

The Parallel Calibration Method

Clifford H. Spiegelman, Jerome F. Bennett, Marina
Vannucci, Michael J. McShane, and Gerard L. Côté

Acknowledgement

- The ANRCP is gratefully acknowledged for their partial support of this research.

Simple is best

- Many of the great advances within the scientific community are the result of deep yet simple concepts. For instance, Newton's three laws replaced Kepler's rather complicated but useful laws of planetary motion. The scientific community has always replaced complicated solutions with equally good or better simple solutions as soon as they become available. In this paper, we present a new calibration method for the analysis of scientific data, called the parallel method, that is simple and often better when compared to standard calibration methods. We compare the parallel method to standard methods such as classical least-squares and partial least-squares on two scientific data sets and on one computer-generated data set. This new method shows better or comparable results for these data sets in terms of mean squared error but more importantly this method is much simpler than the popular partial least-squares approach and does not require a user-defined selection of latent variables.

Univariate PLS

$$Y = \begin{matrix} y_{11} & \cdots & y_{1q} \\ \vdots & \dots & \vdots \\ y_{n1} & & y_{nq} \end{matrix} .$$

$$\hat{\xi}_{PLS} = \left(\frac{x' Y Y' x}{x' (Y Y')^2 x} \right) Y' x, \text{ Where}$$

$$\hat{\xi} = Z \hat{\beta}_{pls} .$$

Basic Model

$$Y_{nxq} = x_{nxp} \beta_{pxq} + \varepsilon_{nxq}$$

x

After Calibration we observe

$$Z = \xi \beta + \varepsilon^* .$$

Parallel calibration method

Let the linear operator A , from $\mathbb{R}^{n \times q}$ to \mathbb{R}^q , be (a_1, \dots, a_n) , where we have

$$A(y_1, \dots, y_n) = \sum_{i=1}^n a_i y_i. \text{ It is assumed that there is a corresponding operator } A^* \text{ from } \mathbb{R}^{n \times p} \text{ to}$$

\mathbb{R}^p such that the unique c^* corresponding to $\sum_{i=1}^n a_i y_i$ is $\sum_{i=1}^n a_i c_i$. Note that except for the fact

that q usually does not equal p , $A=A^*$.

This assumption is satisfied for full rank linear models. Secondly, we require the rank of (y_1, \dots, y_n) be n and that $q > n > p$. The main idea behind the estimation method is as follows. We look for an operator \hat{A} so that $A(Y_1, \dots, Y_n) = Z$ and then estimate c_{new} by

$A^*(c_1, \dots, c_n) = \sum_{i=1}^n a_i c_i$. Thus, Z , a newly measured spectrum, can be estimated by a linear combination of the calibration spectra.

Connection to ridge regression

$$Y_{nxq} = X_{nxp} \beta_{pxq} + \varepsilon$$

$$Z_{1xq} = x_{01xp} \beta_{pxq} + \varepsilon^*$$

$\|Z - \alpha Y\|^2$ is minimized by $\hat{\alpha}$

$$\hat{x}_o = \hat{\alpha} x$$

- Now suppose $p=1$.

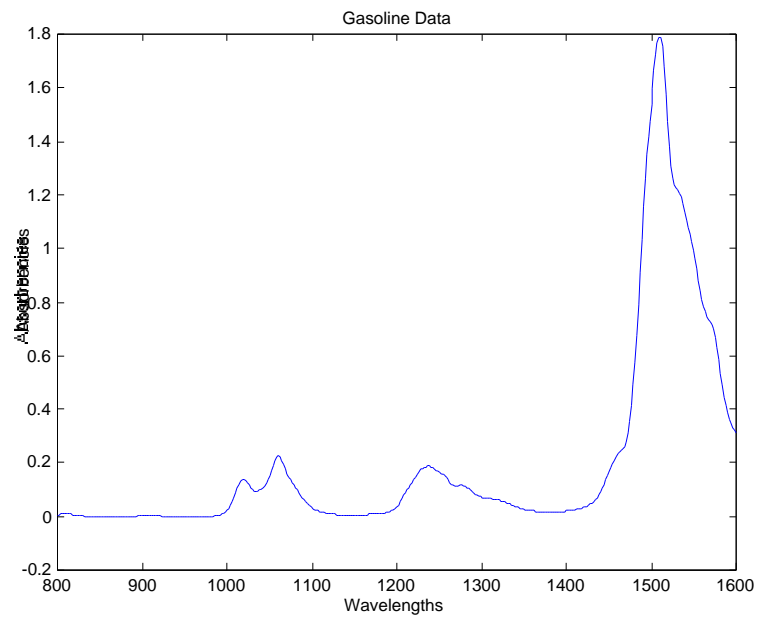
$$\hat{\alpha} = ZY' (YY')^{-1} = x_0 \beta \beta' x' (x \beta \beta' x')^{-1}$$

