

# Analysing spectral lines in *Gaia* low-resolution spectra

M. Weiler<sup>1,2,3</sup>, J. M. Carrasco<sup>1,2,3</sup>, C. Fabricius<sup>1,2,3</sup>, and C. Jordi<sup>1,2,3</sup>

<sup>1</sup> Departament de Física Quàntica i Astrofísica (FQA), Universitat de Barcelona (UB), c. Martí i Franquès 1, 08028 Barcelona, Spain

e-mail: mweiler@fqa.ub.edu

<sup>2</sup> Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (UB), c. Martí i Franquès, 1, 08028 Barcelona, Spain

<sup>3</sup> Institut d'Estudis Espacials de Catalunya (IEEC), c. Gran Capità 2–4, 08034 Barcelona, Spain

Received 18 August 2022 / Accepted 12 November 2022

## ABSTRACT

**Context.** With its third data release, European Space Agency's *Gaia* mission has published the first set of low-resolution spectra for a large number of celestial objects. However, these spectra differ in their nature from typical spectroscopic data, as they do not consist of wavelength samples with associated flux values. Instead, they are represented by a linear combination of Hermite functions.

**Aims.** We derive an approach to studying spectral lines that is robust and efficient for spectra that are represented as a linear combination of Hermite functions.

**Methods.** For this purpose, we combined established computational methods for orthogonal polynomials with the peculiar mathematical properties of Hermite functions and basic properties of the *Gaia* spectrophotometers. In particular, we made use of simple computations for the derivatives of linear combinations of Hermite functions and their roots.

**Results.** A simple and efficient computational method for deriving the position in wavelength, statistical significance, and line strengths is presented for spectra represented by a linear combination of Hermite functions. The derived method is fast and robust enough to be applied to large numbers of *Gaia* spectra without the need for high-performance computing resources or human interference. We present example applications to hydrogen Balmer lines, He I lines, and a broad interstellar band in *Gaia* DR3 low-resolution spectra.

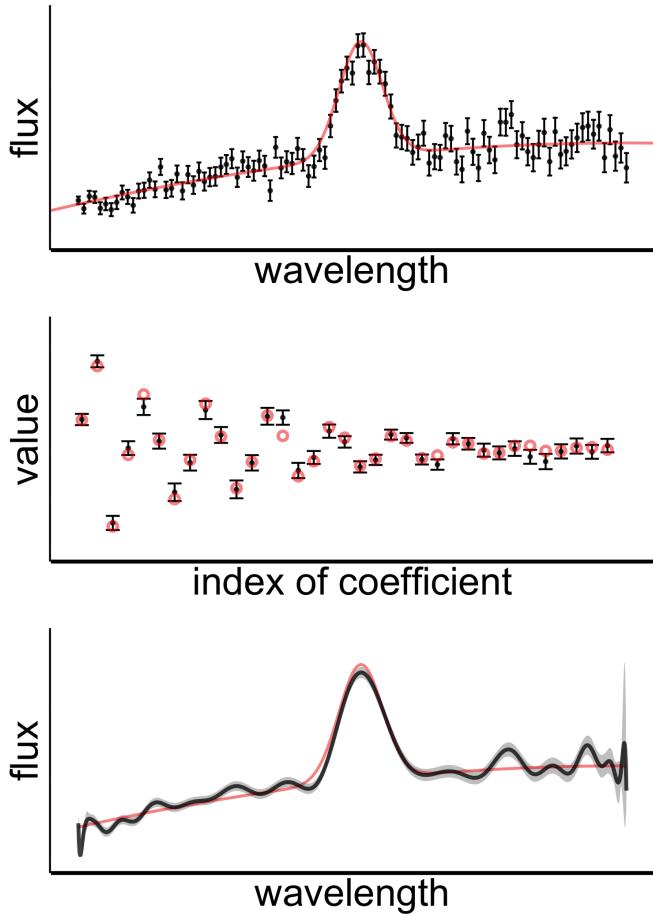
**Key words.** techniques: spectroscopic – methods: data analysis – catalogs – surveys

## 1. Introduction

The European Space Agency's *Gaia* mission (Gaia Collaboration 2016) repeatedly scans the entire sky and observes point-like or nearly point-like objects at magnitudes from approximately  $G = 3^m$  to  $G = 20.7^m$ . These observations include aperture-free low-resolution spectrophotometry covering a wavelength range from approximately 330 nm to 1050 nm, with a spectral resolution of about 30 to 100 (Carrasco et al. 2021). The third *Gaia* Data release (*Gaia* DR3, Gaia Collaboration 2023) includes these spectrophotometric observations for about 219 million astronomical objects (De Angeli et al. 2023). The purpose of these spectra is to lend support to the calibration process for the astrometric observations of *Gaia* and to characterise the observed astronomical objects (Gaia Collaboration 2016). Spectroscopy at very low-resolution is generally not well suited for the study of spectral lines, as the contrast between the line and the underlying continuum is poor, the shape of the lines is not or only partially resolved, and blending of different lines might be strong. However, the exceptional properties of the *Gaia* spectrophotometry, such as the all-sky coverage from a stable space environment, its large degree of completeness, down to  $G = 17.65^m$  in *Gaia* DR3, and the wide wavelength range covered, motivates efforts for analysing spectral lines in *Gaia*'s low-resolution spectra as well. Eventually, *Gaia*'s final data release will comprise hundreds of low-resolution epoch spectra collected over several years, along with a completeness to about  $G = 20.7^m$ .

As a consequence of the self-calibration strategy for *Gaia* spectra, they are represented by a linear combination of basis functions (Carrasco et al. 2021). This distinguishes *Gaia* low-resolution spectra from typical spectra, which consist of individual flux values on a grid of wavelength points. The implications of this difference in spectral representation have to be taken into account when answering questions, such as whether there are spectral lines present in the *Gaia* low-resolution spectrum for a given astronomical object and what the relevant level of statistical significance would be; where the lines are located in terms of wavelength and what the associated uncertainty would be; what the line strength is, whether the line is broad, and what information we might be able to gather from its shape. In the following, we work out the answers to these questions.

The structure of this work is as follows. In Sect. 2, we describe in detail the differences between typical spectra and spectra represented by a linear combination of basis functions, as is the case for *Gaia*. The unconventional properties of the spectra of the latter kind are worked out there. Section 3 summarises the particular properties of *Gaia* DR3 low-resolution spectra. In Sect. 4, we summarise the relevant mathematical properties of Hermite functions and derive mathematical relations required in this work. In Sect. 5, we address relevant aspects of the formation of *Gaia* spectra. In Sect. 6, we discuss the problem of inverting spectra. In the following sections, we discuss spectral lines as local extrema (Sect. 7), the cases of narrow (Sect. 8), and broad lines (Sect. 9), and the use of higher derivatives (Sect. 10). Finally, we provide some example applications to *Gaia* DR3



**Fig. 1.** Sketch illustrating the difference between a typical spectrum and a spectrum represented by a linear combination of basis functions. Top panel: typical spectrum, consisting of sampled flux values on a wavelength grid (black symbols) derived from the spectral photon distribution (red line). Central panel: coefficients representing the same data set by a linear combination of basis functions. Black symbols show the noisy representation, red dots the true coefficients. Bottom panel: linear combination of basis functions (black line) and the  $1-\sigma$  error interval (grey shaded region) as a function of wavelength.

spectra in Sect. 11 and close this work with a summary and discussion in Sect. 12.

## 2. Representing spectra with basis functions

Typical astronomical spectra consist of flux measurements at different wavelength points or bins. The flux measurements usually have low correlations and are significantly correlated only over very short distances in wavelength, if the line spread function (LSF; the point spread function integrated along the direction perpendicular to the dispersion direction) is oversampled. This is illustrated in the sketch in the top panel of Fig. 1. The true spectral photon distribution (SPD, red line) is sampled at certain wavelength points, with uncorrelated normally distributed random noise (black symbols) in this example. A spectral line (shown as an emission line in the sketch) is manifested by flux points deviating from an assumed continuum and the significance of the line can, in principle, be derived from how many standard deviations its data points deviate from the continuum.

If the same data set is instead represented by a linear combination of basis functions, then the coefficients of the development tend to have low correlations over small distances

in coefficient indices. This is shown as a sketch in the central panel of Fig. 1, for the same set of data points. From these coefficients, the spectrum can be constructed by computing the actual linear combination on a continuous wavelength axis. This is shown in the bottom panel of Fig. 1. Here, noise manifests not as scatter of points around a mean, as is the case for the typical spectrum in the top panel, but in continuous, wavy structures. The shaded region in the bottom panel of Fig. 1 gives the formal  $1-\sigma$  error interval around the mean spectrum. It is not trivial to judge visually, which wavy pattern is noise, and which actually is a spectral line. In this work, we derive a formalism to detect lines, ideally from the coefficients describing the spectrum directly, as well as a statistics to quantify the significance of wavy structures in *Gaia*'s low-resolution spectra.

## 3. *Gaia* low resolution spectra

We summarise in this section the properties of *Gaia* low resolution spectra as far as they are relevant in the context of this work. This summary is necessarily brief and incomplete and it is a compilation of information provided by Carrasco et al. (2021) for the principles of the calibration of inner-instrumental effects, by De Angeli et al. (2023) for the actual processing and validation of *Gaia* DR3 data, and by Montegriffo et al. (2023) for the absolute flux and wavelength calibration. We refer to these works for more in-depth discussions.

The *Gaia* low resolution spectrophotometer consists of two parts, producing independent spectra for each observed celestial object on the wavelength intervals of approximately 330–680 nm and 640–1050 nm, respectively. The spectra on the short wavelength range are referred to as ‘blue photometer’ (BP) spectra, and the spectra for the longer wavelength range as ‘red photometer’ (RP) spectra. We use the term XP spectra to refer to both types of *Gaia* spectra. From observed XP spectra of a particular source, transiting *Gaia*'s fields of view repeatedly, the ‘mean’ BP and RP spectra can be constructed. In this process, referred to as the ‘internal calibration’, any inner-instrumental effects – for instance, variations in response or line spread functions with time and position in the focal plane – are suppressed. The mean BP and RP spectra for an astronomical object are represented by a linear combination of 55 Hermite functions in BP and RP, respectively, in *Gaia* DR3. The coefficients of these representations, together with the corresponding errors and correlations, are part of *Gaia* DR3. The argument of the Hermite functions is a linear function of pseudo-wavelength,  $u$ , which is a coordinate along the wavelength dimension in the focal plane common to all BP and RP spectra, respectively, and which corresponds to a mean CCD pixel scale. The dispersion function links the pseudo-wavelength to the actual wavelength,  $\lambda$ . The value of the linear combination of Hermite functions is the flux of the source, in photo-electrons per second, per unit of pseudo-wavelength and within the *Gaia* aperture of  $0.7872 \text{ m}^2$ . The coefficients of the linear combination refer not to Hermite functions directly, but to an orthogonal transformation of Hermite functions. The orthogonal transformations are different for BP and RP, and the transformation matrices are available with *Gaia* DR3.

For the 55 orthogonal transformed Hermite functions, corresponding basis functions in absolute flux and wavelength units are provided with *Gaia* DR3. The same coefficients that represent the internally calibrated BP or RP spectrum also represent the spectral energy distribution when using the corresponding external basis functions. The transformation of internal spectra to external spectra however is intrinsically an ill-conditioned problem (Weiler et al. 2020) and thus sensitive to systematic

**Table 1.** List of notations used throughout this work.

Symbol	Meaning
$\alpha$	Integral over $s_l(\lambda)$
$\mathbf{B}$	Non-standard companion matrix to $\mathbf{c}$
$\mathbf{c}$	Coefficient vector of observational spectrum
$\mathbf{D}^k$	Matrix transforming to $k$ -derivative
$\phi(x)$	Linear combination of Hermite functions
$\varphi_n(x)$	$n$ th Hermite function
$\Phi(x)$	Vector of values of $\varphi_n(x)$ , $n = 0, \dots, N - 1$
$f(u)$	Observational spectrum
$f_c(u)$	Continuum contribution to $f(u)$
$f_l(u)$	Line contribution to $f(u)$
$g^{(k)}(x)$	$k$ th derivative of a function $g(x)$
$\mathbf{i}$	Vector with integrals of Hermite functions
$\lambda$	Wavelength
$\lambda_L$	Wavelength of a narrow line
$L(u, \lambda), L(u, u')$	(pseudo-)line spread function
$N$	Number of coefficients in $\mathbf{c}$
$R(\lambda), R(u')$	Response function
$s(\lambda), s(u')$	Spectral photon distribution
$s_c(\lambda), s_c(u')$	Continuum contribution to $s(\lambda), s(u')$
$s_l(\lambda), s_l(u')$	Line contribution to $s(\lambda), s(u')$
$\Sigma_x$	Covariance matrix for vector $\mathbf{x}$
$u$	Pseudo-wavelength
$u'$	Pseudo-wavelength corresponding to $\lambda$
$u_E$	Pseudo-wavelength for local extremum
$u_L$	Pseudo-wavelength for spectral line
$W$	Equivalent width of a spectral line

errors introduced to what essentially is a deconvolution process. For the analysis of spectral lines, the use of internally calibrated spectra is therefore preferable. Furthermore, the Hermite functions used in the representation of internally calibrated BP and RP spectra provide very useful mathematical properties that we can exploit in the analysis of spectral lines.

## 4. Some mathematical properties of linear combinations of Hermite functions

In this section, we compile and derive the mathematical properties and relationships involving Hermite functions that we need in the context of this work. We separate the mathematical aspects in this section from the discussion of the instrumental and astrophysical aspects. In doing so, we simplify the structure of this work, by not interrupting physical and astronomical discussions later in the work with the introduction of mathematical properties related to Hermite functions. For easy reference, Table 1 provides an overview of the nomenclature that is used throughout this paper.

