Publ. Astron. Obs. Belgrade No. 80 (2006), 81 - 84 *Contributed paper*

AN ALTERNATIVE APPROACH TO SPECTRUM BASE LINE ESTIMATION

SRDJAN BUKVIĆ and DJORDJE SPASOJEVIĆ

Facilty of Physics, Studentski Trg 12-16, 11000 Belgrade, Serbia and Montenegro
E–mail: ebukvic@ff.bg.ac.yu

Abstract. We present a new form of merit function which measures agreement between a given data set and the model function with a particular choice of parameters. Proposed merit function is functional compound of recently introduced Close Points Concept and ordinary least-squares approach. Essentially it measures a density of ordinary least-squares for an arbitrary data set. New merit function is insensitive on outlying points and can be applied to data sets containing no more than few points. Specifics of the proposed merit function, including scale independent behavior, are detailed in several examples. We illustrate efficiency of the presented merit function on the common problem of finding the base line of a spectrum. In particular, when the base line represents continuum radiation, estimation of thermodynamic temperature is considered.

1. INTRODUCTION

The aim of the present paper is to put forward a simple and efficient alternative method for robust estimation of a model function parameters, insensitive to outlying points. Proposed method is suitable in general case, regardless to the number of points in the data set.

2. MERIT FUNCTION

Let us consider a data set of x_i's and y_i's $(i = 1, 2, ...n)$ and a family of model functions $y(x; a_1, a_2, ...a_m)$ depending on x and on the parameters $a_1, .., a_m = \mathbf{a}$, specifying a particular model function from the family. We introduce the *density of the least squares* D_s for a given data set s in the following way:

$$D_s(\mathbf{a}_{ols}) = \frac{\sum\limits_s d_i^2}{d_{max}^k}. \tag{1}$$

Here $\sum\limits_s d_i^2 = \sum\limits_{i=1}^{n} [y_i - f(x_i; \mathbf{a}_{ols})]^2$ is the sum of squared deviations if the model function is calculated with parameters $\mathbf{a} = \mathbf{a}_{ols}$ obtained by ordinary least-squares

(OLS) method. In other words $\sum_s d_i^2$ is the ordinary least-squares sum for a given data set s. With $d_{max} = |y_j - f(x_j, \mathbf{a_{ols}}|_{max}$ we denoted maximum deviation from the model function achieved for some point j. d_{max} specifies *width* of the data set in respect to the model function. Below we will give arguments that natural choice for value of the exponent is $k = 2$.

Therefore, Eq. (1) defines generalized density, D_s, of the least-squares for a given data set s. Let us consider a data set containing n points with *one* outlying point. It is intuitively clear that quantity D_s for the whole data set will be less then D_{ss} for the corresponding *subset* with outlying point removed, because d_{max} in denominator is significantly less if there is no outlying point. We will establish our merit function relying on the following simple fact: *if* the data set s contains outlying points then exists *subset* ss, with outlying points removed, for which density D_{ss} is higher than D_s. Our merit function results from the above analysis and has the following form:

$$\chi = -D_{ss}(\mathbf{a}_{ols}) = -\frac{\sum_{ss} d_i^2}{d_{max}^k}. \tag{2}$$

In relation (2) we *don't seek parameters* $\mathbf{a}$ which minimize value of χ. Instead, we need to *find subset* ss for which density of the least-squares D_{ss} is maximum. This task can not be reduced to standard minimization problem.

We will propose a simple strategy to solve this problem: for a given data set or subset we calculate the best fit model function according to OLS and locate a point (or points) with $d_{max} = |y_j - f(x_j, \mathbf{a_{ols}}|_{max}$, subsequently removing this point. We repeat the same procedure with remaining points, preserving initially obtained value d_{max}, until all points with $d_i \geq d_{max}$ are not removed. Points eliminated applying this procedure represent a *layer of outer points* or, simply, a *layer*. Now, starting from a given data set s we can iteratively remove layer by layer obtaining sequence of data subsets ss with appropriate D_{ss} associated to each. We expect that for certain subset, D_{ss} will attain the maximum value. This subset we will call the best subset, points belonging to the best subset we will call the *close points* while removed points are *distant points* or outliers in respect to the given model function. The best fit coefficients $\mathbf{a}_b$, together with corresponding errors $\Delta\mathbf{a}_b$ and width $d_{max}^{best} \equiv d_b$, one can obtain by applying OLS again, to the close points only. Presented approach we will call Close Points Concept Least-Squares (CPCLS) to emphasize the origin of the merit function (2), (see Bukvić and Spasojević, 2005).

3. CALIBRATION PROCEDURE

Suppose that specific data set is given with really large number of points, n_0, normally distributed around value $y = 0$, i.e. $n(y) \sim n_0 \cdot \exp(-\frac{y^2}{2\sigma^2})$, where σ is a standard deviation. For this particular distribution we don't need to remove layer by layer, we can directly calculate generalized density of the least-squares, D_{ss}, for arbitrary width d in respect to the model function $y = 0$:[1]

[1]We have omitted index max for width d of the data set.

