

Measurement-Based Drift Correction in Spectroscopic Calibration Models

P. Gujral¹ M. Amrhein² D. Bonvin³

Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

¹ paman.gujral@epfl.ch

² michael.amrhein@epfl.ch

³ dominique.bonvin@epfl.ch

Keywords: Drift, drift correction, calibration, projection, space inclusion.

1 Introduction

A spectrometer can be calibrated for an analyte of interest in the presence of interferents whose concentrations do not even have to be known [10]. For dynamic processes such as chemical reactions, data from various batches can be combined for calibration. This way, interferents and prevailing drifts are included in the calibration set. The drifts are caused by instrumental, operational and process changes, such as the effect of temperature, pressure and pH on the instrument, residue accumulation or aging of the instrument, changes in probe alignment, operational offsets, and interactions between species [7, 5]. As an alternative to combining batches, on-line or at-line reference measurements might be added to the off-line database to update the calibration model [9, 6, 1].

The methods that include drift in the calibration set are referred to as *implicit correction models* (ICM). In contrast, *explicit correction models* (ECM), such as Dynamic Orthogonal Projection (DOP) [8, 4], model the drift space based on reference measurements and make the calibration model orthogonal or invariant to that space. In this study, the original DOP algorithm will be modified so as to improve the estimation of the drift space.

Correct prediction of the analyte concentration from a new spectrum without drift is possible provided the spectrum lies in the row space spanned by the calibration spectra (*space-inclusion condition*) [3]. An extended space-inclusion condition, which new spectra possibly corrupted with drift should fulfill, will be proposed for all drift-correction methods.

In addition to the drift component, which causes a bias in the measurements, the measurements are also corrupted by noise. A Monte-Carlo simulation study with additive Gaussian noise will illustrate and compare the performance of the drift-correction methods in the presence of various drift types.

The basic principles of multivariate calibration for spectroscopic data are reviewed in Section 2. Section 3 extends the space-inclusion condition to the case of drifts. The results of a simulated case study are presented in Section 4, and Section 5 concludes the paper. For simplicity of notation, the theoretical results developed in Sections 2 and 3 are for the noise-free case, while the simulation study in Section 4 contains noise.

2 Preliminaries

2.1 Calibration model

Spectral absorbance matrix. Let $\mathbf{x}(m)$ denote the spectral (absorbance) vector of an n_r -channel instrument, and $\mathbf{y}(m)$ the n_s -dimensional concentration vector at the observation instant m , where n_s is the number of

absorbing species. For spectral data depending *linearly* on \mathbf{y} (e.g. when Beer's law is valid), one can write:

$$\mathbf{x}^T(m) = \mathbf{y}^T(m)\mathbf{E} \quad (1)$$

where \mathbf{E} is the $[n_s \times n_r]$ pure-component spectra matrix. For n_c off-line calibration measurements, Eq. (1) can be written in matrix form as:

$$\mathbf{X} = \mathbf{Y}\mathbf{E} \quad (2)$$

with \mathbf{X} being the $[n_c \times n_r]$ spectral (absorbance) matrix and \mathbf{Y} the $[n_c \times n_s]$ concentration matrix.

Calibration. Let n_k be the number of absorbing species for which concentrations are available for calibration (called *species of interest*), and n_u the number of remaining absorbing species (called *interferents*). \mathbf{Y} may be separated into the $[n_c \times n_k]$ known part \mathbf{Y}_k and the $[n_c \times n_u]$ unknown part \mathbf{Y}_u :

$$\mathbf{Y} = [\mathbf{Y}_k \ \mathbf{Y}_u] \quad (3)$$

The inverse calibration model reads [10]:

$$\mathbf{Y}_k = \mathbf{X}\mathbf{B} \quad (4)$$

where \mathbf{B} is a $[n_r \times n_k]$ regressor matrix that can be estimated using a variety of methods such as Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), and Continuum Regression (CR) [10]. The difference between the various methods stems from the way the noise in \mathbf{X} and \mathbf{Y}_k is handled. In the absence of noise, the estimate of \mathbf{B} is the same for each of these methods:

$$\hat{\mathbf{B}} = (\mathbf{X})^+ \mathbf{Y}_k \quad (5)$$

where $^+$ stands for the Moore-Penrose inverse.