### 4.1. Hermite functions and linear combinations thereof

Hermite functions appear in a number of mathematical and scientific contexts. In physics, they own their relevance widely to being the eigenfunctions of the quantum mechanical harmonic oscillator. A detailed discussion of Hermite functions and their properties may therefore be found in most basic text books on quantum mechanics (e.g. Magnasco 2013). The  $n$ th Hermite function we denote with  $\varphi_n(x)$ , with  $n = 0, 1, 2, \dots$ . The index  $n$  we refer to as the ‘order’ of the Hermite function. In the context

of this work, we need two basic recurrence relations for Hermite functions  $\varphi_n(x)$ ,  $n = 0, 1, 2$ , etc. The first recurrence relation is:

$$x \varphi_n(x) = \sqrt{\frac{n}{2}} \varphi_{n-1}(x) + \sqrt{\frac{n+1}{2}} \varphi_{n+1}(x), \quad (1)$$

together with

$$\begin{aligned} \varphi_0(x) &= \pi^{-1/4} e^{-\frac{1}{2} x^2}, \\ \varphi_1(x) &= \pi^{-1/4} \sqrt{2} x e^{-\frac{1}{2} x^2}. \end{aligned}$$

From this relation between Hermite functions of different order  $n$  and the first two functions given explicitly, any Hermite function can easily be computed for any argument,  $x$ .

The second recurrence relation connects the first derivative of a Hermite function with Hermite functions. In this work, we denote the  $k$ th derivative of a function with the superscript “ $(k)$ ”. For Hermite functions, we have the following relation:

$$\varphi_n^{(1)}(x) = \sqrt{\frac{n}{2}} \varphi_{n-1}(x) - \sqrt{\frac{n+1}{2}} \varphi_{n+1}(x). \quad (2)$$

The Hermite functions are a set of orthonormal functions on  $\mathbb{R}$ , i.e. the integrals over the product of two Hermite functions over the entire real axis obeys the orthonormality relation

$$\int_{-\infty}^{\infty} \varphi_i(x) \cdot \varphi_j(x) dx = \delta_{ij}, \quad (3)$$

with  $\delta_{ij}$  the Kronecker delta. Furthermore, we consider linear combinations of Hermite functions, for instance, functions of the form:

$$\phi(x) = \sum_{n=0}^N c_n \varphi_n(x), \quad (4)$$

with the  $c_n$  the coefficients of the linear combination. We may write the  $N + 1$  coefficients as an  $N + 1$  element vector  $\mathbf{c}$ . With the orthonormality expressed in Eq. (3), the coefficients  $c_n$  in Eq. (4) become

$$c_n = \int_{-\infty}^{\infty} \phi(x) \cdot \varphi_n(x) dx. \quad (5)$$

### 4.2. Integrals over linear combinations of Hermite functions

The application of Eq. (3) turns the computation of integrals over the real axis over products of Hermite functions into a trivial problem. Using the relationships of Eq. (2) also makes the computation of integrals over finite intervals  $[a, b]$  on the real axis simple. We get:

$$\int_a^b \varphi_n^{(1)}(x) dx = \sqrt{\frac{n}{2}} \int_a^b \varphi_{n-1}(x) dx - \sqrt{\frac{n+1}{2}} \int_a^b \varphi_{n+1}(x) dx, \quad (6)$$

and therefore

$$\begin{aligned} \int_a^b \varphi_{n+2}(x) dx &= \sqrt{\frac{n+1}{n+2}} \int_a^b \varphi_n(x) dx \\ &\quad + \sqrt{\frac{2}{n+1}} [\varphi_{n+1}(a) - \varphi_{n+1}(b)]. \end{aligned} \quad (7)$$

Together with the indefinite integrals for the first two Hermite functions (Brychkov et al. 1989),

$$\int \varphi_0(x) dx = \frac{1}{\sqrt{2}} \pi^{\frac{1}{4}} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \quad (8)$$

$$\int \varphi_1(x) dx = -\sqrt{2} \pi^{-\frac{1}{4}} e^{-\frac{x^2}{2}}, \quad (9)$$

the integrals over any interval on the real axis can be computed iteratively for all orders. In particular, for integrations of the entire real axis, that is,  $a \equiv -\infty$  and  $b \equiv \infty$ , we obtain

$$\int_{-\infty}^{\infty} \varphi_{2n+1}(x) dx = 0 \quad (10)$$

and the integrals over the even order Hermite functions can be computed from

$$\int_{-\infty}^{\infty} \varphi_0(x) dx = \pi^{\frac{1}{4}} \sqrt{2}, \quad (11)$$

$$\int_{-\infty}^{\infty} \varphi_{2(n+1)}(x) dx = \sqrt{\frac{n+\frac{1}{2}}{n+1}} \int_{-\infty}^{\infty} \varphi_{2n}(x) dx. \quad (12)$$

When arranging the integrals over the interval  $[a, b]$  for the first  $N$  Hermite functions as an  $N$ -element vector  $\mathbf{i}$ , then the integral over the linear combination of Hermite functions  $\phi(x)$  from Eq. (4) over the real axis is

$$\int_{-\infty}^{\infty} \phi(x) dx = \mathbf{i}^T \mathbf{c}. \quad (13)$$

Since the vector  $\mathbf{i}$  only depends on the Hermite functions, but not on the particular linear combination thereof, the integral over any linear combination of Hermite functions can eventually be computed by the cost of one vector multiplication.

#### 4.3. Derivatives of linear combinations of Hermite functions

The recurrence relation given by Eq. (2) has immediate important consequences, as the derivative of an Hermite function is a linear combination of two other Hermite functions, with orders increased and reduced by one. And since differentiation is a linear operation, it follows that the derivative of any linear combination of Hermite functions is another linear combination of Hermite functions. Thus, the  $k$ th derivative of a linear combination of Hermite functions with maximal order  $N$  is a linear combination of Hermite functions of maximal order  $N + k$ .

For linear combinations of Hermite functions, the differentiation is therefore conveniently expressed by a matrix operation. We define the  $(N+1) \times N$  matrix  $\mathbf{D}^1$  as having zero entries except for the two sub-diagonals, whose entries are:

$$\mathbf{D}_{i,i+1}^1 = \sqrt{\frac{i}{2}} = -\mathbf{D}_{i+1,i}^1. \quad (14)$$

Then the derivative of a linear combination of the first  $N$  Hermite functions, with coefficient vector  $\mathbf{c}$ , is the linear combination of the first  $N + 1$  Hermite functions, with the coefficient vector  $\mathbf{c}^{(1)}$  given by  $\mathbf{c}^{(1)} = \mathbf{D}^1 \mathbf{c}$ .

We may generalise the notation to the  $k$ th derivative, if we expand the size of the matrix  $\mathbf{D}^1$  to  $(N+k) \times (N+k)$ , and assume zero-padding for  $\mathbf{c}$ , that is, assuming  $\mathbf{c}$  being of length  $N+k$ , with all entries  $N < n \leq N+k$  being zero. Thus, we obtain:

$$\mathbf{c}^{(k)} = \mathbf{D}^k \mathbf{c}, \quad (15)$$

and the  $k$ th derivative matrix  $\mathbf{D}^k$ , we obtain from  $k$  multiplications of  $\mathbf{D}^1$ ; that is,  $\mathbf{D}^k$  is the  $k$ th power of  $\mathbf{D}^1$ . For  $\mathbf{D}^2 = \mathbf{D}^1 \mathbf{D}^1$  we obtain a matrix whose elements are as follows:

$$\mathbf{D}_{i,j}^2 = \begin{cases} \frac{1}{2} - i & \text{for } i = j \\ \frac{1}{2} \sqrt{i(i+1)} & \text{for } i - j = -2 \\ \frac{1}{2} \sqrt{j(j+1)} & \text{for } i - j = 2 \\ 0 & \text{else.} \end{cases} \quad (16)$$

For illustration, the first  $5 \times 5$  elements of  $\mathbf{D}^1$  and  $\mathbf{D}^2$  are

$$\mathbf{D}^1 = \begin{pmatrix} 0 & \sqrt{\frac{1}{2}} & 0 & 0 & 0 \\ -\sqrt{\frac{1}{2}} & 0 & \sqrt{\frac{1}{2}} & 0 & 0 \\ 0 & -\sqrt{\frac{1}{2}} & 0 & \sqrt{\frac{3}{2}} & 0 \\ 0 & 0 & -\sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{4}{2}} \\ 0 & 0 & 0 & -\sqrt{\frac{4}{2}} & 0 \end{pmatrix} \quad (17)$$

and

$$\mathbf{D}^2 = \begin{pmatrix} -\frac{1}{2} & 0 & \sqrt{\frac{1}{2}} & 0 & 0 \\ 0 & -\frac{3}{2} & 0 & \sqrt{\frac{3}{2}} & 0 \\ \sqrt{\frac{1}{2}} & 0 & -\frac{5}{2} & 0 & \sqrt{3} \\ 0 & \sqrt{\frac{3}{2}} & 0 & -\frac{7}{2} & 0 \\ 0 & 0 & \sqrt{3} & 0 & -\frac{9}{2} \end{pmatrix}. \quad (18)$$

The condition number of  $\mathbf{D}^1$  is larger than one if its size is larger than  $2 \times 2$ , and the condition number increases with the size of the matrix. For a  $56 \times 56$ -sized matrix  $\mathbf{D}^1$ , as required for the computation of the first derivative of Gaia DR3 spectra from 55 coefficients, the condition number is 66.2. As any derivative larger than one is a power of  $\mathbf{D}^1$ , the condition numbers of the derivative matrices increase with the power of the number of derivative. If the coefficients of a linear combination of Hermite functions are affected by noise, the relative noise on the coefficients for higher derivatives consequently increases strongly. It is therefore advisable to work with derivatives of as low an order as possible in the case of noisy data. The highest derivatives we use in the course of this work is the fourth. With the  $k$ th derivative of a linear combination of Hermite functions being a linear combination of Hermite functions, the coefficients of which can be easily computed, we now proceed to the computation of the roots of a linear combination of Hermite functions.

#### 4.4. Roots of linear combinations of Hermite functions and their uncertainties

From the relations in Eq. (1) we see that the  $n$ th Hermite function is actually a polynomial of a degree  $n$ , multiplied by a Gaussian function  $\exp\{-x^2/2\}$ . Since the Gaussian function is a strictly positive function, it does not affect the position of the

root of the Hermite function. The roots of the Hermite functions coincide with the roots of the polynomial. Thus, a linear combination of Hermite functions with maximum order  $N$  has exactly  $N$  roots, which may be real and complex. For the computation of the actual roots, we may also follow the lines of root finding for polynomials. For polynomials in monomial representation, a standard approach to this problem is interpreting the polynomial as the characteristic polynomial of some matrix, called the companion matrix of the polynomial. With finding this matrix, the problem of root finding for a linear combination of monomials is transformed into an eigenvalue problem (Press et al. 2007a), which can take advantage of the variety of fast and robust numerical eigenvalue techniques. For non-monomial representations of polynomials, an elegant generalisation of this approach has been developed for orthogonal polynomials. This approach makes use of the recurrence relationship for orthogonal polynomials, rather than the polynomial properties directly, and it is therefore immediately transferable to Hermite functions, satisfying a suitable recurrence relationship expressed in Eq. (1). Here, we follow the line discussed by Day & Romero (2005), with the trivial transfer from orthogonal polynomials to Hermite functions. We only present the required results here and avoid the repetition of the not too complicated proofs and derivations, as it saves space and as they have already been presented by Day & Romero (2005) in detail.

Here, we consider a linear combination of Hermite functions according to Eq. (4). With  $\Phi_N(x)$ , we denote the  $N$ -element vector of functions, containing the first  $N$  Hermite functions, with indices from 0 to  $N - 1$ . Furthermore, with  $\mathbf{c}_N$  we denote the vector of the first  $N$  entries of the coefficient vector  $\mathbf{c}$ , and with  $\mathbf{e}_N$  the  $N$  element vector with a value of one as the  $N$ th entry and zero else; that is, we have:

$$\Phi_N(x) := [\varphi_0(x), \varphi_1(x), \dots, \varphi_{N-1}(x)]^\top, \quad (19)$$

$$\mathbf{c}_N := [c_0, c_1, \dots, c_{N-1}]^\top, \quad (20)$$

$$\mathbf{e}_N := [0, 0, \dots, 0, 1]^\top. \quad (21)$$

We assume  $c_N \neq 0$  here, which can be ensured by trimming the length of  $\mathbf{c}$  such that the highest order entry is non-zero if necessary. With these conventions, the function  $\phi(x)$  from Eq. (4) is written as:

$$\phi(x) = \Phi_N^\top(x) \mathbf{c}_N + c_N \varphi_N(x). \quad (22)$$

For the moment, it is not necessary to assume that the functions  $\varphi_n(x)$  are actually the Hermite functions and we may only assume that these functions are satisfying a general recurrence relation of the form:

$$x \varphi_{n-1}(x) = \sum_{i=0}^n \varphi_i(x) h_{i,n-1}. \quad (23)$$

Then we can form the  $N \times N$  matrix  $\mathbf{H}$ , whose elements are the  $h_{i,j}$  from the recurrence relation above, with  $i$  and  $j$  running from 0 to  $N - 1$ . We define the  $N \times N$  matrix  $\mathbf{B}$ , denoted the non-standard companion matrix, by

$$\mathbf{B} := \mathbf{H} - \frac{h_{N,N-1}}{c_N} \mathbf{c}_N \mathbf{e}_N^\top. \quad (24)$$

The value  $x_0$  is a root of the linear combination of the first  $N + 1$  functions  $\varphi_n(x)$  with coefficient vector  $\mathbf{c}$  – if and only if  $x_0$  is an eigenvalue of the matrix  $\mathbf{B}$ . Thus, the problem of root finding is again transformed into an eigenvalue problem for

linear combinations of functions satisfying a suitable recurrence relation.

