Pattern Recognition Letters 65 (2015) 152–156

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Stochastic complexity-based model selection with false alarm rate control in optical spectroscopy☆

Julien Fade*

Institut de Physique de Rennes, Université de Rennes 1, CNRS, Campus Beaulieu, Rennes 35042, France

ARTICLE INFO

Article history:
Received 14 March 2015
Available online 31 July 2015

Keywords:
model selection
stochastic complexity
hypothesis testing

ABSTRACT

Stochastic complexity-based penalization criteria can prove efficient and robust in spectroscopy applications for unsupervised identification and concentration estimation of spectrally interfering chemical components. It is shown here how the so-called Normalized Maximized Likelihood (nMDL) introduced in [17] can be tailored to provide control of the detection performances in terms of probability of false alarm. Numerical experiments conducted on realistic simulated optical spectroscopy signals evidence that the nMDL approach outperforms standard information criteria in terms of model selection performances. Moreover, the ability to control false alarm rates with the proposed modified nMDL criterion is demonstrated on simulations.

1. Introduction

Model selection is a very common issue in the field of data analysis and regression. Many questions usually have to be answered simultaneously: does the data reveals the presence of a significant feature (regressor), or not? If so, how many and which regressors must be selected in the linear regression model to better explain the observations? Without any model selection step, the most exhaustive regression model would include all potential regressors, and may lead to misleading and imprecise (if not incorrect) results, mostly due to overfitting of the noise. This can have harmful consequences in some situations, for instance in the context of trace gas detection by optical spectroscopy as addressed by this paper as an illustration.

To avoid such undesirable situations, and to provide unsupervised model selection strategies, many penalized regression methods have been proposed using various penalization criteria, such as the classical information criteria (Mallow's C_p [13], Akaike's AIC and variants [1,10], Schwarz's BIC [20], RIC [5], etc.). However, many "best" penalization criteria have been introduced in the literature to refine these standard selection rules, so as to optimize the quality of model selection depending on the problem at hand [22]. This raises the question of the generality of such penalization strategies. In that context, since its introduction by Rissanen's seminal work [15], the Minimum Description Length (MDL) principle is an interesting and fruitful attempt to build a general theoretical framework to interpret model complexity and to provide unsupervised model selection rules. The MDL principle states that the best description of the data must be given by the model leading to the shortest code length, or stochastic complexity (SC) (expressed in bits or in *nats* (1 *nat*= $\ln 2$ bits)) required to describe both the model and the data [9,15,18]. The MDL principle has found wide applicability in very distinct contexts, such as model selection [9], data clustering [8], but also radar signal processing [3] or image segmentation [6].

In this paper, we focus on a sophisticated form of the SC, referred to as Normalized Maximized Likelihood (nMDL) [17]. We show how this penalization criterion can be modified so as to provide probability of false alarm (Pfa) control in the context of model selection, and we illustrate this property on realistic numerical simulations of an unsupervised optical spectroscopy experiment. The paper is organized as follows: in Section 2, the expression of the nMDL criterion is first recalled, and a modified version of this criterion is derived, allowing to control the Pfa in the model selection procedure. Then, a simulated experiment of optical spectroscopy is described in Section 3 and numerical simulation results allow us to compare the quality of the standard nMDL criterion with respect to more standard information criteria. The possibility of Pfa control using the proposed modified nMDL criterion is finally illustrated on simulated data, before the conclusion of the paper is given in Section 4.

2. Normalized-Maximized Likelihood (nMDL) criterion and false alarm rate control

Throughout this paper, we shall consider the simple problem of linear regression, with m-dimensional observation vector $\tilde{\mathbf{y}}$ modeled as

$$\tilde{\mathbf{y}} = \mathbf{H} \cdot \mathbf{c} + \mathbf{n}, \tag{1}$$

☆ This paper has been recommended for acceptance by Dr. G. Moser.

* Tel.: +33 223 235 215; fax: +33 223 236 717.
E-mail address: julien.fade@univ-rennes1.fr

http://dx.doi.org/10.1016/j.patrec.2015.07.022
0167-8655/

with $m \times k$ regressor matrix $\mathbf{H}$ and unknown parameters vector $\mathbf{c}$. We further assume the m components of the additive noise vector $\mathbf{n}$ to be independent realizations of a centered Gaussian random variable. Under this hypothesis, applying a penalized criterion in the model selection procedure corresponds to minimize the following quantity:

$$-\ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathbf{H}) + \mathcal{C}, \tag{2}$$

where the expression of $\mathcal{C}$ depends on the penalization criterion used (AIC, BIC, etc.), and where the log-likelihood is directly related to the regression residual sum of squares (RSS) through $\ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathbf{H}) = -(m/2) \ln \mathrm{RSS}$.

