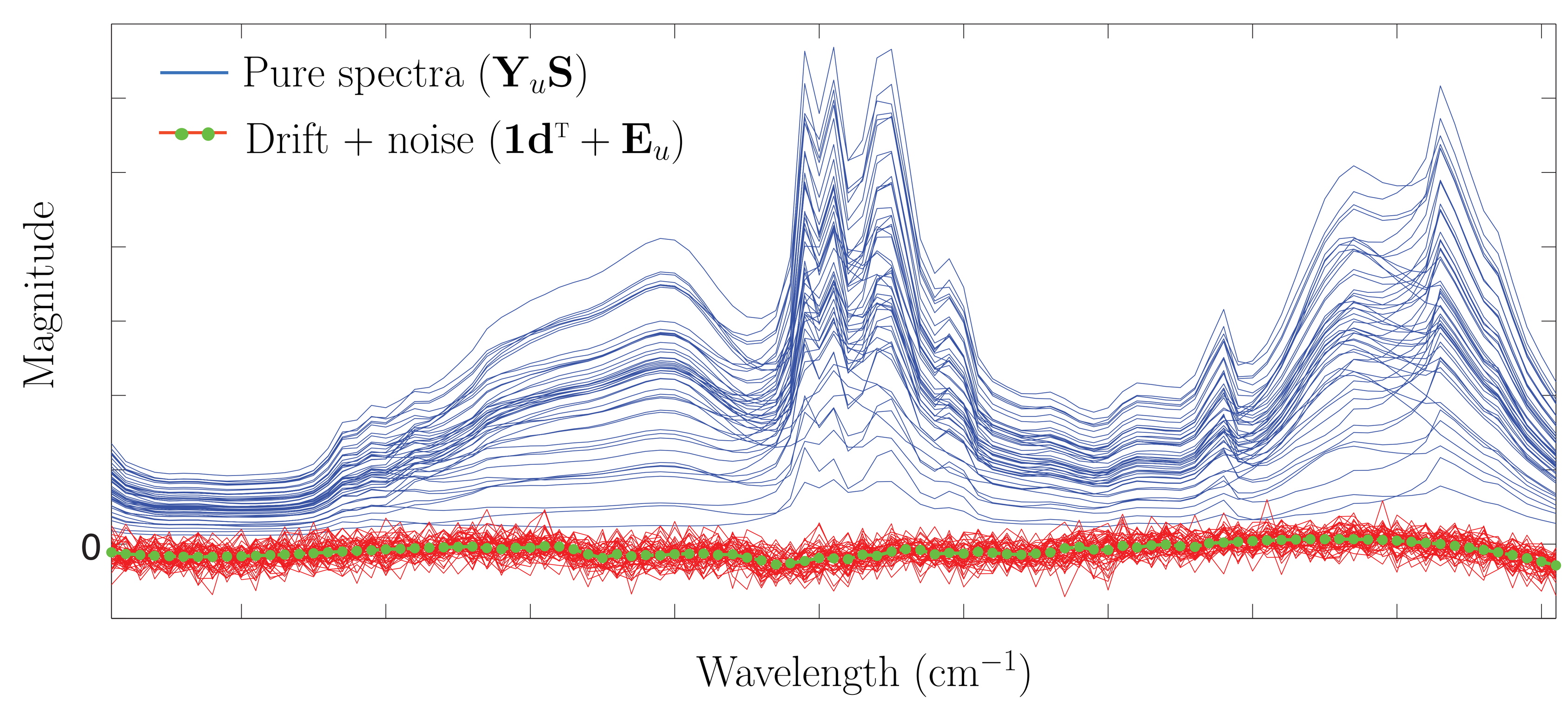

Fig. 3: One realization of data with $||\mathbf{d}||_2$ in region R2 (see Fig. 2).

Even very little drift can offset the gains from using unlabeled data

Conclusions

This study shows, via Monte Carlo simulations, the trade-off in prediction error between (i) smaller variance due to improved estimation of the loading space, and (ii) larger bias due to the reduction of the angle between $\mathbf{b}$ and the drift components. The bias-variance trade-off is unfavorable in the presence of very small amounts of drift in unlabeled data. The latter may often not be verifiable in advance. Hence, Edwards' PCR is recommended only with extra $\mathbf{X}$ measurements collected during calibration, not with prediction data available off-line.

E-mail of corresponding author: paman.gujral@epfl.ch