For Hermite functions, the general recurrence relation in Eq. (23) takes the form given by Eq. (1), and we find

$$h_{i,j} = \begin{cases} \sqrt{\frac{i+1}{2}} & \text{for } j = i + 1 \\ \sqrt{\frac{i}{2}} & \text{for } j = i - 1 \\ 0 & \text{otherwise.} \end{cases} \quad (25)$$

As an example, for the case of  $N = 5$ , matrix  $\mathbf{B}$  takes the following form:

$$\mathbf{B} = \begin{pmatrix} 0 & \sqrt{\frac{1}{2}} & 0 & 0 & -\sqrt{\frac{5}{2}} \frac{c_0}{c_5} \\ \sqrt{\frac{1}{2}} & 0 & \sqrt{\frac{2}{2}} & 0 & -\sqrt{\frac{5}{2}} \frac{c_1}{c_5} \\ 0 & \sqrt{\frac{2}{2}} & 0 & \sqrt{\frac{3}{2}} & -\sqrt{\frac{5}{2}} \frac{c_2}{c_5} \\ 0 & 0 & \sqrt{\frac{3}{2}} & 0 & \sqrt{\frac{4}{2}} - \sqrt{\frac{5}{2}} \frac{c_3}{c_5} \\ 0 & 0 & 0 & \sqrt{\frac{4}{2}} & -\sqrt{\frac{5}{2}} \frac{c_4}{c_5} \end{pmatrix}. \quad (26)$$

The non-standard companion matrix  $\mathbf{B}$  is thus easily generated for any given coefficient vector  $\mathbf{c}$ . The matrix  $\mathbf{B}$  is of upper Hessenberg form and the computations of the eigenvalues of a real Hessenberg matrix is particularly fast and simple (see, e.g. Chap. 7.4 in Golub & Van Loan 2013, or Press et al. 2007b).

We may, however, consider the case in which the coefficient vector,  $\mathbf{c}$ , is not exact, but comes with an error, expressed in a covariance matrix,  $\Sigma_c$ . For finding an estimate on the uncertainties of the roots of the linear combination of Hermite functions, we again may apply the approach developed for polynomial roots. Here, we follow the formulations in Sect. 3.2.2 of Stetter (2004) – again with a trivial transfer of argument from polynomials to Hermite functions.

The root of a linear combination  $\phi(x)$  as in Eq. (4) is a non-linear function of the coefficients in the vector  $\mathbf{c}$ . A linearisation of the problem is thus required for an efficient computation of the uncertainties of the roots. The roots of  $\phi(x)$  we denote as  $z_i$ ,  $i = 1, \dots, N$ , and we consider the  $z_i$  functions of the coefficient vector  $\mathbf{c}$ , as we do for the linear combination  $\phi(x)$ . We thus write:

$$\phi(z_i(\mathbf{c}), \mathbf{c}) = 0, \quad (27)$$

and compute the derivative of this expression with respect to  $\mathbf{c}$ , resulting in

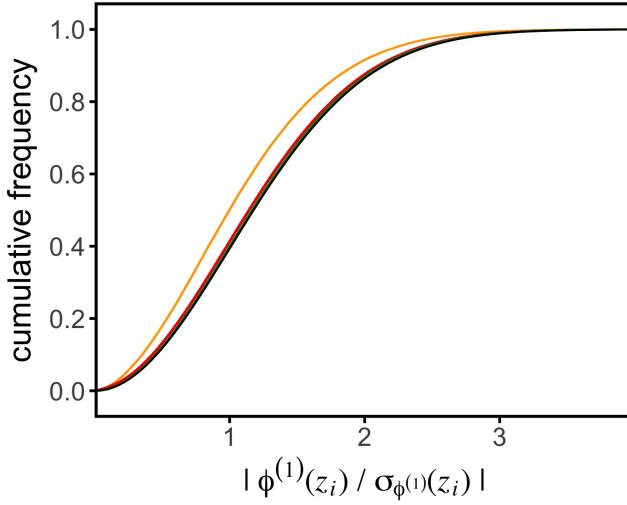
$$\phi^{(1)}(z_i(\mathbf{c}), \mathbf{c}) \cdot \frac{d z_i(\mathbf{c})}{d \mathbf{c}} + \frac{\partial \phi(z_i(\mathbf{c}), \mathbf{c})}{\partial \mathbf{c}} = 0. \quad (28)$$

We thus obtain for each component  $c_j$  of  $\mathbf{c}$ , the derivative:

$$\frac{\partial z_i}{\partial c_j} = -\frac{\varphi_j(z_i)}{\phi^{(1)}(z_i)}, \quad (29)$$

which are the elements  $i, j$  of the Jacobian of the roots with respect to the coefficients of the linear combination, which we denote as  $\mathbf{J}$ . In a linear approximation, we obtain the covariance matrix of the roots,  $\Sigma_z$ , as

$$\Sigma_z = \mathbf{J} \Sigma_c \mathbf{J}^\top. \quad (30)$$



**Fig. 2.** Empirical cumulative distribution functions for the absolute value of the derivative, normalised to its error, at the roots of linear combinations of Hermite functions, for  $N = 5$  (orange), 55 (red), and 200 (green). The black curve shows the cumulative distribution function for a chi distribution with two degrees of freedom.

#### 4.5. Significance of local extrema

When considering a linear combination of Hermite functions, with coefficients affected by normally distributed random noise, we want to find out whether a given local extremum is statistically significant or whether it is a noise-induced feature that can be removed within some boundaries given by the covariance matrix on the coefficients. A local extremum is removed if the second derivative at the extremum is zero. In this case, the extremum is transformed to an inflection point. We might therefore consider the second derivative at the local extremum, normalised to its error, with the probability of this value being zero. This probability might tell us about the statistical significance of the local extremum.

To investigate this approach quantitatively, we used Monte Carlo simulations. We generated random covariance matrices and for each of the random covariance matrices, we drew a large number of sets of random coefficients with zero mean. We then computed the roots of the corresponding random linear combination of Hermite functions, and the values of the derivative divided by the error of the value of the derivative, at the roots. We then investigated the distribution function. We carried out this procedure for a variety of different numbers of coefficients. Figure 2 shows resulting empirical cumulative distribution functions as examples, for different numbers  $N$ , and  $5 \times 10^5$  random realisations. These results only include roots within the outermost extrema of the highest order Hermite function. With increasing  $N$ , the distribution approaches a chi distribution with two degrees of freedom.

Given the cumulative distribution function of a chi distribution with two degrees of freedom, we can therefore estimate the significance of a local extremum of a linear combination of Hermite functions at  $z_i$  by attributing an approximate  $p$ -value given by:

$$p(z_i) \approx 1 - \exp \left\{ -\frac{1}{2} \left( \frac{\phi^{(2)}(z_i)}{\sigma_{\phi^{(2)}(z_i)}} \right)^2 \right\}. \quad (31)$$

#### 4.6. Linear substitutions

We close this section on the mathematical tools with a remark on the substitution  $x \rightarrow (u - \Delta\theta)/\Theta$  in the argument of the Hermite functions. Such a transformation has been applied to *Gaia* DR3 low-resolution spectra to adjust the widths of the Hermite functions to the widths of the spectra (De Angeli et al. 2023). In this case,

$$\int_{-\infty}^{\infty} \varphi_n \left( \frac{u - \Delta\theta}{\Theta} \right) \varphi_m \left( \frac{u - \Delta\theta}{\Theta} \right) du = \Theta \delta_{nm}. \quad (32)$$

To maintain orthonormality, the Hermite functions after linear substitution need thus to be multiplied by  $1/\sqrt{\Theta}$ . The derivative of a Hermite function after substitution is

$$\frac{d}{du} \varphi_n \left( \frac{u - \Delta\theta}{\Theta} \right) = \frac{1}{\Theta} \frac{d}{dx} \varphi_n(x). \quad (33)$$

For the  $k$ th derivative with respect to  $u$ , an additional factor  $\Theta^{-k}$  needs to be applied to the result.

Taking the additional factors depending on  $\Theta$  into account when differentiating and integrating Hermite functions with respect to  $u$ , we can use all results derived for linear combinations of Hermite functions also for Hermite functions after a linear substitution of its argument has been applied. For general substitutions in the form  $f(x) \rightarrow f(y)$ , we write  $f(y)$  for  $f(y(x)) dy/dx$ .

### 5. Elements of the formation of spectra

Before entering into the analysis of spectral lines, we first resumed certain aspects of the formation of the spectra in this section to establish the relationship between the low-resolution observational *Gaia* spectrum and the underlying SPD. In particular, we worked out the role of the LSF and the advantages resulting from their efficient parameterisation.

An observational spectrum  $f(u)$  is often related to the SPD  $s(\lambda)$  by an integral transformation of the form:

$$f(u) = \int_0^{\infty} L(u, \lambda) R(\lambda) s(\lambda) d\lambda, \quad (34)$$

or a modified formulation thereof. Here,  $L(u, \lambda)$  denotes the LSF, including the dispersion, and  $R(\lambda)$  is the response function. In optical theory, the LSF is the inverse Fourier transform of the optical transfer function, integrated over the spatial direction (Goodman 1996). For *Gaia* DR3 low-resolution spectra, this concept of an LSF is however not appropriate. The BP and RP spectra are obtained as a finite linear combination of Hermite functions constructed in the internal self-calibration process (Carrasco et al. 2021) and are not directly related to an optical transfer function describing the actual spectroscopic instrument. However, the XP spectra can still be linked to the SPD via an integral transformation of the general form of Eq. (34) if the function  $L(u, \lambda)$  is given a more specific form. Setting

$$L(u, \lambda) := \sum_{n=0}^N l_n(\lambda) \varphi_n(u) \quad (35)$$

and inserting it into Eq. (34), and also using Eq. (5), we obtain the following for *Gaia* DR3 low-resolution spectra:

$$f(u) \equiv \sum_{n=0}^N c_n \varphi_n(u) = \int_0^\infty L(u, \lambda) R(\lambda) s(\lambda) d\lambda \quad (36)$$

and

$$c_n = \int_0^\infty l_n(\lambda) R(\lambda) s(\lambda) d\lambda. \quad (37)$$

Thus, the function  $L(u, \lambda)$  is taking the role of the LSF, in the sense that it represents the instrument's response to a monochromatic signal. It is however not directly linked to the optical transfer function. The most important consequence of this is that  $L(u, \lambda)$  is not necessarily a positive function. The response to a monochromatic input, when represented by a finite linear combination of Hermite functions, might show oscillational behaviour and be partly negative as well. We refer to  $L(u, \lambda)$  as the ‘pseudo-LSF’ hereafter for convenience. The vector for all  $l_n(\lambda)$ ,  $n = 0, \dots, N$ , we denote as  $\mathbf{l}(\lambda)$ .

In carrying out this work, it appears convenient to allow for a fast computation of the vector  $\mathbf{l}(\lambda)$ . Therefore, we considered developing  $L(u, \lambda)$ , not only in Hermite functions in the dependency on  $u$ , but also developing the wavelength dependency in basis functions. Doing so requires a suitable choice for the basis functions describing the wavelength dependency. In searching for a good choice for the basis in wavelength, we note that we can write:

$$l_n(\lambda) = \int_{-\infty}^\infty L(u, \lambda) \varphi_n(u) du. \quad (38)$$

This is essentially equivalent to applying the smoothing effect of the pseudo-LSF  $L(u, \lambda)$  on the Hermite function  $\varphi_n(u)$ . If the dispersion relation linking  $u$  and  $\lambda$  is applied to substitute  $\lambda$  with  $u$ , then the resulting function  $l_n(u)$  might be very similar to the  $n$ th Hermite function, particularly for small  $n$ . We therefore expect an efficient development of the function  $L(\lambda, u)$  when substituting the wavelength with the pseudo-wavelength, which we denote  $u'$  to distinguish it from  $u$ ; and we use Hermite functions for the development in  $u'$  as well. We thus assume as a good approximation:

$$L(u, u') \approx \sum_{n=0}^N \sum_{m=0}^M \Lambda_{n,m} \varphi_n(u) \varphi_m(u'), \quad (39)$$

with  $M$  only moderately larger than  $N$ . With  $\mathbf{\Lambda}$  as the matrix whose elements are the  $\Lambda_{n,m}$ , we thus obtain:

$$\mathbf{l}(u') = \mathbf{\Lambda} \Phi(u'), \quad (40)$$

with

$$\Phi(u') := [\varphi_0(u'), \varphi_1(u'), \dots, \varphi_M(u')]^\top. \quad (41)$$

We use Eq. (40) in Sect. 6 of this work. As further preparation for Sect. 6, we consider the computation of some integrals involving the pseudo-LSF over  $u'$ , namely, integrals of the form:

$$\mathcal{L}_{k,l}(u_0, u_1) := \int_{-\infty}^\infty L^{(k)}(u_1, u') \cdot (u' - u_0)^l du', \quad (42)$$

for  $k, l = 0, 1, 2, \dots$ . Since the pseudo-LSF has an exponential decrease, these integrals exist. The choice of Hermite functions for the description of the dependency on both  $u$  and  $u'$  allows us to make use of results from Sect. 4 to find a simple solution to these integrals. For the case of  $l = 0$ , from combining Eqs. (13), (15), and (40), we immediately obtain:

$$\mathcal{L}_{k,0}(u_0, u_1) = \mathbf{i}^\top \mathbf{D}^k \mathbf{\Lambda}^\top \Phi(u_1). \quad (43)$$

For the case of  $l = 1$ , we use the recurrence relation from Eq. (1) to develop the linear combination of Hermite functions multiplied by  $(u' - u_0)$  in a linear combination of Hermite functions. The coefficients of this linear combination are obtained by multiplication with a tridiagonal matrix:  $\mathbf{P}(u_0)$ , with the elements

$$\mathbf{P}_{i,j}(u_0) = \begin{cases} -u_0 & \text{for } i = j \\ \sqrt{\frac{i}{2}} & \text{for } i - j = -1 \\ \sqrt{\frac{j}{2}} & \text{for } i - j = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (44)$$

For the case of  $l > 1$ , we obtain the result by multiplying  $\mathbf{P}(u_0)$   $l$  times with itself, which we express as:  $\mathbf{P}^l(u_0)$ . For the general case, we get:

$$\mathcal{L}_{k,l}(u_0, u_1) = \mathbf{i}^\top \mathbf{P}^l(u_0) \mathbf{D}^k \mathbf{\Lambda}^\top \Phi(u_1), \quad (45)$$

we use this equation in Sect. 6. It is an apt demonstration for how the recurrence relations expressed in Eqs. (1) and (2) allow for the efficient computation of integrals, as in Eq. (42), by making use of matrix algebra.