$D_{ss} = \frac{f}{d^k}$
where
$f \sim n_0 \cdot \int_0^d y^2 n(y) dy$.
D_{ss} will have a maximum if $\partial D_{ss}/\partial d = 0$, or:

$$k = \frac{f\,'}{f}\, d = \frac{\frac{1}{\sqrt{\pi}} \exp\left(-\frac{d^2}{2\sigma^2}\right)\left(\frac{d^2}{\sigma^2} - 1\right) + \frac{1}{2}\,\mathrm{erf}\,'\left(\frac{d}{\sigma\sqrt{2}}\right)}{-\frac{1}{\sqrt{2\pi}}\frac{d}{\sigma} \cdot \exp\left(-\frac{d^2}{2\sigma^2}\right) + \frac{1}{2}\,\mathrm{erf}\left(\frac{d}{\sigma\sqrt{2}}\right)} \cdot \frac{d}{\sigma\sqrt{2}}\,. \tag{3}$$

Eq. (3)[2] relates exponent k with a single variable d/σ. Note that set of normally distributed data has no distinct outlying points, instead we need to decide which points will be considered as outlying/close points. It is common to accept as a boundary value $d = \sigma$. Consequently, points with $y_i < \sigma$ will be considered as the close points while points with $y_i > \sigma$ are outlying points. Inserting $d = \sigma$ in Eq. (3) one can obtain corresponding value of the exponent k, i.e. $k \approx 2.44$.

Removing layer by layer from the initial data set, sooner or latter we will have to deal with data subsets containing small number of points. Consider the following example: data set of $n + 1$ points is given. n points belong to the same horizontal straight line while one is out of the line. Our model function is a straight line also. We immediately find that first layer contains outlying point. When we remove this layer, n points remain laying on the same straight line. Note that our merit function (2) is undetermined in that case. Both, $\sum d_i$ and d_{max} are equal to zero. However, it should be underlined that our points are not exact, they are result of the measurement performed with a limited accuracy. Therefore, our n points are not necessarily on the same line and we must admit presence of some unknown sub distribution below the resolution limit of measurement. For our purpose it is sufficient to characterize this sub distribution with average deviation of data points from model function. Suppose that average deviation $d_1 = d_2 =, ..., = d_n = \frac{d_{max}}{\sqrt{m}}$ where m is a small number. It follows that for remaining n points

$$D_{ss} \equiv D_n = \frac{\frac{n d_{max}^2}{m}}{d_{max}^k}. \tag{4}$$

It is obvious that choice $k = 2$ simplifies above relation and we obtain: $D_n = n/m$ where n is the number of points belonging to the straight line, while m is a subject to choice. Note that in our example $D_{n+1} = (1 + n)/n$. Thus, a choice $m = q$ ensures that single outlying point will be recognized as an outlier if the number of points belonging to the straight line satisfies $n/q > (n+1)/n$. Otherwise, the single outlying point will be included in the set of $n + 1$ close points.

It is simple to show that $k = 2$ is the most suitable choice for arbitrary model function. According to Eq. (3) it will introduce $\sim 1.4\sigma$ equivalent width of the best subset if applied on a data set with large number of normally distributed points.

[2]Here, $\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ is the standard error function, while its first derivative is $\mathrm{erf}\,'(x) = \frac{2}{\sqrt{\pi}} \exp(-x^2)$.

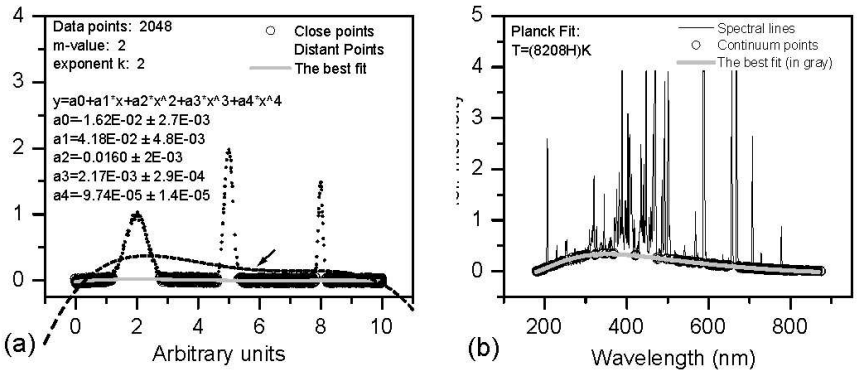

Figure 1: (a) Typical spectrum base line and (b) continuum processed by CPCLS.

4. EXAMPLES

In Fig. 1a an artificially generated spectrum with three spectral lines is shown. Magnitude of the applied Gaussian noise is $\sigma = 0.04$. Full line in Fig 1a represents the best fit polynomial according to CPCLS. It is interesting to note that width of the best subset $d_b = 0.039$ coincides well with the applied Gaussian noise. This value can serve as a discrimination level, points with $d < d_b$ are included in the noise while the points with $d > d_b$ are formally speaking outliers, in our case these points belong to the spectral lines.

Finally we will consider the real world example, a high voltage pulsed discharge spectrum obtained by miniature Ocean Optics spectrograph. Created argon plasma is in a state close to the thermodynamic equilibrium, (see Djeniže and Bukvić, 2001) generating spectrum with number of spectral lines and strong continuum. Appropriate model function follows from the Planck's law of black body radiation. Within this simple example we will use Planck's formula without any corrections in the following way: $I(\lambda, T) = \frac{c_1}{\lambda^5} \cdot \frac{1}{\exp(\frac{hc}{\lambda kT})-1} + c_2$, where parameter c_1 comprises all fundamental constants present in Planck's law, including unknown sensitivity of the spectrograph while parameter c_2 takes care of the instrumental offset. The third fit parameter, T, is thermodynamic temperature. Other symbols have their usual meaning. In Fig. 1b we present the best fit Planck's function according to the proposed methodology. The estimated temperature is approximately 8200K.

Acknowledgements. This work was supported by the Serbian Ministry of Science and Environmental Protection under projects 141010 and 141014.

References

Bukvić, S., Spasojević, Dj.: 2005, *Spectrochim. Acta B*, **60**, 1308-1315.
Djeniže, S., Bukvić, S.: 2001, *Astron. Astrophys.*, **365**, 252-257.