Evaluation of empirical chemometric models for calibration and classification

Kurt Varmuza

Vienna University of Technology, Institute of Chemical Engineering, Austria
kvarmuza@email.tuwien.ac.at, www.lcm.tuwien.ac.at

Autumn School of Chemoinformatics, The University of Tokyo, 16 November 2011

1. Introduction

A fundamental task in science and technology is searching for relationships between data sets, e. g., modeling a PROPERTY y by one or several VARIABLES x. Often a desired property cannot be determined directly or only with high cost. In contrary some x-data may be easily available. Depending on how well known and how strictly defined the relationship between x and y is, we can distinguish different levels of creating and applying models that predict y from x [1].

- The relationship is described by a fundamental scientific law (a FIRST PRINCIPLE), formulated as a relative simple mathematical equation with all parameters known. An example is for instance the time, y, a falling stone needs for a given height, x; the gravity constant, g, is known and y can be easily calculated by $(2x/g)^{0.5}$ - if effects like air friction are ignored.

- The relationship is well described by a relatively simple mathematical equation - usually based on physical/chemical knowledge - but the parameters are not known. An example is Lambert-Beer's law. The concentration, c, of a light-absorbing substance is given by $A/(a \cdot L)$ with L being the path length, a the absorption coefficient, and A the absorbance defined by $\log(I_0 / I)$ with I_0 the incident light intensity, and I the light intensity after passing the sample. I_0, I, and L can be measured easily but the absorption coefficient is in general not known. It has to be determined from a set of reference samples with known concentrations and by application of a regression method - a so called calibration procedure. This method becomes very powerful if many wavelengths in the IR or NIR range are used [2], and it is one of the main applications of chemometrics [3-5].

- In many cases of practical interest no theoretically based mathematical equations exist for the relationship between x and y. We sometimes know but often only assume that a relationship exists. In this case we call the model 'EMPIRICAL' or 'DATA DRIVEN'. Examples are for instance modeling of the boiling point or the toxicity of chemical compounds by variables derived from the chemical structures (MOLECULAR DESCRIPTORS). Development of quantitative structure-property or structure-activity relationships (QSPR, QSAR) by this approach requires multivariate calibration methods. For such purely empirical models - often with many x-variables - the COMPLEXITY of the model and the PREDICTION PERFORMANCE have to be estimated very carefully and cautiously. Also the variability of used measures has to be estimated. Typical for problems in chemistry are a small number of cases (objects) and a large number of x-variables.

2. Multiple linear models

We focus here on linear models of the form

$$y = f(x_1, x_2, ..., x_m) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m = b_0 + \boldsymbol{b}^T \boldsymbol{x}$$

in which a property y is modeled by a set of m x-variables, with b_0 the INTERCEPT (zero for mean-centered data), b_j (j = 1 ... m) the REGRESSION COEFFICIENTS (forming vector $\boldsymbol{b}$), x_j (j = 1 ... m) the variables (forming vector $\boldsymbol{x}$) [6]. Aim of model creation is finding a vector $\boldsymbol{b}$ from a set of n_{CALIB} cases (calibration objects) that gives low errors (residuals) $e = y - \hat{y}$ (y is the given/true value of a property, $\hat{y}$ is the value estimated by the model) for cases not used for the estimation of $\boldsymbol{b}$. Usually, a good fit for the data of the CALIBRATION SET is not sufficient but an optimum performance for an independent TEST SET is required [7].

All regression methods aim at the minimization of residuals. The simple standard method ORDINARY LEAST SQUARES (OLS) regression estimates $\boldsymbol{b}$ directly as $(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$. This equation holds for mean-centered data, with matrix $\boldsymbol{X}$ ($n_{CALIB} \times m$) for the calibration set data, and vector $\boldsymbol{y}$ with the properties of the n_{CALIB} objects [8]. OLS is not applicable for most data sets typical in chemistry because the matrix inversion cannot be performed for data with $m>n$ or highly correlating variables as is often the case in chemistry. Furthermore, the complexity of the model cannot be optimized with OLS.

Most powerful methods like PRINCIPAL COMPONENT REGRESSION (PCR) or PARTIAL LEAST-SQUARES REGRESSION (PLS) work with a small set of intermediate linear latent variables (COMPONENTS). This approach overcomes the above mentioned drawbacks of OLS. For PCR the x-data are transformed by PRINCIPAL COMPONENT ANALYSIS (PCA). The first PCA component is a linear combination of all variables, so that the values (scores) of this new variable have maximum variance. Further PCA components are orthogonal and again possess maximum possible variance. These new variables (PC scores) are uncorrelated and a set of them is used for OLS. PLS is less strictly defined and various methods are implemented in software; the basic concept of PLS are components with maximum covariance between the scores and the modeled y.