We close this section with the simple observation that the linearity of the integration implies that if  $s(\lambda)$  can be expressed as the sum of two SPDs, say  $s(\lambda) = s_c(\lambda) + s_l(\lambda)$ , the corresponding observational spectrum is the sum of the observational spectra of  $s_c(\lambda)$  and  $s_l(\lambda)$ ,

$$f(u) = \int_0^\infty L(u, \lambda) R(\lambda) [s_c(\lambda) + s_l(\lambda)] d\lambda \quad (46)$$

$$= f_c(u) + f_l(u). \quad (47)$$

If  $s_c(\lambda)$  and  $s_l(\lambda)$  separate the SPD onto a contribution from a continuum SPD and a SPD of a spectral line, respectively, we can thus separate also the observational spectra into a continuum contribution and a line contribution. For a convenient notation in later sections, we introduce the abbreviation:

$$t(\lambda) := R(\lambda) s(\lambda). \quad (48)$$

## 6. Inversion of observational spectra

After having derived the relevant mathematical properties of linear combinations of Hermite functions in Sect. 4 and having combined the results of this section with the observational spectra in Sect. 5, we must then finally link all results to the external spectrum, that is, we have to consider the inversion of internally calibrated *Gaia* spectra to externally calibrated SPDs. This inversion is required to reduce the smearing effect of the pseudo-LSF when quantifying the line strengths, as discussed later. The inversion of observational spectra is intrinsically a complex problem. However, as far as this work is concerned, we may make some simplifying assumptions that will allow us to derive a simple

and efficient formalism. First, for the study of a spectral line, we require a reliable inversion of the observational spectrum only in the vicinity of the spectral line. We do not need a reliable inversion of the entire observational spectrum. Far from the position of the line, the inversion might be arbitrarily wrong for our purposes. Second, we are mainly interested in the inversion of the continuum spectrum, where the spectral lines are removed. For the continuum, we may well assume smoothness on the inverted spectrum. We are thus left with the problem of finding a local inversion to a smooth external spectrum, which is considerably easier than the problem of finding a global inversion to a general external spectrum.

A smooth local approximation to the inverted spectrum calls for a polynomial approximation in the vicinity of the line position. Naturally, the use of a Taylor expansion around the line position follows. This approach we line out in the following as the first of two approaches to the local smooth inversion.

We consider the product of the response curve and the continuum flux at some value  $u_0$ . To obtain  $s_c(u_0) R(u_0) =: t_c(u_0)$ , we use Eq. (34), with Eq. (48) and transformed from wavelength to pseudo-wavelength, which is:

$$f(u_0) = \int_{-\infty}^{\infty} L(u_0, u') t_c(u') du'. \quad (49)$$

We introduce a Taylor expansion of  $t(u')$  centred on  $u_0$ , resulting in:

$$f(u_0) = \int_{-\infty}^{\infty} L(u_0, u') \left[ \sum_{l=0}^{\infty} \frac{t_c^{(l)}(u_0)}{l!} (u' - u_0)^l \right] du' \quad (50)$$

$$= \sum_{l=0}^{\infty} \frac{t_c^{(l)}(u_0)}{l!} \int_{-\infty}^{\infty} L(u_0, u') (u' - u_0)^l du' \quad (51)$$

$$= \sum_{l=0}^{\infty} \frac{t_c^{(l)}(u_0)}{l!} \mathcal{L}_{0,l}(u_0, u_0), \quad (52)$$

with  $\mathcal{L}_{k,l}(u_0, u_0)$  as defined in Eq. (42). If we wish to approximate  $t_c(u_0)$  in a higher order approximation, we need to impose additional conditions. We can derive these conditions from higher derivatives at the same point,  $u_0$ .

We compute the  $k$ th derivative of  $f(u)$  with respect to  $u$ , and we obtain an analogous expression,

$$f^{(k)}(u_0) = \sum_{l=0}^{\infty} \frac{t_c^{(l)}(u_0)}{l!} \mathcal{L}_{k,l}(u_0, u_0). \quad (53)$$

If we truncate the Taylor expansion at the  $K$ th term, we obtain a matrix equation:

$$\mathbf{f}(u_0) = \mathbf{\Omega}(u_0) \mathbf{t}(u_0), \quad (54)$$

with

$$\mathbf{f}(u_0) = [ f^{(0)}(u_0) f^{(1)}(u_0) f^{(2)}(u_0) \dots f^{(K)}(u_0) ]^T, \quad (55)$$

$$\mathbf{t}(u_0) = [ t_c^{(0)}(u_0) t_c^{(1)}(u_0) t_c^{(2)}(u_0) \dots t_c^{(K)}(u_0) ]^T, \quad (56)$$

and the  $(K+1) \times (K+1)$  matrix  $\mathbf{\Omega}(u_0)$ , whose elements are

$$\Omega_{i,j}(u_0) = \frac{\mathcal{L}_{(i-1),(j-1)}(u_0, u_0)}{(j-1)!}. \quad (57)$$

Equation (54) is a formulation for the inversion of the observational spectrum by polynomial approximations to the product of the response function and the SPD.

The matrix  $\mathbf{\Omega}(u_0)$  takes a simple form, namely,

$$\Omega_{i,j}(u_0) = 0 \text{ for } i > j \quad (58)$$

$$\Omega_{i,i+1}(u_0) = \Omega_{i+1,i+2}(u_0) \text{ for } i \geq 0. \quad (59)$$

Equation (58) states that the matrix  $\mathbf{\Omega}$  is of upper triangular type. Equation (59) states that the entries of the diagonal and the secondary diagonals are all identical, respectively. This implies that the  $(K+1) \times (K+1)$  matrix  $\mathbf{\Omega}$  has only  $K+1$  independent elements that need to be computed. This allows for a fast computation of the matrix and its inverse and makes the usage computationally simple.

For the computation of the formal error on  $t_c(u_0)$ , which we can use to judge the quality of the inversion of Eq. (54), it is convenient to write the vector  $\mathbf{f}(u_0)$  as:

$$\mathbf{f}(u_0) = \mathbf{H}(u_0) \mathbf{c}, \quad (60)$$

with the  $(K+1) \times (N+K)$  matrix  $\mathbf{H}(u_0)$  whose  $l$ th row is  $\Phi(u_0) \mathbf{D}^{(l-1)}$ . Then,

$$\mathbf{t}(u_0) = \mathbf{\Omega}^{-1} \mathbf{H}(u_0) \mathbf{c}, \quad (61)$$

and the covariance matrix for  $\mathbf{t}(u_0)$  is

$$\Sigma_t = \mathbf{\Omega}^{-1} \mathbf{H}(u_0) \Sigma_c \mathbf{H}^T(u_0) (\mathbf{\Omega}^{-1})^T. \quad (62)$$

The disadvantage of this method comes from the dependency on higher order derivatives of the spectrum at  $u_0$ . As discussed in the previous section, the signal-to-noise ratio (S/N) of higher derivatives decreases strongly with increasing derivative and local inversion becomes dominated by noise easily. In practice, only  $K=1$  or  $K=2$  might result in good estimates for  $t_c(u_0)$ .

Here, it appears useful to replace the conditions on higher derivative by conditions on the zeroth derivative at different pseudo-wavelength positions in order to become less sensitive to random noise in the inversion. The change is straight forward, replacing the derivative conditions in Eq. (53) by conditions of  $f^{(0)}(u_i)$  at different points,  $u_i$ , in the vicinity of  $u_0$ . The results obtained with this approach however depend strongly on the choice of the  $u_i$  and are not necessarily better than the results with derivative conditions. A more suitable approach results from adapting a technique for the global inversion from Weiler et al. (2020) to the problem of a local smooth inversion.

Thus, we selected a set of basis functions for the SPD, such as Legendre polynomials and, assuming sufficient knowledge of the instrument, we computed the observations spectra corresponding to the basis functions using Eq. (37). Locally approximating the observations spectrum around a pseudo-wavelength  $u_0$  with these observational basis functions corresponds to approximating the SPD around a wavelength,  $\lambda(u_0)$ , by Legendre polynomials. To avoid conditioning issues, we may adjust the argument of the Legendre polynomials to the range where these functions are orthonormal, namely, in the range of  $[-1, 1]$ . This approach is the same as used for the Hermite functions describing the Gaia DR3 XP spectra. For sake of simplicity, we refer to this linear substitution for the Legendre polynomials  $P_i(x)$  as  $x = g(\lambda)$ . For the approximation of the observational spectrum, we do not have to ensure an exact solution, but do a regression instead and reduce the influence of random noise. We chose an interval  $\Delta$  in pseudo-wavelength, centred on  $u_0$ , and then chose a grid of

$M$  equidistant points on this interval. We arranged the grid point as a vector  $\mathbf{u}$ . Since the errors of the observational spectrum at the set of  $M$  points  $u_i$ ,  $i = 1, \dots, M$  are likely to be correlated, we have to find a linear combination of observational spectra for the first  $K$  Legendre polynomials, the coefficients of which are arranged as a vector  $\mathbf{b}$ , that minimise the squared Mahalanobis distance:

$$D_m^2 = \left( f(\mathbf{u}) - \Phi^\top(u) \mathbf{L} \mathbf{b} \right)^\top \Sigma_f^{-1} \left( f(\mathbf{u}) - \Phi^\top(u) \mathbf{L} \mathbf{b} \right). \quad (63)$$

Here,  $\mathbf{L}$  is the  $N \times K$  matrix of coefficients of the observational spectra corresponding to the first  $K$  Legendre polynomials, developed in Hermite functions. This is a generalised least squares problem with the solution (Kariya & Kurata 2004):

$$\mathbf{b} = \left( \left( \Phi^\top(\mathbf{u}) \mathbf{L} \right)^\top \Sigma_f^{-1} \left( \Phi^\top(\mathbf{u}) \mathbf{L} \right) \right)^{-1} \left( \Phi^\top(\mathbf{u}) \mathbf{L} \right)^\top \Sigma_f^{-1} f(\mathbf{u}). \quad (64)$$

The linear combination:

$$s(\lambda_0) = \sum_{i=0}^{K-1} b_i \cdot P_i(g(\lambda_0)) \equiv \mathbf{P}_L(\lambda_0) \mathbf{b} \quad (65)$$

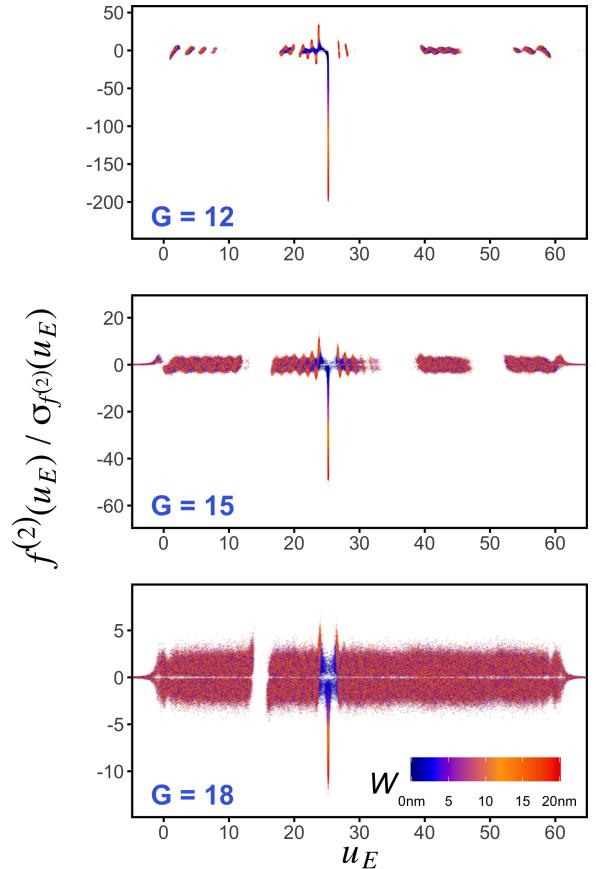
is then an approximation to the SPD at wavelength  $\lambda_0$ . The results of the inversion of the observational spectrum with this approach depend in the choice of  $\Delta$ ,  $K$ , and  $M$ . A good guideline for the choices is  $M > K$ , and the conditioning of the matrix to be inverted in Eq. (64) being small enough to be well off any numerical problems. For the tests carried out in this work, we used  $K = 5$ ,  $\Delta$  around 6 and  $M$  around 15.

Both methods for local smooth inversion of observational spectra have their advantages and disadvantages. The great advantage of the second approach, namely, the inversion using Legendre polynomials is the lower sensitivity to random noise. This approach, however, requires the explicit use of the response functions and systematic errors in the response function immediately affect the inversion. This is not the case of the first approach, which solves for the product of the external spectrum and the response function directly. As we show in Sect. 8, this is an advantage in the treatment of narrow spectral lines. So, generally, the second approach might be more advisable. In cases of strong lines and good S/N on the observational spectrum, potential systematic errors on the response function might become larger than the random noise and the first approach might be more suitable. We re-use the second approach in Sect. 9.2 – this time not for the purposes of inversion, but for a continuum estimation.

## 7. Spectral lines as local extrema

A spectral line might result in a local extremum of the observational spectrum  $f(u)$ . In particular, we assume that an emission line results in a local maximum and an absorption line in a local minimum. A necessary condition for a local extremum of  $f(u)$  is a root of  $f^{(1)}(u)$ . The position of such a root of the first derivative we denote with  $u_E$ . For  $f^{(2)}(u_E) < 0$ , the observational spectrum has a local maximum at  $u_E$ , and for  $f^{(2)}(u_E) > 0$ , a local minimum. We omit higher order derivative tests for simplicity here.