2.1. nMDL criterion

The Normalized Maximized Likelihood (nMDL) is a recently introduced form of SC [17] which has proved efficient in various practical problems [4,9,18] and which presents various optimality properties [18]. The nMDL theory suggests the following penalization terms to be introduced in the criterion given in Eq. (2), depending on the hypothesis considered [18]:

$$\mathcal{C}^{(n)}_{|\mathcal{H}_0} = \frac{m}{2} \ln \pi - \ln \Gamma\left(\frac{m}{2}\right) + \ln \ln \frac{b}{a} \tag{3}$$

and

$$\mathcal{C}^{(n)}_{|\mathcal{H}_1^{(k)}} = \frac{k}{2} \ln \frac{kF}{m-k} + \frac{m}{2} \ln \pi - \ln \left[\Gamma\left(\frac{k}{2}\right) \Gamma\left(\frac{m-k}{2}\right) \right] + \mathcal{L}_k + 2 \ln \ln \frac{b}{a}, \tag{4}$$

where hypothesis $\mathcal{H}_0$ refers to the null-model (no regressor) and hypothesis $\mathcal{H}_1^{(k)}$ to a selected model containing k regressors among K_m potential regressors. In this last equation, F denotes the standard F-ratio, which depends on the RSS of the regression through $F = (m-k)[\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - RSS]/(k\, RSS)$. The code length $\mathcal{L}_k$ needed for encoding model $\mathcal{H}_1^{(k)}$ is given by $\mathcal{L}_k = \ln \binom{K_m}{k} + \ln k + \log_2 \ln (e\, K_m)$ [18]. Lastly, it must be noted that the nMDL approach requires two hyperparameters a and b to be estimated. According to indications in [18], they can be respectively estimated with the regression sum of squares (i.e., $\tilde{\mathbf{y}}^T \tilde{\mathbf{y}} - RSS$) obtained with the most exhaustive model $\mathcal{H}_1^{(K_m)}$ and the regression sum of squares under null hypothesis $\mathcal{H}_0$.

As with standard information criteria, model selection is finally carried out by identifying the set of regressors which minimizes the penalized criterion. This operation can be performed by exhaustive search, or by appropriate stepwise procedures, in which the number of regressors is gradually increased until no further decrease of the criterion can be reached. Such a stepwise method will be used in Section 3.

In the remainder of this section, we show and illustrate how the nMDL criterion can be tailored so as to provide control of the false alarm rate (Pfa) in a model selection procedure.

2.2. Pfa control and model selection

The link between detection performance and model selection is a known result for standard information criteria. For instance, it is quite straightforward to understand that using the AIC or BIC criteria for model selection is equivalent to applying a standard generalized likelihood ratio test (GLRT) with given threshold depending on the penalization considered [23]. For example, to discriminate between hypothesis $\mathcal{H}_0$ and hypothesis $\mathcal{H}_1$ (at least one regressor included in the selected model), this decision rule can be summarized as

$$\ell_{glrt}(\tilde{\mathbf{y}}) = \ln \left[\frac{P(\tilde{\mathbf{y}}|\mathcal{H}_1)}{P(\tilde{\mathbf{y}}|\mathcal{H}_0)} \right] = \ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathcal{H}_1) - \ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathcal{H}_0) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau, \tag{5}$$

where the value of the threshold τ fixes the Pfa. For a Gaussian noise model, one simply has

$$\ell_{glrt}(\tilde{\mathbf{y}}) = \frac{m}{2} \ln \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}. \tag{6}$$

A similar property has been recently analyzed in the case of the nMDL criterion in [7], where the authors evidenced that the application of the nMDL criterion is formally equivalent to a GLRT with fixed threshold. In the following, we show how to exploit this property so as to control the Pfa in a model selection procedure. This is made possible by introducing a slightly modified version of the nMDL criterion.