Prediction. Let $\text{rank}(\mathbf{E}) = n_s$ and $\mathbf{x}(m)$ be a new spectrum obeying (1). The concentrations of the n_k species are predicted correctly from $\mathbf{x}(m)$ using

$$\hat{\mathbf{y}}_k^T(m) = \mathbf{x}^T(m)\hat{\mathbf{B}} \quad (6)$$

if $\mathbf{x}(m)$ satisfies the space-inclusion condition, i.e. $\mathbf{x}(m) \in \mathcal{S}_r(\mathbf{X})$, where $\mathcal{S}_r(\mathbf{X})$ denotes the row space of \mathbf{X} . This condition, which is necessary and sufficient in the absence of interferents, is only sufficient in the presence of unknown interferents. (See [3] for a proof)

2.2 Update of calibration model using on-line reference measurements

Let us assume that, during on-line operation, the measured spectrum $\mathbf{x}(m)$ is given by:

$$\mathbf{x}^T(m) = \mathbf{y}_k^T(m)\mathbf{E}_k + \mathbf{d}^T(m) \quad (7)$$

where \mathbf{E}_k contains the $[n_k \times n_r]$ pure-component spectra of the n_k known species. \mathbf{d} is a n_r -dimensional drift component lying in a rank- r subspace with $r \ll n_r$, i.e. the drift components of several observations can be linearly dependent. The first term in Eq. (7) models the spectrum according to the known species, while the second term corresponds to the spectrum of the interferents together with the drift caused by (i) baseline shift, (ii) the differences in pH or interactions like hydrogen bonding that result in positional peak shifts and change the shape of the peak (physico-chemical interactions), and (iii) the multiple path lengths of the light reaching the exit slit of the instrument (stray light).

The drift component causes a bias in the prediction given by Eq. (6). Hence, the calibration model needs to be corrected using on-line reference measurements. Let \mathbf{X}_τ contain n_τ reference measurements and \mathbf{Y}_τ be the corresponding analyte concentrations that are measured by reference analytics. Eq. (7) gives:

$$\mathbf{X}_\tau = \mathbf{Y}_\tau \mathbf{E}_k + \mathbf{D} \quad (8)$$

where \mathbf{X}_τ , \mathbf{Y}_τ and \mathbf{D} are of size $[n_\tau \times n_r]$, $[n_\tau \times n_k]$, and $[n_\tau \times n_r]$, respectively. With ICM, the data pair $\{\mathbf{X}_\tau, \mathbf{Y}_\tau\}$ collected during the run is periodically appended to the calibration database [1, 6], and the regressor matrix \mathbf{B} is re-estimated using the data pair $\left\{ \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_\tau \end{bmatrix}, \begin{bmatrix} \mathbf{Y}_k \\ \mathbf{Y}_\tau \end{bmatrix} \right\}$ and Eq. (5). With ECM, \mathbf{D} is estimated from the data pair $\{\mathbf{X}_\tau, \mathbf{Y}_\tau\}$ as $\hat{\mathbf{D}}$ [4, 8]. The $[n_r \times n_r]$ orthogonal projection matrix $\hat{\mathbf{N}} = (\mathbf{I} - \hat{\mathbf{D}}^+ \hat{\mathbf{D}})$ is computed, and with each new on-line reference measurement, the regressor matrix \mathbf{B} is re-estimated using the data pair $\{\mathbf{X}\hat{\mathbf{N}}, \mathbf{Y}_k\}$ and Eq. (5).

In DOP, the estimation of $\hat{\mathbf{D}}$ is based on a kernel approach [4, 8]. A modified kernel is used in this work, which will be detailed in the full paper.

3 Extended space-inclusion condition in the presence of unknown drifts

3.1 Space-inclusion condition for ICM

Proposition 1 Consider the measured spectrum $\mathbf{x}(m)$ that is affected by the unknown drift component $\mathbf{d}(m)$ as given by Eq. (7). Let $\text{rank} \left(\begin{bmatrix} \mathbf{E}_k \\ \mathbf{D} \end{bmatrix} \right) = n_k + r$, $\mathbf{d}(m) \in \mathcal{S}_r(\mathbf{D})$, and on-line reference measurements $\{\mathbf{X}_\tau, \mathbf{Y}_\tau\}$ be available. Then, the concentrations of the n_k species can be predicted correctly from $\mathbf{x}(m)$ using

$$\hat{\mathbf{y}}_k^T(m) = \mathbf{x}^T(m) \hat{\mathbf{B}}^* \quad (9)$$

where

$$\hat{\mathbf{B}}^* = \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_\tau \end{bmatrix}^+ \begin{bmatrix} \mathbf{Y}_k \\ \mathbf{Y}_\tau \end{bmatrix} \quad (10)$$

if $\mathbf{x}(m) \in \mathcal{S}_r \left(\begin{bmatrix} \mathbf{X} \\ \mathbf{X}_\tau \end{bmatrix} \right)$.

(Proof in the full paper). Proposition 1 says that, assuming that the drift component lies within the implicitly included drift space but not in the row space of the n_k pure-component spectra, prediction will not be impaired for a new spectrum if it lies in the row space of the augmented calibration set.

