Chemometrics and Intelligent Laboratory Systems 107 (2011) 124–138

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

The informative converse paradox: Windows into the unknown

Harald Martens *

Centre for Integrative Genetics CIGENE, Dept. of Mathematical Sciences and Technology, Norwegian University of Life Sciences, N-1430 Ås, Norway
Nofima Mat AS, N-1432 Ås, Norway

ARTICLE INFO

Article history:
Received 23 December 2010
Received in revised form 17 February 2011
Accepted 19 February 2011
Available online 26 February 2011

Keywords:
Causal modeling
Incomplete
Residuals
Informative converse
Spectroscopy
Optimized EMSC

ABSTRACT

Model-based interpretation of empirical data is useful. But unanticipated phenomena (interferences) can give erroneous model parameter estimates, leading to wrong interpretation. However, for multi-channel data, interference phenomena may be discovered, described and corrected for, by analysis of the lack-of-fit residual table — although with a strange limitation, which is here termed the *Informative Converse* paradox: When a data table (rows × columns) is approximated by a linear model, and the model-fitting is done by row-wise regression, it means that only the column-wise interference information can be correctly obtained, and vice versa. These "windows into the unknown" are here explained mathematically. They are then applied to multi-channel mixture data — artificial simulations as well as spectral NIR powder measurements — to demonstrate discovery after incomplete row-wise curve fitting and column-wise multivariate regression. The analysis shows how the Informative Converse paradox is the basis for selectivity enhancement in multivariate calibration. Data-driven model expansion for statistical multi-response analyses (ANOVA, N-way models etc.) is proposed.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Incomplete knowledge is the rule

Mathematical models of real-world systems are often incomplete, either because they reflect incomplete knowledge or because they are consciously simplified for a given purpose. Incomplete models can be very useful, e.g. to encapsulate knowledge and to study specific aspects of a system. However, when an incomplete mathematical model is fitted to empirical data, unmodeled phenomena can cause grave systematic alias errors in the resulting parameter estimates. This is a general and well known problem in e.g. linear statistical modeling, but it is handled differently in different model types. Some pragmatic modeling methods automatically correct for unidentified interferences by purely data-driven means, for instance within the field of multivariate calibration [1]. More causally oriented methods require conscious, explicit modeling of interferences. This difference warrants clarification. More importantly, there is a need for generic methodology that reveals and describes unanticipated phenomena in data, beyond what is already understood and modeled — i.e. windows into the unknown.

This paper demonstrates the potentially damaging effect of unanticipated phenomena on parameter estimates due to incomplete modeling. But it also shows a way to study these phenomena, and how this is used implicitly in multivariate regression modeling. The methodology to be presented is generic for models fitted by unrestricted linear least squares, and probably for other situations as well. Multivariate calibration of spectroscopic data of chemical samples will be used as example, since it is easy to visualize.

1.2. How does multivariate calibration compensate for unidentified interferences?

In multivariate calibration [1], multichannel input data that are highly non-selective due to more or less unidentified interferences, may be converted into selective output information about analytes. In most cases the interference handling is done implicitly. It would be desirable if unexpected and unidentified interferences could be more easily discovered and characterized from calibration data, because multivariate calibration could then give more valuable contributions to the over-all scientific process of knowledge generation.

Traditional univariate calibration employs a very simple model $c \approx f_1(y)$ relating c, the concentration of the analyte of interest, to y, a given property measured, e.g. light absorbance at a certain wavelength. Once the calibration model $f_1(\cdot)$ has been established, it may be used for predicting c from y in new samples, $\hat{c} = f_1(y)$. The calibration model is often linear and based on a presumed causal relationship between the two: $y \approx f_2(c)$, e.g. in spectroscopy.