A local extremum is not, however, necessarily caused by a spectral line, as noise in the coefficients representing the spectra can be manifested in wavy structures with many local extrema. As a local extremum requires a root in the derivative, noise is



**Fig. 3.** Error-normalised value of the second derivative of the BP spectrum as a function of  $u_E$  for random spectra consisting of a  $2 \times 10^4$  K black body continuum and a narrow H $\beta$  line with random equivalent width  $W$ . The three panels correspond to  $G$ -magnitudes of 12 $m$ , 15 $m$ , and 18 $m$ , from top to bottom.

more likely to result in local extrema in regions of an observational spectrum where the first derivative of the true, noise-free spectrum is small, that is, where the gradient of the spectrum is low. With decreasing S/N, more and more regions with increasing gradient of the spectrum become likely to show local extrema due to noise. We therefore need an estimator quantifying the statistical significance of a local extremum, to distinguish between noise patterns and local extrema that actually are produced by the presence of a spectral line. For this purpose, we can use the value of the second derivative, normalised by their error, as discussed in Sect. 4.5. For noise-induced local extrema, this normalised value of the second derivative at the position of an extremum follows approximately a chi-distribution with two degrees of freedom, from which we can compute a  $p$ -value for a local extremum.

The noise-induced local extrema are illustrated for simulated BP spectra in Fig. 3. Here, the value of the second derivative at the root of the first derivative, normalised to its error, is shown for simulated spectra. The simulations in this work are done with the simple *Gaia* XP simulator described in Weiler et al. (2020). The simulated spectra are composed of a black body continuum with  $2 \times 10^4$  K, and a narrow H $\beta$  line with a random chosen equivalent width,  $W$ , which we define as

$$W := \int_0^\infty \frac{s(\lambda) - s_c(\lambda)}{s_c(\lambda)} d\lambda. \quad (66)$$

This definition of  $W$  has the opposite sign than in usually used definitions (e.g. Carroll & Ostlie 2017). We use this definition here to apply it to absorption and emission lines alike; we prefer emission lines to have a positive equivalent width and absorption lines to have a negative one. The result for  $2 \times 10^5$  simulated BP spectra for  $G$ -band magnitudes for the continuum of 12, 15, and 18, respectively, are shown. The colour scale shows the random value of  $W$  in each of the  $10^5$  spectra.

With increasing magnitude, and thus decreasing S/N, the number of local extrema increases. For the lowest magnitude, these local extrema caused by noise form predominantly in parts of the spectrum where the gradient is small. With increasing magnitude, the noise-related local extrema spread out. For  $G = 18^m$ , the only part in the BP spectrum free of noise-related local extrema is a narrow region around the wavelength cut-off position, where the gradient of the spectrum is largest. The simulated H $\beta$  line shows very clearly against the noise background of local extrema. And the larger the equivalent width of the simulated H $\beta$  line, the smaller the value of the second derivative at the root of the first derivative, normalised to its error.

If the spectral line is weak, namely, it has a small absolute value of  $W$ , and the slope of the spectrum is sufficiently large, an emission or absorption line might not result in a local extremum of the spectrum, however. A line might manifest itself in this case as a deformation of the spectrum. To detect such lines which are weak compared to the slope of the spectrum, we might consider higher order derivatives. The deformation of the spectrum results in a local extremum in curvature of the spectrum, namely, in  $f^{(2)}(u)$ . In this work, we first discuss the case of a spectral line as a local extremum in  $f(u)$  in detail. The following sections analyse this case for narrow and broad lines. Having worked out the necessary formalism for this case, it is easy to transfer the results to extrema in the curvature of the spectrum by simply increasing all derivatives involved by two. This latter case is analysed in detail in Sect. 10.

## 8. The limit of narrow lines

In this section, we consider the case of narrow spectral lines, where we understand ‘narrow’ in the sense that the intrinsic width of the spectral line in the SPD (measured e.g. by its FWHM) is small compared to the width of the LSF, expressed in wavelength. In this case we can neglect the width of the spectral line in the SPD, and write it as the sum of a continuum SPD and a weighted Dirac delta distribution:

$$s(\lambda) = s_c(\lambda) + s_l(\lambda) \quad (67)$$

$$\approx s_c(\lambda) + \alpha \delta(\lambda - \lambda_L). \quad (68)$$

Here,  $s_c(\lambda)$  is the continuum SPD,  $s_l(\lambda)$  the line SPD, and  $\lambda_L$  is the wavelength of the line. Using Eq. (34) and using the substitution of  $u'$  for the wavelength  $\lambda$ , we obtain in this case for the observed spectrum:

$$f(u) = \int_{-\infty}^{\infty} L(u, u') R(u') [s_c(u') + \alpha \delta(u' - u_L)] du' \quad (69)$$

$$= f_c(u) + \alpha R(u_L) L(u, u_L) \quad (70)$$

$$= f_c(u) + f_l(u). \quad (71)$$

The observed spectrum in the limit of a narrow spectral line is thus the sum of the observed continuum spectrum  $f_c(u)$  and a weighted LSF as the observed line  $f_l(u)$ .

### 8.1. Systematic errors in line positions

As we are considering spectral lines as local extrema in the internally calibrated spectrum, we have to analyse the relationship between the position of the spectral line,  $u_L$ , and the position at which the corresponding local extremum is occurring,  $u_E$ . Using Eq. (70), we obtain, for a local extremum, the necessary condition:

$$f^{(1)}(u_E) + \alpha R(u_L) L^{(1)}(u_E, u_L) = 0. \quad (72)$$

We assume  $\alpha \neq 0$ , and for convenience we introduce the abbreviation

$$\zeta(u) := -\frac{f_c^{(1)}(u)}{R(u_L) s_c(u_L)}. \quad (73)$$

At this point, it appears convenient to express the strength of the line with respect to the underlying continuum and use the equivalent width as defined in Eq. (66). In the approximation of a narrow line, and using Eq. (68), the equivalent width becomes:

$$W = \frac{\alpha}{s_c(\lambda_L)}. \quad (74)$$

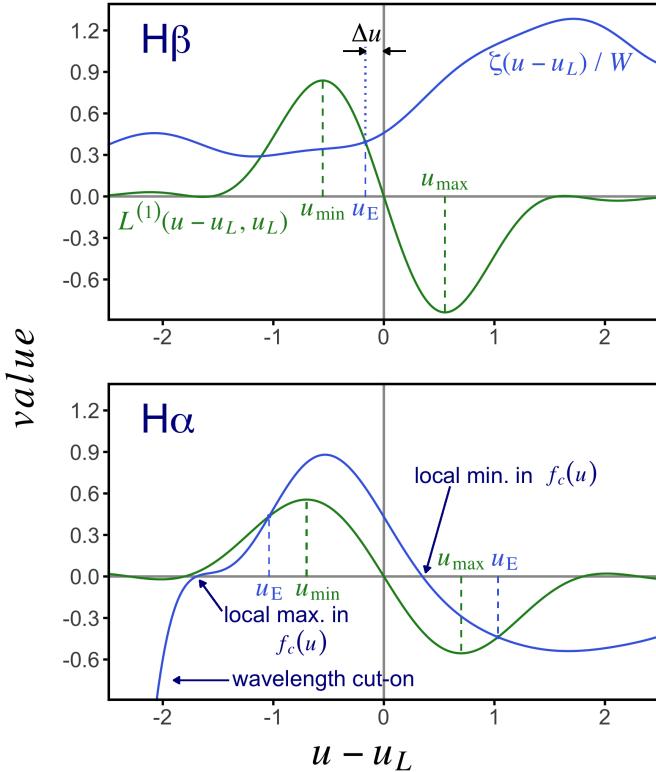
Substituting with  $u_L$ , we obtain

$$L^{(1)}(u_E, u_L) = \frac{\zeta(u_E)}{W}. \quad (75)$$

This is a strongly non-linear equation for the position of the local extremum,  $u_E$ , given the line position,  $u_L$ . Although this equation is not useful for practical computations in relating  $u_E$  and  $u_L$ , it nevertheless allows for a graphical illustration. In the limit of an extremely strong line, namely, for the limit  $W \rightarrow \infty$ , the right hand side of Eq. (75) becomes zero and the position of the extremum coincides with the position of the line. In the limit of negligible line strength, i.e.  $W \rightarrow 0$ , the right hand side becomes arbitrarily large and no more solution to Eq. (75) may exist. In this case, the line is not introducing a local extremum in the observational spectrum anymore. Between these two extreme cases, the position of the local extremum is shifted with respect to the line position.

This is illustrated with a simulated spectrum, again consisting of a black body continuum of  $2 \times 10^4$  K and narrow H $\alpha$  (in RP) and H $\beta$  (in BP) lines. The functions  $L^{(1)}(u, u_L)$  and  $\zeta(u)/W$  are shown in Fig. 4. A local extremum forms where the two functions intersect, this position is indicated  $u_E$  in the figure. A good range to look for extrema associated with a line is the range between the first local extrema of  $L^{(1)}(u, u_L)$ . The shift  $\Delta u$  is towards higher values in the underlying continuum spectrum for emission lines and towards lower values for absorption lines. As a criterion for associating a local extremum with a line at known wavelengths, and thus with known  $u_L$ , we might therefore require a shift  $\Delta u$  which is within the extrema of the first derivative of  $L(u, u_L)$ ; and towards lower values for a local minimum and towards larger values for a local maximum.

For H $\alpha$ , the situation is more complicated. For a faint line, no local extremum within the expected range of  $[u_{\min}, u_{\max}]$  forms. Instead, two local extrema, that is, a wavy-like structure, forms around the line position. This is a consequence of the more complex structure of the response curve close to the wavelength cut-on position of RP. This filter cut-on is realised by an interference filter on the RP prism, and interference filters tend to have a wavy transmission pattern in particular close to strong gradients as a cut-on flank. This pattern, which is present in the



**Fig. 4.** Illustration of the shift between the line position,  $u_L$ , and the position of local extrema,  $u_E$ , for a  $2 \times 10^4$  K black body background and an H $\beta$  line (top panel) and an H $\alpha$  line (bottom panel). The green lines show  $L^{(1)}(u - u_L, u_L)$  and the blue lines show  $\zeta(u - u_L)$  divided by  $W$ . We refer to the text for more details.

nominal response curves for the *Gaia* instruments (Jordi et al. 2010), interfere with the H $\alpha$  line. Even if the nominal pre-launch response curves might not be fully representative for the calibrated instrument, such patterns (as they are based on physical properties of the instrument) might still occur and complicate the analysis of faint H $\alpha$  lines.

In practice, the function  $\zeta(u_E)$  is unknown, and with it the shift  $\Delta u$  between the position of the extremum and of the line position. However, we may expect the shift  $\Delta u$  to be with the local extrema of the first derivative of the pseudo-LSF, at  $u_{\min}$  and  $u_{\max}$ . With an assumption for the pseudo-LSF, we can compute these positions and consider a local extremum in agreement with a line position if  $u_{\min} < u_E < u_{\max}$ .

## 8.2. Continuum model for narrow lines

Having  $f(u)$  available, we have to separate the continuum spectrum from the line contribution. This requires the construction of a continuum spectrum which is to be subtracted from the observational spectrum  $f(u)$ . To simplify this procedure, we use a simple approximation. We assume that the continuum spectrum is characterised by

$$f_c^{(2)}(u_E) = 0, \quad (76)$$

that is, we assume that the curvature of the observed continuum spectrum is zero at the maximum intensity of the spectral line. Differentiating Eq. (70) twice with respect to  $u$  results in

$$f^{(2)}(u) = f_c^{(2)}(u) + \alpha R(u_L) L^{(2)}(u, u_L). \quad (77)$$

With Eq. (76), this becomes

$$\alpha R(u_L) \approx \frac{f^{(2)}(u_E)}{L^{(2)}(u_E, u_L)} \quad (78)$$

and

$$f_c(u) \approx f(u) - \frac{f^{(2)}(u_E)}{L^{(2)}(u_E, u_L)} L(u, u_L). \quad (79)$$

With Eq. (76) as a condition for estimating the continuum spectrum, we can thus estimate the line spectrum from the ratio of the curvature of the observed spectrum at the position where the line is strongest in the observational spectrum and the curvature of the LSF at the same position. In case the true line position,  $u_L$  is not known, we may use the approximation  $u_L \approx u_E$ .

Equation (79) translates, with Eq. (40), into an equation for the vector of continuum coefficients,  $\mathbf{c}_c$ ,

$$\mathbf{c}_c \approx \mathbf{c} - \frac{f^{(2)}(u_E)}{L^{(2)}(u_E, u_L)} \mathbf{l}(u_L). \quad (80)$$

## 8.3. Equivalent widths for narrow lines

For a spectral line, we are not particularly interested in the strength of the line scaled with the response function,  $\alpha R(u_0)$ , as provided by Eq. (78), but, rather, we look for the equivalent width  $W$ . To obtain an expression for  $W$ , we use Eq. (74), with a substitution of  $u_L$ , and we combine this equation with Eq. (78), to obtain

$$W = \frac{f^{(2)}(u_E)}{L^{(2)}(u_E, u_L)} \frac{1}{s_c(u_L) R(u_L)} \equiv \frac{f^{(2)}(u_E)}{L^{(2)}(u_E, u_L)} \frac{1}{t_c(u_L)}. \quad (81)$$

For the determination of the equivalent width from the observational spectrum, we thus need to derive  $s_c(u_0) R(u_0)$  from the continuum spectrum provided by Eq. (79). Here, we can apply the inversion approaches as discussed in Sect. 6. Equation (54) or (65), together with a response function, can be combined with Eq. (81) to obtain the equivalent width of a spectral line.