3.2 Space-inclusion condition for ECM

Proposition 2 Consider the measured spectrum $\mathbf{x}(m)$ that is affected by the unknown drift component $\mathbf{d}(m)$ as given by Eq. (7). Let $\text{rank} \left(\begin{bmatrix} \mathbf{E}_k \\ \hat{\mathbf{D}} \end{bmatrix} \right) = n_k + r$, $\mathbf{d}(m) \in \mathcal{S}_r(\hat{\mathbf{D}})$, and on-line reference measurements $\{\mathbf{X}_\tau, \mathbf{Y}_\tau\}$ be available. Then, the concentrations of the n_k species can be predicted correctly from $\mathbf{x}(m)$ using

$$\hat{\mathbf{y}}_k^T(m) = \mathbf{x}^T(m) \hat{\mathbf{B}}^* \quad (11)$$

where

$$\hat{\mathbf{B}}^* = (\mathbf{X}\hat{\mathbf{N}})^+ \mathbf{Y}_k \quad (12)$$

if $\mathbf{x}^T(m) \hat{\mathbf{N}} \in \mathcal{S}_r(\mathbf{X}\hat{\mathbf{N}})$.

(Proof in the full paper). Proposition 2 says that, assuming that the drift component lies within the explicitly estimated drift space but not in the row space of the n_k pure-component spectra, prediction will not be impaired for a new corrected spectrum if it lies in the row space of the corrected calibration set.

4 Simulated batch reactor

4.1 Data generation

Drift correction is illustrated via spectral measurements from a simulated isothermal, constant-volume batch reactor involving $n_s = 4$ absorbing species and 2 independent reactions (example taken from [2]). Reactant P

is converted to the desired product S following a catalyzed two-step reaction:



The mole balances for the batch reactor read:

$$\frac{dy_P}{dt} = -2\kappa_1 y_P^2 \quad (14)$$

$$\frac{dy_Q}{dt} = \kappa_1 y_P^2 - \kappa_2 y_Q y_R. \quad (15)$$

All four species are assumed to absorb and obey Eq. (1). The numerical values for the rate constants are $\kappa_1 = 2.45 \text{ l mol}^{-1} \text{ h}^{-1}$ and $\kappa_2 = 21.33 \text{ l mol}^{-1} \text{ h}^{-1}$. Only P , R and S are assumed to be known during calibration ($n_k = 3$), Q being an unknown interferent ($n_u = 1$). Spectra at $n_r = 101$ channels, measured for $n_c = 49$ mixture samples are used for calibration using principal component regression (PCR) and n_s latent vectors. The on-line measurements $\mathbf{x}(m)$ are constructed according to Eq. (7) from simulated concentrations, pure-component spectra and drift models for (i) baseline shift, (ii) physico-chemical interactions, and (iii) stray light. Zero-mean Gaussian noise with standard deviation σ_c and σ_p is added to the calibration and prediction data, respectively. The value of σ_c is chosen such that the noise level during calibration, defined as σ_c/σ_x , where σ_x is the standard deviation of the calibration spectral matrix averaged over all channels, is 10%. The value of σ_p is varied over a wide range to determine the drift correction ability of ICM and ECM in the presence of different noise levels. Ten on-line reference measurements ($n_r = 10$) are collected at equal intervals during the reaction, and r is chosen so as to capture at least 95% variation in $\hat{\mathbf{D}}$.

4.2 Results and discussion

ICM and ECM give exactly the same prediction when the calibration and prediction data and the on-line reference measurements are noise free (results not shown). In the absence of noise, the space-inclusion conditions can be checked by calculating the Q-statistics [10], which should be zero. Fig. 1 shows that, even after drift correction by ECM and ICM, the Q-statistics are non-zero initially, and thus prediction during this stage is inaccurate. However, after a sufficient number of on-line reference measurements, the Q-statistics goes to zero, thereby leading to correct prediction.

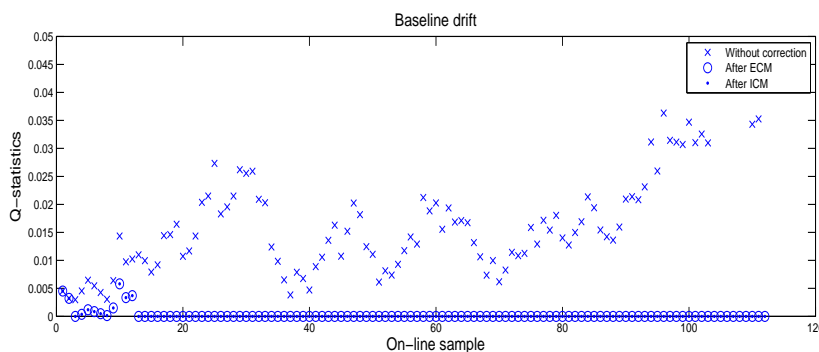


Figure 1: Q-statistics for the noise-free case without correction and with correction based on ECM and ICM for analyte S and for baseline shift.