To give correct results, such a univariate calibration requires a selective, one-to-one correspondence between c and y. Chemical substances that interfere with measurement y have to be removed *physically* from the samples prior to measurement, otherwise they will masquerade as "analyte signal" and thus cause alias errors in the predicted value $\hat{c}$. In contrast, multivariate calibration [1] allows many selectivity problems to be removed *mathematically*, and thereby allows samples to be

* Centre for Integrative Genetics CIGENE, Norwegian University of Life Sciences, N-1430 Ås, Norway. Tel.: +47 95075025; fax: +47 64970333.
E-mail address: harald.martens@nofima.no.

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.02.007

measured without any physical sample clean-up. Multivariate calibration is therefore used extensively in chemistry and many other fields. Linear calibration models are often used, because of their simplicity and their good approximation ability. Several different linear modeling methods are used for multivariate calibration, but they often give similar results, and they all may be formulated as predictors of the form $\hat{c} = f_1(\mathbf{y})$, where $\mathbf{y}$ $(1 \times q)$ is vector of measurements at q different channels.

Multivariate calibration can provide clean information from dirty data — selective concentration estimates $\hat{c}$ from non-selective measurements $\mathbf{y}$. But what makes it work so well? Why are the concentration predictions $\hat{c}$ free of alias errors, in spite of gross, unidentified interferences in $\mathbf{y}$? And beyond the selectivity enhancement clean-up, can the nature of the "dirt" itself be identified from the data?

Statistically speaking, the multivariate calibration methods work — implicitly or explicitly — by compensating for response similarities between analyte and interferences (unexpected chemical constituents and other chemical or physical effects present). This can be done in different ways, all of which seek to estimate and correct for analyte/interference covariances, in terms of between-variables spectral overlap and/or in terms of between-sample intercorrelations. But statistical covariance correction is difficult to understand, at least for non-statisticians. This paper shows that multivariate residual analysis simplifies the covariance correction and offers additional discovery tools, applicable even in more classical statistical analyses.

1.3. Paradox: what we know the least about can be seen most clearly

The general problem addressed in this paper is the following: Assume that we want to analyze a system that is controlled by several causal phenomena, but where we only anticipate some of these phenomena — the others are totally unexpected. How will the latter affect our ability to quantify and interpret the former? To what extent can we discover and describe even these unexpected phenomena?

This paper shows how new insight about the "dirt" — the unexpected, unidentified interferences — can be obtained from multi-channel data. In fact, it proves that we can get more accurate information about the unexpected interferences, about which we know nothing, than about the expected analytes, about which at least we know the spectra or the concentrations. Moreover, the nature of this accurate, new information about the unexpected interference phenomena is the *converse* of the background knowledge that we used when modeling the observed data. This means that if our knowledge about the analytes consists of *rows* (analyte spectra), then the accurate information about the interferences consists *only* of the converse *columns* (interference concentrations), and vice versa. All other parameters will have alias errors in their estimates. Therefore it is here called the *Informative Converse* (*IC*) paradox.

Fig. 1a) illustrates a simple situation where the IC paradox applies: A set of q different properties (attributes, channels; also called "variables") have been measured in n different samples (locations in time and space, individuals; "objects"), and collected in a data table $\mathbf{Y}$ $(n \times q)$. We expect the variations in the measured signals $\mathbf{Y}_{\text{expected}}$ to have been caused by varying levels of a recognized phenomenon (e.g. a chemical constituent) — or of several such expected phenomena — plus random measurement noise. But unknown to us, there are other, completely unexpected sources of variation $\mathbf{Y}_{\text{unexpected}}$ in the system that also affect the measured data $\mathbf{Y}$. The IC analysis shows what information can and cannot be derived from $\mathbf{Y}$, depending on what is known about the analyte(s) in $\mathbf{Y}_{\text{expected}}$ and on the nature of $\mathbf{Y}_{\text{unexpected}}$.