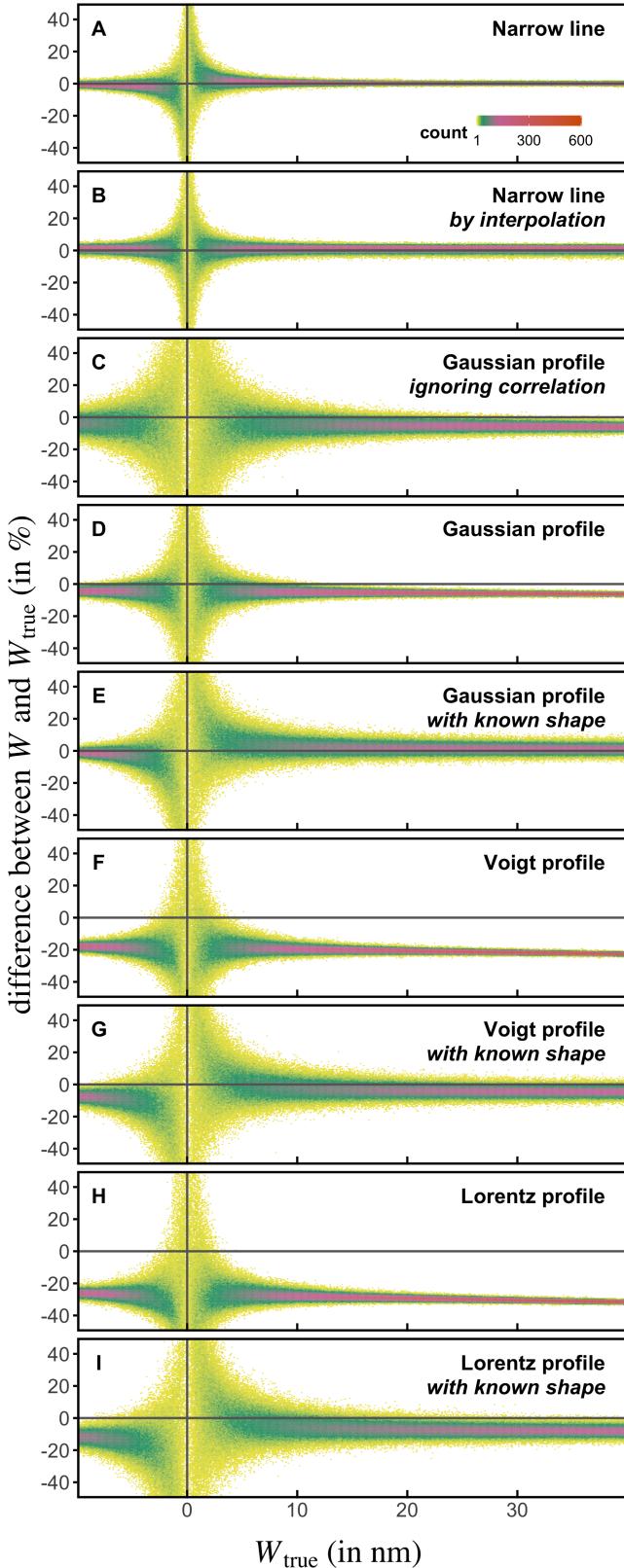
We can compute the uncertainty on the equivalent width in linear approximation using the Jacobian of  $W$  with respect to the coefficients of the XP spectrum. The elements of this are:

$$\frac{\partial W}{\partial c_i} = \frac{1}{L^{(2)}(u_E, u_L)} \left[ \frac{1}{t_c(u_E)} \frac{\partial}{\partial c_i} f^{(2)}(u_E) - \frac{f^{(2)}(u_E)}{t_c^2(u_L)} \frac{\partial}{\partial c_i} t_c(u_L) \right]. \quad (82)$$

The derivative of  $f^{(2)}(u_E)$  with respect to  $c_i$  is the  $i$ th element of the vector  $\Phi^T(u_E) \mathbf{D}^2$ . The expression for the derivative of  $t_c(u_L)$ , with respect to  $c_i$  depends on the approach chosen for inversion. Using Eq. (61), the derivative is the  $i$ th element of the first row of the matrix,  $\mathfrak{L}^{-1} \mathbf{H} - \mathfrak{L}^{-1} \mathbf{H} \Phi^T(u_E) \mathbf{D}^2 \mathbf{l}(u_L)$ . Using Eqs. (64) and (65), the derivative of  $t_c(u_L)$  with respect to  $c_i$  is the  $i$ th entry of

$$R(u_L) \mathbf{P}_L(\mathbf{u}) \left( (\Phi^T(\mathbf{u}) \mathbf{L})^T \Sigma_f^{-1} (\Phi^T(\mathbf{u}) \mathbf{L}) \right)^{-1} (\Phi^T(\mathbf{u}) \mathbf{L})^T \Sigma_f^{-1} \Phi(\mathbf{u}) \\ [1 - \mathbf{D}^2 \mathbf{l}(u_L)]. \quad (83)$$

A simple example computation for the illustration of the approach for narrow lines is shown in panel A of Fig. 5. As previously, simulations for a black body continuum for a



**Fig. 5.** Distributions of the relative difference between the computed equivalent width ( $W$ ) and the true equivalent width ( $W_{\text{true}}$ ) of  $\text{H}\beta$  lines with a  $2 \times 10^4$  K black body continuum for different assumptions on the line shape (narrow lines, Gaussian, Voigt, and Lorentz profiles) and different computation techniques. Each panel shows the result from  $5 \times 10^5$  random realisations of a spectrum for  $G = 12^{\text{m}}$  and ten transits. Positive values in the horizontal axis correspond to  $\text{H}\beta$  emission, negative values to  $\text{H}\beta$  absorption. For details see text in Sects. 8.3, 9.3, and 9.4.

temperature of  $2 \times 10^4$  K and a  $G$ -magnitude of  $12^{\text{m}}$  was used, and a narrow  $\text{H}\beta$  line was added, with a randomly chosen equivalent width. Then, the BP spectrum was computed and the coefficients of the Hermite representation computed from ten transits. The equivalent width was computed from these coefficients, and from it the relative difference to the true equivalent width. This was repeated for  $5 \times 10^5$  random realisations. The distribution of the relative difference for these simulations is shown as a function of the true equivalent width. In this computation, no a priori information on the line position was used, namely, the approximation  $u_L = u_E$  was used. Even with this approximation, a very good determination of  $W$  is possible, with systematic errors on the level of a percent.

#### 8.4. Blended narrow lines

In the following, we consider the case that two or more narrow lines are not sufficiently separated in pseudo-wavelength to neglect the overlap between them. We refer to this case as a blending of lines. The approach applied here is very similar to the case of unblended narrow lines, we can still use the condition expressed by Eq. (76) in this case. However, we have to modify Eq. (77) in such a way that it includes the contributions from other lines to a particular line under consideration. Taking  $n_l$  lines into account, with the line fluxes  $\alpha_i$ ,  $i = 1, \dots, n_l$  and a pseudo-wavelength position  $u_{L,i}$  for the  $i$ th line, we obtain:

$$f^{(2)}(u) = \sum_{i=1}^{n_l} \alpha_i R(u_{L,i}) L^{(2)}(u, u_{L,i}). \quad (84)$$

Evaluating this expression at the line positions  $u_{L,j}$  and writing it as a matrix equation, we get

$$\mathbf{f}^{(2)}(\mathbf{u}) = \mathbf{L}^{(2)} \mathbf{r}(\mathbf{u}), \quad (85)$$

with  $\mathbf{L}^{(2)}$  as the matrix with the elements  $L_{i,j}^{(2)} = L(u_{L,i}, u_{L,j})$ ,  $\mathbf{f}^{(2)}(\mathbf{u})$  as the vector containing the elements  $f^{(2)}(u_{L,i})$ , and  $\mathbf{r}(\mathbf{u})$  as the vector containing the elements  $\alpha_i R(u_{L,i})$ ,  $i = 1, \dots, n_l$  in both cases. Equation (85) is then solved for  $\mathbf{r}(\mathbf{u})$ , and the computation of the equivalent widths for the  $n_l$  lines continues in the same way as discussed for the unblended case previously.

#### 8.5. Upper limits for narrow lines

In the case of a non-detection of a particular spectral line, it might be of interest to derive an upper limit for the equivalent width of the non-detected line from the coefficients of the *Gaia* DR3 spectrum. Given the covariance matrix for the coefficients of the spectrum,  $\Sigma_c$ , and a statistical level of significance,  $p$ , we can compute what line, given by  $\alpha R(u_L)$ , can be added to the spectrum within the level of significance. With the coefficients for the pseudo-LSF at the position of the line,  $\mathbf{l}(u_L)$  given by Eq. (40), the distance between a spectrum and a spectrum with a narrow line added, expressed in numbers of standard deviations, is given by the Mahalanobis distance:

$$D_M = |\alpha R(u_L)| \sqrt{\mathbf{l}^\top(u_L) \Sigma_c^{-1} \mathbf{l}(u_L)}. \quad (86)$$

This statistics follows a chi-distribution with  $n$  degrees of freedom. With  $Q(p, n)$  the quantile function of the chi-distribution with  $n$  degrees of freedom and the coefficients for the

pseudo-LSF at the position of the line,  $\mathbf{I}(u_L)$ , given by Eq. (40), we obtain

$$|\alpha_{\text{limit}} R(u_L)| = \frac{Q(p, n)}{\sqrt{\Gamma(u_L) \Sigma_c^{-1} \mathbf{I}(u_L)}}. \quad (87)$$

This result can be used with Eq. (81) to derive an upper limit on the absolute value of the equivalent width of a non-detected line:

$$W_{\text{limit}} = \frac{|\alpha_{\text{limit}} R(u_L)|}{t(u_L)}. \quad (88)$$

With the cumulative distribution function for the chi distribution,  $P(n/2, p^2/2)$ , with  $P(k, x)$  the incomplete gamma function (Press et al. 2007a), the 0.6827 quantile is approximately 7.7094, while the 0.95 and the 0.99 quantiles are approximately 8.5622 and 9.0715, respectively.

## 9. Broad lines

### 9.1. Detection of broad lines

We now consider broad lines, with which we mean that the approximation of the line by a Delta function is no longer suitable. Instead, the intrinsic width of the line is so large that it cannot be neglected even for the low spectral resolution of the *Gaia* spectrophotometers. For the distinction between narrow and broad lines we require a measure for the line width. We can base this measure on roots of derivatives of the spectrum, if we consider inflection points that are related to a local extremum. A local extremum has an inflection point on either side, which we denote  $u_{i,1}$  and  $u_{i,2}$ , and base the estimator for the line width on the distance between these two points. As inflection points are roots of the second derivative of the observational spectrum, for a narrow line the distance between the inflection points is in a first-order approximation the distance between the innermost inflection points of the pseudo-LSF. With increasing intrinsic line width, the distance between the inflection points increases with respect to the corresponding distance for the pseudo-LSF. We therefore normalise the distance between the innermost two inflection points of the pseudo-LSF, and use this dimensionless quantity, which we denote as  $\omega$ ,

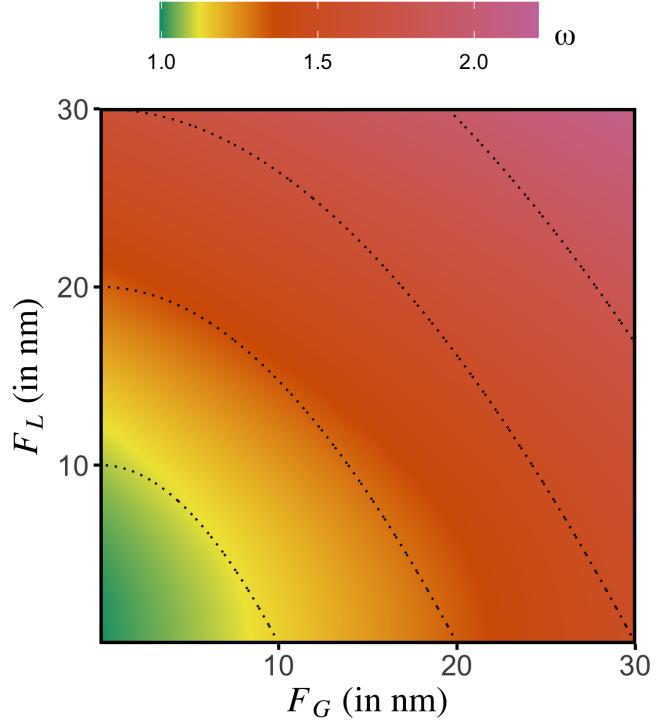
$$\omega = \frac{u_{i,2} - u_{i,1}}{u_{i,2}^{[\text{LSF}]} - u_{i,1}^{[\text{LSF}]}} , \quad (89)$$

as the measure for line widths. If this quantity is significantly larger than one, we may consider the line broad. For the uncertainty on  $\omega$ , we have to take the correlations between the errors on  $u_{i,1}$  and  $u_{i,2}$  into account, which we obtain from the corresponding elements of the covariance matrix on the roots of the second derivative,

$$\sigma_\omega = \frac{\sqrt{\Sigma_{11} + \Sigma_{22} - 2\Sigma_{12}}}{u_{i,2}^{[\text{LSF}]} - u_{i,1}^{[\text{LSF}]}} . \quad (90)$$

Figure 6 illustrates  $\omega$  for Voigt profiles, as a function of the full width at half maximum (FWHM) of the Gaussian and Lorentz contributions. On the horizontal axis is the FWHM of the Gaussian,  $F_G$ , on the vertical axis is the FWHM of the Lorentz profile,  $F_L$ . The FWHM of the resulting Voigt profile,  $F_V$  is in good approximation (Olivero & Longbothum 1977)

$$F_V = 0.5356 \cdot F_L + \sqrt{0.2166 \cdot F_L^2 + F_G^2} . \quad (91)$$



**Fig. 6.** Line width parameter,  $\omega$ , for the Voigt profiles, as a function of the FWHM of the Gaussian contribution (on the horizontal axis) and the Lorentz contribution (on the vertical axis). The dotted lines indicate iso-contours in FWHM of 10, 20, 30, and 40 nm. See text for details.