2.3. Thresholded nMDL criterion for Pfa control

Although one of the main concerns that underlies MDL approaches is to minimize the number of user-defined parameters in the criterion, we propose to introduce a fixed threshold in the application of the nMDL criterion. Meanwhile, this avoids resorting to the hyperparameters a and b included in the former criterion, which is easily obtained by setting $a = be$ in Eqs. (3) and (4). Such modified nMDL criterion for discrimination between null/non-null hypotheses leads to the following decision rule:

$$\begin{aligned} \Delta\mathcal{C}^{(n)}(\tilde{\mathbf{y}}, m, k) &= -\ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathcal{H}_1^{(k)}) + \mathcal{C}^{(n)}_{|\mathcal{H}_1^{(k)}} + \ell_{\tilde{\mathbf{y}}}(\tilde{\mathbf{y}}; \mathcal{H}_0) - \mathcal{C}^{(n)}_{|\mathcal{H}_0} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \delta_{\mathrm{Pfa}} \\ \Leftrightarrow\ \Delta\mathcal{C}^{(n)}(\tilde{\mathbf{y}}, m, k) &= -\ell_{glrt}(\tilde{\mathbf{y}}) + \mathcal{C}^{(n)}_{|\mathcal{H}_1^{(k)}} - \mathcal{C}^{(n)}_{|\mathcal{H}_0} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \delta_{\mathrm{Pfa}}, \end{aligned} \tag{7}$$

illustrating that the nMDL criterion will reject hypothesis $\mathcal{H}_0$ if $\Delta\mathcal{C}^{(n)}(\tilde{\mathbf{y}}, m, k)$ is lower than a given threshold.

Let us now analyze the relationship between the Pfa and the value of the user-defined threshold δ_{Pfa}. First, by noticing that $\ell_{glrt} = m/2 \ln[1 + kF/(m-k)]$, a few calculation steps allow us to show that the quantity $\Delta\mathcal{C}^{(n)}$ can be rewritten as a function of the generalized log-likelihood test (GLRT) ℓ_{glrt} as (see [7])

$$\Delta\mathcal{C}^{(n)}(\ell_{glrt}, m, k) = g(\ell_{glrt}, m, k) + \eta(m, k) + \mathcal{L}_k, \tag{8}$$

with $\eta(m, k) = -\ln B[k/2, (m-k)/2]$, where the Beta function reads $B(x, y) = [\Gamma(x)\, \Gamma(y)]/\Gamma(x+y)$, and with

$$g(x, m, k) = -x + \frac{k}{2} \ln[e^{2x/m} - 1]. \tag{9}$$

$\mathcal{L}_k$ still denotes the code length needed to encode the model. Using Stirling's approximation of the Beta function, this modified nMDL criterion can be rewritten $\Delta\mathcal{C}'^{(n)}(\ell_{glrt}, m, k) = g(\ell_{glrt}, m, k) + \eta'(m, k) + \mathcal{L}_k$, with

$$\eta'(m, k) = \frac{1}{2} \left[f(m) - f(m-k) - f(k) + \ln \left[\frac{k(m-k)}{4\pi m} \right] \right] \tag{10}$$

and $f(x) = x \ln(x)$.

From this expression, it can first be observed that $\Delta\mathcal{C}'^{(n)}(\ell_{glrt}, m, k)$ does not evolve monotonously as a function of ℓ_{glrt}. Nevertheless, following a similar reasoning as in [7], it can be shown that the function $\Delta\mathcal{C}'^{n}(\ell_{glrt}, m, k)$ is concave and takes on positive values when ℓ_{glrt} lies within an interval $[\ell^-_{glrt}, \ell^+_{glrt}]$ (see Fig. 1), $\forall k \in [1; m]$, as soon as $K_m > 2$. It is now quite obvious that the application of the modified (thresholded) nMDL decision rule given in Eq. (7) corresponds to a fixed value of the false alarm rate, which is equal to the probability that ℓ_{glrt} lies outside the interval $[\ell^-_{\delta_{\mathrm{Pfa}}}; \ell^+_{\delta_{\mathrm{Pfa}}}]$ when hypothesis $\mathcal{H}_0$ is true, i.e.,

$$\mathrm{Pfa} = \Pr(\ell_{glrt} < \ell^-_{\delta_{\mathrm{Pfa}}} | \mathcal{H}_0) + \Pr(\ell_{glrt} > \ell^+_{\delta_{\mathrm{Pfa}}} | \mathcal{H}_0). \tag{11}$$

This is illustrated in Fig. 1 where the obtained Pfa corresponds to the darkened areas under the red dashed curve representing the probability density function (pdf) of the log-likelihood ratio ℓ_{glrt} under hypothesis $\mathcal{H}_0$. From this last relation, it is now clear that provided the