The NUMBER OF USED COMPONENTS (PCA or PLS) determines the complexity of a model and has to be optimized. Fig. 1 explains this essential aspect for empirical models. The more complex a model is (the more components are considered), the better is the fit for the calibration set data. Thus, the prediction error for the calibration set in general decreases with increasing complexity of the model. An appropriate highly complicated model would fit almost any data with almost zero residuals. It is evident that such model is not necessarily useful for new cases, because it is probably OVERFITTED; that means it is well adapted to the calibration data but do not include sufficient GENERALIZATION. The prediction errors for new cases (test set, objects not used in model generation) mostly show a minimum at a medium model complexity. Complexity (number of independent parameters of a model) becomes higher e. g., by (1) increasing the number of variables, (2) adding derived variables such as nonlinear functions of original variables, (3) implementing nonlinearities into the model, (4) increasing the number of components in PLS and PCA. Determination of the optimum complexity of a model is an important but not always easy task. It should be performed by using only calibration data but not test set data.

Note that for PCR or PLS - independent from the number of considered components - the final model $y = f(x_1, x_2, ..., x_m)$ contains all variables. Variable selection is another topic not discussed here.

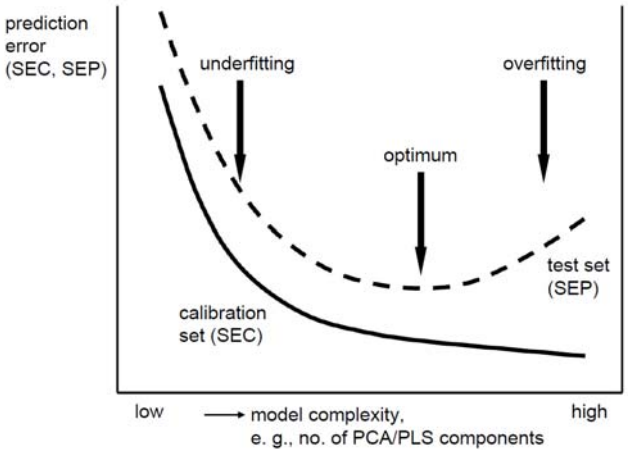

Fig. 1. Model complexity (number of PCA or PLS components) versus prediction error for calibration set and for test set (schematically).

3. Performance measures for calibration and classification models

In CALIBRATION MODELS a continuous property is modeled and the residuals (prediction errors), $e = y - \hat{y}$, obtained from test set objects have to be used to estimate the prediction performance and its variability. The distribution of e_i ($i = 1 \ldots n_{TEST}$) is often similar to a normal distribution and then can be characterized by its mean and standard deviation. This mean is usually called BIAS and should be near zero; the standard deviation is usually called STANDARD ERROR OF PREDICTION (SEP) - if calculated for test set object. In the case the prediction errors are approximately normal distributed about 95% of the prediction errors are within the tolerance interval $\pm 2 \cdot$ SEP. The measure SEP and the tolerance interval are given in the units of y, and are therefore most useful for model applications.

The STANDARD ERROR OF CALIBRATION (SEC) is identical to SEP but calculated from prediction errors of the calibration set object. SEC is usually a too optimistic estimation of the prediction errors for new cases. Other performance criteria are e. g., MEAN SQUARED ERROR (MSE), the arithmetic mean of the squared errors; ROOT MEAN SQUARED ERROR (RMSE), the square root of MSE; and PREDICTED RESIDUAL ERROR SUM OF SQUARES (PRESS), the sum of the squared errors. All these criteria can be calculated for calibration sets (often within a cross validation) or test sets - a clear definition is essential.

Widely used is the SQUARED PEARSON CORRELATION COEFFICIENT, R^2, between y, the given/true values, and $\hat{y}$, the predicted values. R^2 near 1 indicates a good model; however, a visual inspection of diagnostic plots $\hat{y}$ versus y or e versus y is highly recommended. For a comparison of models containing different numbers of variables, m, the ADJUSTED SQUARED CORRELATION COEFFICIENT

$$_{ADJ}R^2 = 1 - (n - 1)(1 - R^2) / (n - m - 1)$$

is convenient because it penalized models with a large number of variables; n is the number of objects [1].