The pseudo-LSF used here is based on the optical LSF for *Gaia*, blurred with a Gaussian with a standard deviation of 0.64, which is the LSF model used in Sect. 11 of this work. The LSF changes with wavelength, and for this example we use the wavelength of H $\beta$ . For lines with FWHM below about 20 nm, the increase in  $\omega$  is only a few ten per cent, showing the intuitive result that low resolution spectra are therefore not particularly sensitive to line widths.

### 9.2. Continuum model for broad lines

In the case of broad lines, estimating the continuum underlying a spectral line requires some form of interpolation. We would like to do the interpolation on internally calibrated spectra, where we detect a line, but at the same time, we would like to ensure that the interpolation on the internally calibrated spectrum corresponds to a smooth interpolation on the externally calibrated spectrum as well. To achieve this, we may adapt the approach already discussed in Sect. 6 and choose smooth interpolation functions in the external regime, then compute the observational spectra corresponding to these interpolation functions. We can then use these artificial observational spectra of the interpolation functions to find an interpolation over some pseudo-wavelength interval for an XP spectrum. For the interpolation functions, we may choose again the first  $M$  Legendre polynomials over a wavelength interval that essentially covers the entire range of non-zero response in BP and RP, respectively.

With the interpolation functions at hand, we have to decide about the interpolation interval  $[u_{\text{lower}}, u_{\text{upper}}]$  under the line. The choice of this interval is crucial for a good estimate of the continuum. Choosing this interval too narrow will result in an under-subtraction of the continuum, while choosing it too wide is likely to make the interpolation unreliable. We decided to tie

the interval to the widths of the line, as expressed by the distance between the two inflection points on either side of the local extremum,  $u_{i,2} - u_{i,1}$ , as discussed in the previous subsection. We chose the interpolation interval symmetric around the local extremum, with a width corresponding to a factor of  $x$  times  $\Delta$ . In experimenting with simulated spectra with different black body continua and Voigt profiles for the lines, we found a value around  $x \approx 5.5$  is a good choice.

For the interpolation, we again have to impose conditions on the function at the boundaries of the interpolation interval and different points of the function itself on either side of the interval. So we selected a small number of  $m$  points of either side of the interval, separated by a value  $\delta$ , such that we can obtain a sequence of  $2m + 2$  points  $\mathbf{u} := u - m\delta, u - (m-1)\delta, \dots, u_{\text{lower}}, u_{\text{upper}}, u + \delta, \dots, u + m\delta$ . We might not force an exact solution for the interpolation, but chose  $2m + 2 > M$ , and do a regression, as it reduces the noise. The noise on  $f(\mathbf{u})$  however is likely to be correlated, with a covariance matrix  $\Sigma_f = \Phi^T(\mathbf{u}) \Sigma_c \Phi(\mathbf{u})$ . We thus have to determine the coefficient vector for the interpolation functions,  $\mathbf{b}$ , such that it minimised the squared Mahalanobis distance as a generalised least squares problem, as already discussed in Sect. 6. The product,  $\mathbf{L}\mathbf{b}$ , then provides the coefficients for the continuum approximation on the interval  $[u_{\text{lower}}, u_{\text{upper}}]$ .

### 9.3. Equivalent widths for broad lines

Having separated the line and the continuum, the computation of the equivalent width requires the inversion of both components of the spectrum. The inversion of the continuum is trivial, as the coefficients,  $\mathbf{b}$ , have been constructed from the observational spectra corresponding to Legendre polynomials in the regime of the absolute calibration. However, the inversion of the line signal is intrinsically an ill-posed problem and any inversion might be subject to significant systematic errors. We therefore consider a simplification and compute an estimate for the equivalent width without inversion of the internally calibrated spectrum, namely, using the approximation

$$W \approx \int_{u_1}^{u_2} \frac{f(u) - p(\mathbf{b}, u)}{p(\mathbf{b}, u)} \frac{d\lambda}{du} du, \quad (92)$$

with  $p(\mathbf{b}, u)$  the continuum approximation. The approximate computation of  $W$  from the internally calibrated spectrum instead of the SPD has been tested using simulations for narrow lines and Voigt profiles with different underlying continua. The resulting systematic errors were in the order of some percent. In all test cases, the systematic errors from continuum interpolation were at least three times larger than the error resulting from the approximate computation. This approximation might be sufficiently good for general broad lines.

Example computations for the determination of  $W$  with this approach are shown in Fig. 5. The simulated case is analogous to the case considered for narrow lines, a black body continuum for  $2 \times 10^4$  K and an H $\beta$  line are used, for  $G = 12^{\text{m}}$  and ten transits, and  $5 \times 10^5$  random realisations of the BP spectrum with randomly selected values for  $W$ . Panel B shows the distribution of the relative difference between the computed and the true equivalent width as a function of true equivalent width. The performance of the broad line approach is similar to the use of the narrow line approach, shown in panel A. The random noise is slightly higher, however.

Panel C of the same figure shows the distribution for an H $\beta$  line with a Gaussian profile and a FWHM of 15 nm. In this case, the Legendre interpolation was done without taking correlations into account. A systematic error in  $W$  can be seen, resulting mainly from a systematic underestimation of the continuum in the fit. Panel D shows the same case with taking correlations into account. The random noise is suppressed as compared to the case without including correlations, but the systematic errors from the continuum interpolation remain. Panel F of Fig. 5 shows the results for a Voigt profile for the H $\beta$  line, with a FWHM of 15 nm and equal FWHM for the Gaussian and Lorentz contributions to the Voigt profile. The systematic errors from the continuum interpolation increase as compared to a Gaussian line profile. Panel H shows the distribution of relative difference in  $W$  for a Lorentz profile with a FWHM of 15 nm. Here, the systematic errors from continuum interpolation increase even more. The reason for the increasing systematic errors from Gaussian to Voigt to Lorentz profile are the increasingly flat wings of the spectral line, which make the choice of a suitable interpolation interval for the continuum increasingly difficult. As a result, the fit is likely to over- or under-subtract the line from the continuum. This problem can be strongly reduced if the shape of the line can be assumed known a-priori, a case we discuss in the following section.

### 9.4. Broad lines with known line shapes

In cases where the position and shape of a broad line can be assumed to be known a priori, the analysis of the line in *Gaia* DR3 XP spectra is simplified. In this case, the interpolation step for the continuum can be replaced by an approach more similar to the case of narrow lines. Assuming a line profile  $p(\lambda)$ , normalised such that its integral is unity, we compute the observational spectrum corresponding to this profile:

$$P(u) := \int_0^\infty L(u, \lambda) R(\lambda) p(\lambda) d\lambda. \quad (93)$$

Analogously to the approach in Sect. 8.2, we then use the second derivative at the line position to determine the scaling factor  $\alpha$  that describes the line strength:

$$\alpha = \frac{f^{(2)}(u_L)}{P^{(2)}(u_L)}. \quad (94)$$

The approximation for the continuum spectrum is:

$$f_c(u) \approx f(u) - \alpha P(u). \quad (95)$$

As the line profile is assumed to be known, no numerical inversion of the profile is required. An inversion of the continuum over the pseudo-wavelength range covered by the line is, however, required. As the continuum can be assumed smooth, the inversion for  $t_c(u)$  using the Taylor approximation as lined out in Sect. 6 might allow for a good smooth approximation on a grid of points in the interval of interest. With knowledge of the response function, we can thus obtain an estimate for  $s_c(\lambda)$  and compute the equivalent with by numerical integration of the expression:

$$W = \alpha \int_{\lambda_1}^{\lambda_2} \frac{p(\lambda)}{s_c(\lambda)} d\lambda. \quad (96)$$

The inversion of the continuum point-wise is increasing the computational effort, but in exchange, it is possible to remove systematic errors resulting from the interpolation of the continuum. Illustrations are shown in panels E, G, and I of Fig. 5, for a Gaussian, a Voigt, and a Lorentz line profile, respectively. The equivalent width of an H $\beta$  line with a black body continuum of  $2 \times 10^4$  K and a FWHM of 15 nm is computed assuming the line position and profile known. The distribution of the relative difference between the computed and true  $W$  is shown for  $5 \times 10^5$  random generations for  $G = 12^m$  and ten transits. The systematic error in  $W$  that results when interpolating the continuum without using a priori knowledge about the line, a case shown in panels D, F, and H of the same figure, is strongly reduced. The reduction in systematic error is strongest for a Gaussian line profile and lowest for the Lorentz profile. The differences in the remaining systematic error mainly arise from the interval of integration. From Gaussian and Voigt to Lorentz profile, the wings of the line become wider, and the systematic errors from an integration over a fixed interval become larger. The reduction in systematic error however comes with an increase in the random error. If the position and shape of the line is known a priori, it is possible to compute an upper limit for a non-detected line, following the approach discussed for narrow lines in Sect. 8.5. Furthermore, we may also treat the problem of blended lines analogously to the case of narrow lines described in Sect. 8.4.

## 10. Lines as extrema in higher derivatives

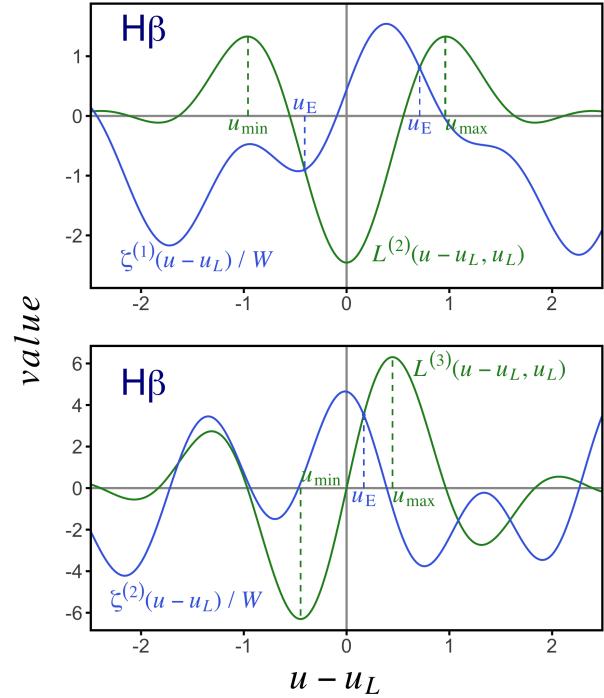
As we have seen in Sect. 8.1, no local extremum in the observational spectrum is formed by a line if the function  $\zeta(u)/W$  is not intersecting with the function  $L^{(1)}(u, u_L)$ , the first derivative of the LSF or pseudo-LSF. As can be seen from Eq. (73), large values for  $\zeta(u)/W$  are a result of an overly large ratio between the slope of the continuum spectrum,  $f_c^{(1)}(u)$ , and the line's equivalent width,  $W$ . Thus, a large slope and a weak line, namely, a line with a small absolute value of  $W$ , might prevent the formation of a local extremum. However, a line might still be able to produce a local extremum in higher derivatives. We are therefore considering higher derivatives of Eq. (75), stating the condition for a local extremum.

The function  $L^{(2)}(u, u_L)$  has no root at  $u_L$ , and a line might show as two local extrema in  $f^{(1)}(u)$  on both sides of  $u_L$ . These two points correspond to inflection points of the observational spectrum. This situation is illustrated in the top panel of Fig. 7, for a simulated H $\beta$  line with a black body continuum of 6000K. The shapes of the higher derivatives of the pseudo-LSF and  $\zeta(u)/W$  become more complex, and the positions of  $u_E$  might vary strongly. A good indicator for inflection points being related to a line is whether the values of  $u_E$  are within the first local extrema of  $L^{(2)}(u_L)$ , indicated by  $u_{\min}$  and  $u_{\max}$  in Fig. 7.

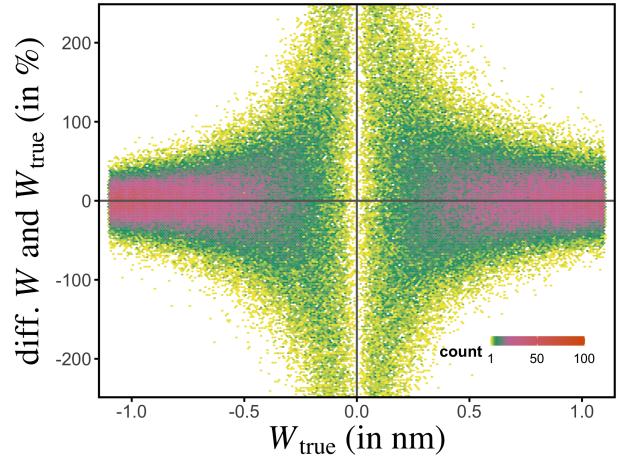
The function  $L^{(3)}(u, u_L)$  has a root at  $u_L$ , as the curvature of the pseudo-LSF is maximal at its centre. We might therefore consider the case of local extrema in  $f^{(2)}(u)$ . This case is illustrated in the bottom panel of Fig. 7. For the determination of the equivalent width of a narrow line, we might therefore use the approach of Eq. (81), with the orders of the involved derivatives increased by two,

$$W = \frac{f^{(4)}(u_E)}{L^{(4)}(u_E, u_L) t_c(u_L)} \frac{1}{W} \quad (97)$$

As an example, we again consider a simulated H $\beta$  line in BP, with an underlying black body continuum for a temperature of 6000 K. This simple test case does not result in a local extremum



**Fig. 7.** Illustration of the positions of local extrema in the first derivative (top panel) and second derivative (bottom panel) for a simulated H $\beta$  line with a 6000K black body continuum. See text for details.