Since $\beta \beta'$ is a scalar

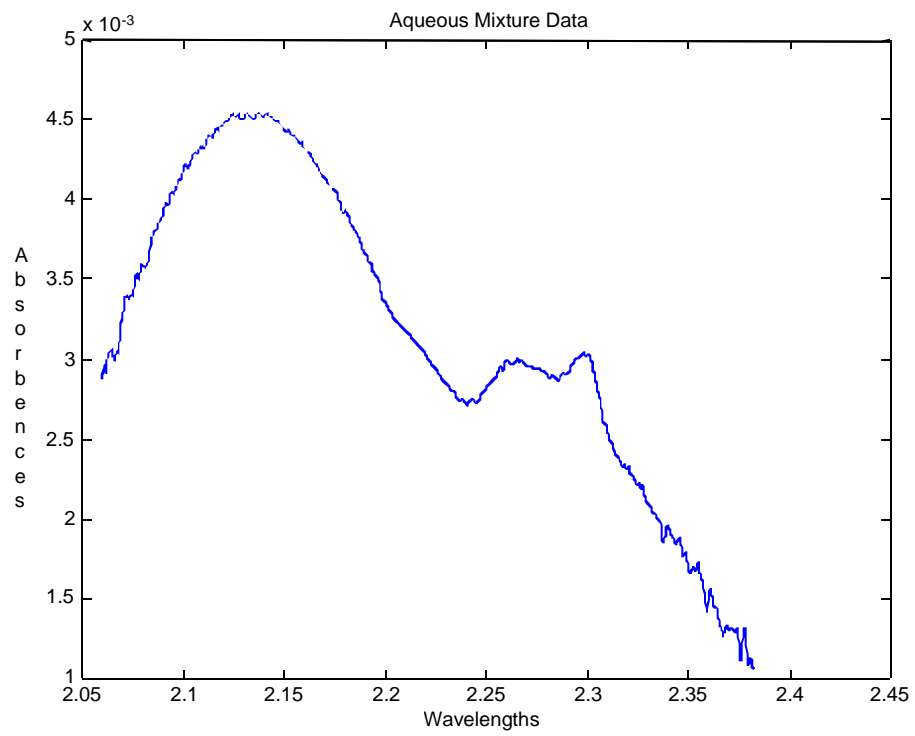
$$\hat{\alpha} = x_0 x' (x x')^{-1}$$

Connection to ridge regression is through the error structure

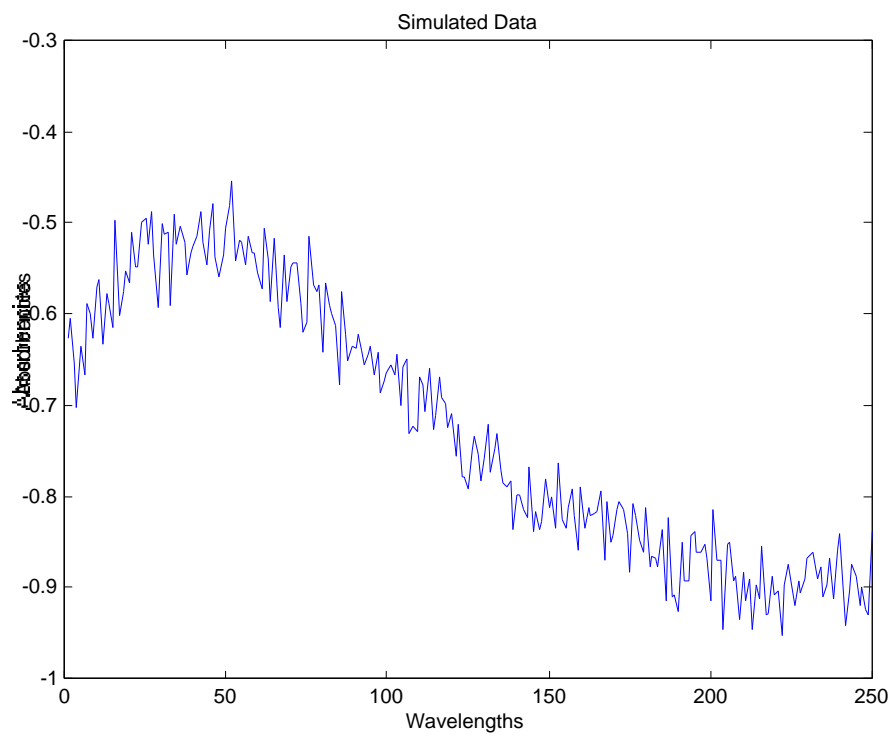
Gasoline data



Aqueous mixture data



Simulated data



Results for gasoline

- Table 1.) Mean squared errors for pseudo-gasoline data. Leave-one-out cross validation was used to estimate mean squared error.

• Method	G1	G2	G3	G4	G5	Ave*	Corr†
• PLS	0.45	0.51	0.41	0.37	1.84	0.72	0.99
• PAR	0.08	0.10	0.13	0.07	0.07	0.09	1.00
• CLS	9.78	74.87	13.96	39.97	25.19	32.75	0.59

- * 'Ave' is the average MSE over the ingredients for each method.
- † 'Corr' is the correlation between the predicted concentrations and the true concentration for each method.

Results for aqueous mixture

- Table 2.) Mean squared errors for aqueous mixture data. Mean squared error is computed for a test set comprised of 20% of the total sample size.

• Method	Glucose	Lactate	Ammonia	Glutamate	Glutamine	Ave*	Corr†
• PLS	0.41	0.19	0.08	0.26	0.33	0.25	1.00
• PAR	0.12	0.01	0.00	0.01	0.12	0.05	1.00
• CLS	3.62	0.13	0.09	0.23	2.70	1.36	1.00

- * 'Ave' is the average MSE over the ingredients for each method.
- † 'Corr' is the correlation between the predicted concentrations and the true concentration for each method.

Results for simulation

- Table 3.) Simulation with $\sigma = 0.03$. Mean squared error is computed for a test set comprised of 50% of the total sample size.

• Method	ingr1	ingr2	ingr3	Ave*	Corr†
• PLS	3.37	4.97	13.64	7.33	0.91
• PAR	2.68	4.17	9.93	5.59	0.92
• CLS	119.7	192.8	449.3	254.0	0.34

- * 'Ave' is the average MSE over the ingredients for each method.
- † 'Corr' is the correlation between the predicted concentrations and the true concentration for each method.