Technically, the basic IC analysis combines two well-known techniques, linear regression (projection) to model the empirical data $\mathbf{Y}$ in terms of prior knowledge about the analytes, followed by a bi-linear decomposition of the lack-of-fit residuals in $\mathbf{Y}$. The IC analysis to be presented here is generic for additive systems. It will be illustrated in a calibration framework, but applies equally well to other linear regression-based modeling methods used in statistics, such as multi-response analysis of variance (ANOVA).

The IC analysis to be presented here is graphically oriented and mathematically simple — almost a banality — and not even new. To perform Principal Component Analysis (PCA) of residuals in $\mathbf{Y}$ after fitting a model $\mathbf{Y} \approx f(\mathbf{X})$ by e.g. Ordinary Least Squares (OLS) or Partial Least Squares (PLS) regression, has surely been done by numerous researchers in statistics and chemometrics. But the specific interpretation opportunities that the residual PCA offers, does not seem to be well known, or at least not used much in practice. The IC paradox associated with the interpretation of lack-of-fit residuals by PCA was,

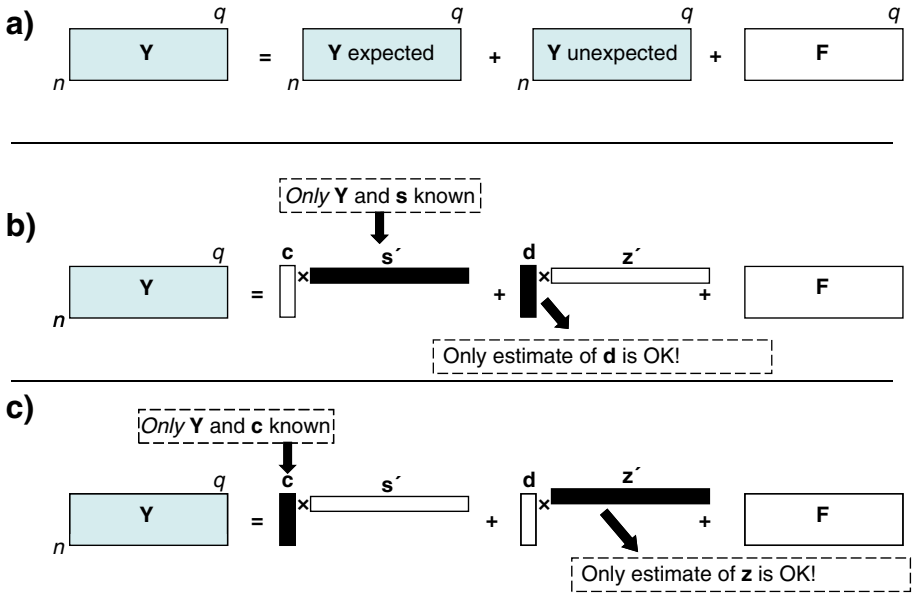

Fig. 1. The Informative Converse, illustrated for the linear two-constituent linear mixture model $\mathbf{Y} = \mathbf{cs}' + \mathbf{dz}' + \mathbf{F}$. a) Mixtures with unanticipated interference problems: multichannel data table $\mathbf{Y}$ is a sum of an expected analyte contribution, an unexpected, but systematic interference contribution and random errors $\mathbf{F}$. b) Hypothesis H1: modeling mixture data $\mathbf{Y}$ by known analyte spectrum $\mathbf{s}$ yields correct interference concentration estimates $\mathbf{d}$, but erroneous estimates of analyte concentrations $\mathbf{c}$ and interference spectrum $\mathbf{z}$. c) Hypothesis H2: modeling mixture data $\mathbf{Y}$ by known analyte concentrations $\mathbf{c}$ yields correct interference spectrum $\mathbf{z}$, but erroneous estimates of analyte spectrum $\mathbf{s}$ and interference concentrations $\mathbf{d}$.

Download English Version:

https://daneshyari.com/en/article/1179901

Download Persian Version:

https://daneshyari.com/article/1179901

Daneshyari.com