In the noisy case, a test statistics based on both the Q-statistics and the T^2 -statistics is used [10] (results not shown). In Fig. 2, the average standard error of prediction (SEP) of analyte S from 500 Monte-Carlo simulations with different σ_p/σ_c is shown for each of the three drifts. It can be seen that ICM and ECM perform similarly in the presence of noise. As σ_p/σ_c increases, the average SEP after drift correction approaches the average SEP of the calibration model without correction. Similar results are obtained for species P and R .

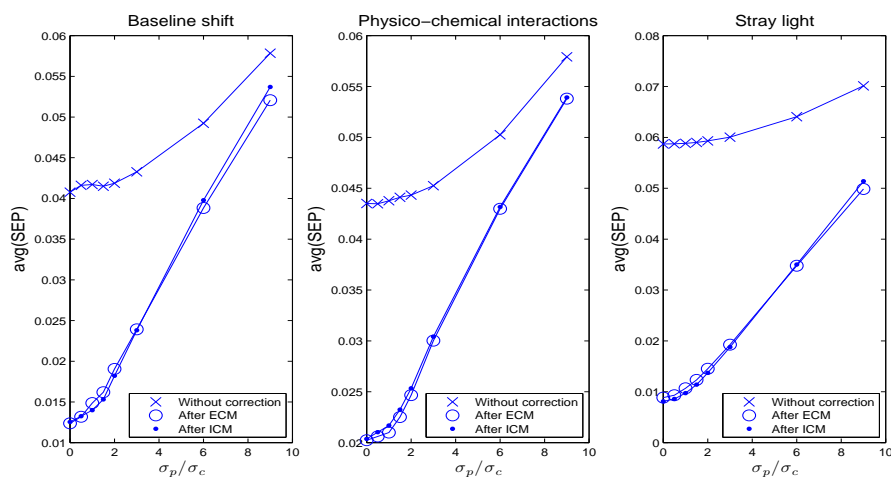


Figure 2: Average SEP without correction and with correction based on ECM and ICM for analyte S and for three drift types: (i) baseline shift, (ii) physico-chemical interactions, and (iii) stray light.

5 Conclusions

If the measurement-based drift-correction methodologies such as ICM and ECM satisfy the space-inclusion conditions, they are equivalent in the absence of noise. The difference between the various methods stems from the way they handle noise. This study has shown that, for additive Gaussian noise, ICM and ECM exhibit similar performance. As the prediction noise increases relative to the calibration noise, the correction methodologies are less able to extract drift information from on-line reference measurements, and the corresponding SEPs approach the SEP of the calibration model without correction.

References

- [1] J. Flores-Cerrillo, J. F. MacGregor: Within-batch and batch-to-batch inferential-adaptive control of semibatch reactors: A Partial Least Squares approach. *Industrial and Engineering Chemistry Research*, 42(14) : 3334-3345, 2003.
- [2] M. Amrhein, B. Srinivasan, D. Bonvin, M. M. Schumacher: On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics and Intelligent Laboratory Systems*, 33(1) : 17-33, 1996.
- [3] M. Amrhein, B. Srinivasan, D. Bonvin, M. M. Schumacher: Calibration of spectral reaction data. *Chemometrics and Intelligent Laboratory Systems*, 46(2) : 249-264, 1999.
- [4] M. Dabros, M. Amrhein, P. Gujral, U. von Stockar: On-line recalibration of spectral measurements using metabolite injections and dynamic orthogonal projection. *Applied Spectroscopy*, 61(5) : 507-513, 2007.
- [5] M. J. Saiz-Abajo, B. H. Mevik, V. H. Segtnan, T. Naes: Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Analytica Chimica Acta*, 533(2) : 147-159, 2005.
- [6] M. R. Riley, M. A. Arnold, D. W. Murhammer: Matrix-enhanced calibration procedure for multivariate calibration models with near-infrared spectra. *Applied Spectroscopy*, 52(10) : 1339-1347, 1998.
- [7] M. S. Larrechi, M. P. Callao: Strategy for introducing NIR spectroscopy and multivariate calibration techniques in industry. *TrAC Trends in Analytical Chemistry*, 22(9) : 634-640, 2003.
- [8] M. Zeaiter, J. M. Roger, V. Bellon-Maurel: Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems*, 80(2) : 227-235, 2006.
- [9] N. Gallagher, B. Wise, S. Butler, D. White, G. Barna: Benchmarking of multivariate statistical process control tools for a semiconductor etch process: Improving robustness through model updating. *International Symposium on Advanced Control of Chemical Processes (IFAC ADCHEM'97)*, 1997, pp. 78-83.
- [10] T. Naes, T. Isaksson, T. Fearn, T. Davies: *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, 2002.