For CLASSIFICATION MODELS the fraction (or percent) correctly classified objects is an evident performance measure. It is essential to estimate these numbers separately for each class (PREDICTIVE ABILITIES OF CLASSES, P_1, P_2, ...). The mean of them, P, AVERAGED PREDICTIVE ABILITY, may be used as a single number that characterizes the overall performance. However, the proportion of all correct classifications (all classes together) should be avoided because it depends on the relative frequencies of the classes and may be misleading.

4. Evaluation of models

Empirical models should be evaluated carefully and cautiously (t. m., conservative) because a problem-relevant interpretation of the model parameters is often not possible and the model has to be used as a gray box. The two main aspects in model generation are as follows.

- Estimation of the OPTIMUM MODEL COMPLEXITY (more general, the optimum for one or several model parameters) in order to avoid underfitting and overfitting, but aim at a model which is optimal for new cases. Typically such parameters are the number of PLS components, or the number of neighbors in KNN classification.

- Estimation of the MODEL PERFORMANCE for new cases (not used for model optimization and creation). Typical measures are SEP or the averaged predictive ability.

For a proper model generation and evaluation these two goals must be performed independently. In other words, no performance measures obtained during optimization of the model complexity should be finally used; and no adjustment of the complexity should be done based on results from test sets. The available data have to be split by some strategy into a calibration set and a test set. Furthermore, the distribution or a confidence interval should be given for the optimum complexity and for the performance measure; thus repetitions of the evaluation are needed with different random splits (essential for comparing models).

Two re-sampling strategies are common for a repeated and separate estimation of the optimum model complexity and of the model performance. Both are applicable to data with a rather low number of objects (however, ca >20). One strategy is the DOUBLE BOOTSTRAP based on sampling with replacement [9]. The other strategy is REPEATED DOUBLE CROSS VALIDATION (rdCV) [10], combining systematic manner and randomness. Several applications of rdCV in chemistry have been reported [11,12] and free software is available [13,14] for the programming environment ℝ [15].

[1] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, 2009.
[2] T. Naes, T. Isaksson, T. Fearn, T. Davies, A user-friendly guide to multivariate calibration and classification, NIR Publications, Chichester, United Kingdom, 2004.
[3] R.G. Brereton, Chemometrics - Data analysis for the laboratory and chemical plant, Wiley, Chichester, United Kingdom, 2006.
[4] D.L. Massart, B.G.M. Vandeginste, L.C.M. Buydens, S. De Jong, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics: Part A, Elsevier, Amsterdam, The Netherlands, 1997.
[5] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. De Jong, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics: Part B, Elsevier, Amsterdam, The Netherlands, 1998.
[6] B.F.J. Manly, Multivariate statistical methods: A primer, Chapman and Hall, London, United Kingdom, 2000.
[7] T. Hastie, R.J. Tibshirani, J. Friedman, The elements of statistical learning, Springer, New York, NY, 2001.
[8] B. Flury, H. Riedwyl, Multivariate statistics: A practical approach, Chapman & Hall, Boca Raton, FL, USA, 1988.
[9] B. Efron, R.J. Tibshirani, An introduction to the bootstrap, Chapman & Hall, London, United Kingdom, 1993.
[10] P. Filzmoser, B. Liebmann, K. Varmuza, Repeated double crosss validation, J. Chemometr. 23 (2009) 160-171.
[11] Y. Felkel, N. Dörr, F. Glatz, K. Varmuza, Determination of the total acid number (TAN) of used gas engine oils by IR and chemometrics applying a combined strategy for variable selection, Chemom. Intell. Lab. Syst. 101 (2010) 14-22.
[12] B. Liebmann, A. Friedl, K. Varmuza, Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics, Anal. Chim. Acta 642 (2009) 171-178.
[13] P. Filzmoser, K. Varmuza, R package chemometrics, http://cran.at.r-project.org/web/packages/chemometrics/index.html, Vienna, Austria, 2010.
[14] K. Varmuza, P. Filzmoser, B. Liebmann, R software for chemometrics and chemoinformatics, http://www.lcm.tuwien.ac.at/R, Vienna, Austria, 2011.
[15] R, A language and environment for statistical computing, R Development Core Team, Foundation for Statistical Computing, www.r-project.org, Vienna, Austria, 2011.