**Fig. 8.** Distribution of the difference between true and derived equivalent width for simulated weak and narrow H $\beta$  lines. See Sect. 10 for details.

for  $-1.1 \text{ nm} \leq W \leq 1.1 \text{ nm}$ . We derive the equivalent width from the extremum in the second derivative of  $f(u)$ . The resulting distribution of differences between the true and the computed  $W$ , as a function of true  $W$ , is shown in Fig. 8. This figure shows the result from  $2 \times 10^5$  random realisations of an object, with  $G = 12^m$  and ten transits. Unbiased recovery of the equivalent width is possible, with some tens of percent uncertainty, also from the second derivative of a spectrum. As expected for faint lines and the use of higher derivatives, the random noise is considerably larger than for the case of deriving equivalent widths from local extrema. However, only if  $|W|$  falls below about 0.5 nm, the noise becomes so large that no meaningful interpretation of the spectrum around the H $\beta$  line becomes possible in the simulated case.

## 11. Example applications to *Gaia* DR3 spectra

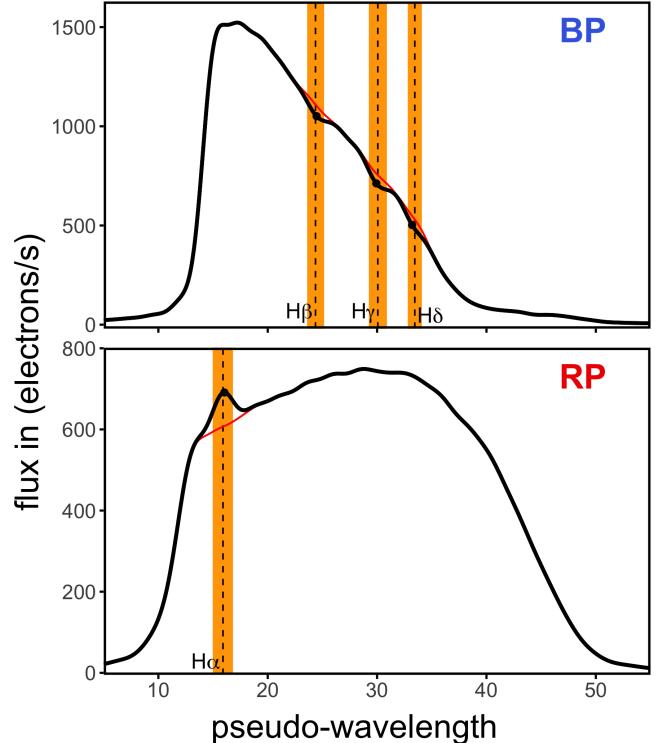
In the following, we apply the methods derived in the previous sections to some example cases among the *Gaia* DR3 low-resolution spectra. These examples were not selected for their scientific output, but serve solely for illustrating the computational method and illustrating its potential. To cover a wide range of applications, the examples in this section include the analysis of lines in the zero and second derivative, absorption and emission lines, narrow and broad lines, and blended and non-blended cases, for hydrogen Balmer lines, He I lines, and broad interstellar absorption bands.

For the instrument model, we used the dispersion functions and response curves from the configuration files in the *Gaia* software tool GaiaXPy<sup>1</sup>. For easy computation, we represent the tabulated dispersion functions by spline functions. The configuration files contain no model for the (pseudo-)LSF. To construct a simple substitute, we measure the widths  $\omega$  of H $\alpha$  and H $\beta$  lines in a large set of sources known for strong hydrogen absorption and emission with respect to the optical LSF of *Gaia*, broadened with pixel sampling and TDI effects (cf. Weiler et al. 2020). We then applied a Gaussian blur to this LSF until the width of the resulting blurred LSF matched the observed widths for H $\alpha$  and H $\beta$ . The standard deviation of the Gaussian blur applied is 0.64, independent of wavelength. From this simple LSF model, we compute the pseudo-LSF by development in Hermite functions. In the following, we show that this simplistic approach is already sufficient to obtain good results. However, the results presented should be understood as preliminary, as they are likely to improve with the use of a realistic full instrument model when available in the future.

The analysis of the H $\alpha$  line of a large set of sources with strong H $\alpha$  emission showed a systematic difference between the pseudo-wavelength at which the RP dispersion function predicts this line and the true line position by about 0.3 in pseudo-wavelength. We corrected for this systematic effect by artificially shifting the value of  $u_L$  for H $\alpha$  by +0.3 with respect to  $u_L$  predicted by the dispersion relation used. A similar effect seems to apply at long wavelengths for the BP instrument, and we applied a correction of +0.35 for the He I line at 587.7 nm as well.

### 11.1. Example: Balmer emission and absorption lines

As a first example, we considered the star *Gaia* DR3 505171240162594560, IRAS 02058+5719, an object with  $G \approx 14.187^m$  and  $G_{BP} - G_{RP} \approx 0.495^m$ . This object has been selected from the IGAPS data set (Monguió et al. 2020), where H $\alpha$  emission is reported. The analysis of this object is illustrated with the BP and RP spectra as a function of pseudo-wavelength in Fig. 9. The expected positions of the H $\alpha$  to H $\delta$  lines are indicated by dashed lines in this figure. The intervals  $[u_{\min}, u_{\max}]$ , within which an extremum is consistent with resulting from the Balmer lines, are highlighted. The intervals are based on the two inflection points of the LSF on both sides of the maximum position of the LSF, respectively, as discussed in Sect. 8.1. For this star, we find a local maximum in RP consistent with H $\alpha$ , along with three local extrema in the second derivative, consistent with H $\beta$ , H $\gamma$ , and H $\delta$ . All four local extrema have  $p$ -values larger than 0.999 and it is thus very unlikely that these are noise patterns. The red lines show the continuum model for narrow lines according to Eq. (79). We use Eq. (81) for the computation of the



**Fig. 9.** Example for the analysis of H $\beta$  to H $\delta$  in BP (top panel) and H $\alpha$  in RP (bottom panel) for source *Gaia* DR3 505171240162594560. The nominal line positions are indicated by dashed lines, the shaded regions indicate the range in pseudo-wavelength within which a local extremum is considered in agreement with being a hydrogen Balmer line. For BP, the wavelength increase from right to left (with decreasing pseudo-wavelength) for RP from left to right (with increasing pseudo-wavelength). Black curves show the XP spectra, the red curves show the continuum model. Black dots indicate the positions of local extrema in the second derivative (H $\beta$  to H $\delta$ ) and zero derivative (H $\alpha$ ).

equivalent widths of the lines, using the inversion according to Eq. (54). The results for the four Balmer lines in the XP spectra of this source are summarised in Table 2. The difference between the wavelength corresponding to the local extrema,  $\lambda_e$ , and the wavelengths of the corresponding Balmer lines is well within the range of expected systematic deviations. The estimated line widths,  $\omega$ , are in a good agreement with the assumption of narrow lines. Only the H $\gamma$  line shows a deviation above a  $1-\sigma$  level, but mildly so.

### 11.2. Example: H $\alpha$ absorption for CALSPEC stars

In order to test the accuracy of the equivalent widths derived from *Gaia* XP spectra with the method presented in this work, we compared the values of  $W$  with values derived from higher resolution spectra of the same sources. For our set of stars with high-quality spectra with higher resolution, we used the CALSPEC data (Bohlin et al. 2014, 2020) and we considered the H $\alpha$  line. In the visible wavelength range, the CALSPEC spectra are obtained with the *Hubble* Space Telescope/Space Telescope Imaging Spectrograph (HST/STIS) and the resolving power at the H $\alpha$  line is approximately 673. The data set provides 44 stars with XP spectra available in DR3 and a detection of H $\alpha$  in the RP spectrum with a  $p$ -value larger than 0.8.

In the CALSPEC spectrum of each star, a third-degree polynomial was fitted to continuum intervals on both sides of the

<sup>1</sup> <https://gaia-dpc.github.io/GaiaXPy-website/>, see De Angeli et al. (2023).

**Table 2.** Line, XP instrument (BP or RP), position in pseudo-wavelength ( $u_E$ ), and shift with respect to expected line position ( $\Delta u$ ), wavelength ( $\lambda_e$ ) for the local extrema, line width parameter,  $\omega$ , estimated equivalent width ( $W$ ), and  $p$ -value for the lines in the XP spectra of individual sources analysed in this work.

Line	XP	$u_E$	$\Delta u$	$\lambda_e$ (nm)	$\omega$	$W$ (nm)	$p$
Gaia DR3 505171240162594560 (IRAS 02058+5719)							
H $\alpha$	RP	$16.034 \pm 0.037$	0.127	$657.34 \pm 0.25$	$0.984 \pm 0.030$	$2.35 \pm 0.19$	1.000
H $\beta$	BP	$24.466 \pm 0.093$	0.100	$485.20 \pm 1.01$	$1.062 \pm 0.106$	$-1.13 \pm 0.32$	1.000
H $\gamma$	BP	$29.927 \pm 0.073$	-0.139	$435.25 \pm 0.57$	$1.229 \pm 0.154$	$-0.98 \pm 0.26$	1.000
H $\delta$	BP	$33.206 \pm 0.057$	-0.242	$411.86 \pm 0.37$	$0.972 \pm 0.036$	$-1.47 \pm 0.28$	1.000
Gaia DR3 426306363477869696 (HT Cas)							
He I (706.7 nm)	RP	$21.841 \pm 0.333$	-0.458	$702.95 \pm 2.70$	$0.76 \pm 0.12$	$0.55 \pm 2.31$	0.420
H $\alpha$	RP	$16.005 \pm 0.048$	0.098	$657.14 \pm 0.33$	$1.004 \pm 0.050$	$12.09 \pm 1.89$	1.000
He I (587.7 nm)	BP	$16.932 \pm 0.550$	-0.621	$586.25 \pm 10.10$	$0.65 \pm 0.25$	$1.51 \pm 2.26$	0.270
H $\beta$	BP	$24.319 \pm 0.086$	-0.048	$486.80 \pm 0.93$	$0.961 \pm 0.078$	$6.93 \pm 1.61$	1.000
He I (447.3 nm)	BP	$27.816 \pm 0.274$	-0.624	$452.67 \pm 2.40$	$0.764 \pm 0.079$	$2.17 \pm 1.40$	0.700
H $\gamma$	BP	$29.706 \pm 0.161$	-0.360	$436.98 \pm 1.27$	$0.858 \pm 0.077$	$5.86 \pm 1.99$	0.969
H $\delta$	BP	$33.067 \pm 0.169$	-0.380	$412.77 \pm 1.12$	$0.859 \pm 0.091$	$5.65 \pm 2.01$	0.957
H $\epsilon$	BP	$35.768 \pm 0.126$	0.174	$396.12 \pm 0.73$	$0.833 \pm 0.048$	$3.30 \pm 0.89$	0.978
Gaia DR3 2067888218857234304 (BD +41 3807, ALS 11438)							
770 nm DIB	RP	$28.968 \pm 0.105$	-0.303	$766.92 \pm 1.04$	$1.15 \pm 0.16$	$-1.38 \pm 0.55$	1.000

line and then used as an approximation of the continuum. With the continuum approximation, the FWHM of the line is derived and the equivalent width computed by applying Eq. (66). The 44 stars cover a wide range of parameters, with equivalent widths between -0.2 nm and -4.4 nm, FWHM of the lines ranging from 1.5 nm to 7.8 nm,  $G$  band magnitudes from 14.7<sup>m</sup> to 4.3<sup>m</sup>, and  $G_{\text{BP}} - G_{\text{RP}}$  colours ranging from -0.54 to 0.91.

A comparison between the equivalent widths derived from RP spectra and those from the CALSPEC spectra is shown in Fig. 10. The upper panel of this figure shows the equivalent widths derived with the narrow line approximation as described in Sect. 8.3. While this approximation is providing a good agreement with the reference values from CALSPEC spectra for weak absorption lines, systematic deviations are obvious for stronger absorption lines. Already for equivalent widths around -1.5 nm the systematic error is clearly visible.

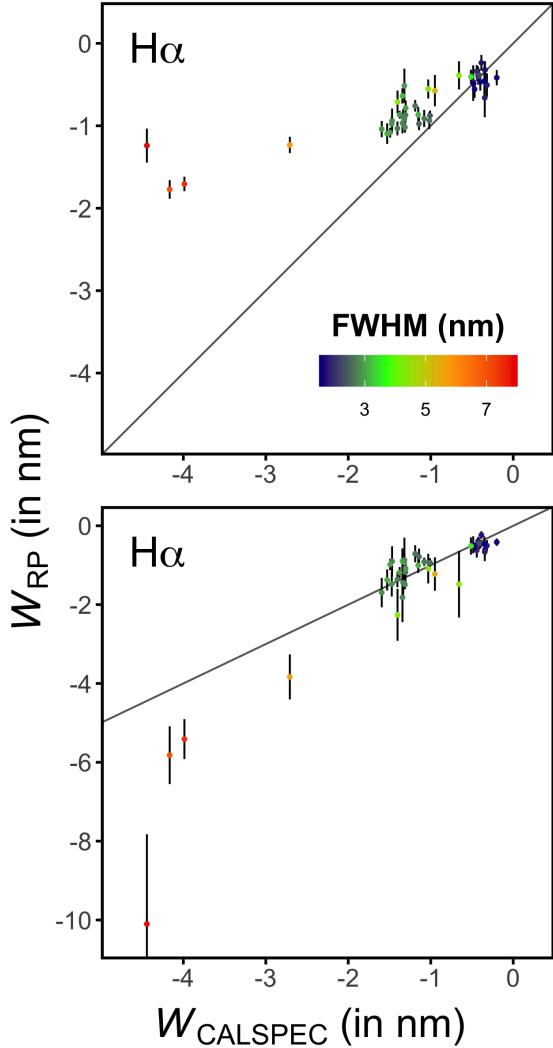
The bottom panel of Fig. 10 shows the comparison for equivalent widths computed by assuming a Lorentz line shape for the H $\alpha$  line for all lines with a value of  $\omega$  larger than one and using the approach described in Sect. 9.4. The width of the Lorentz profile was chosen such that the value of  $\omega$  derived for each source from the RP spectrum is met by the model Lorentz line. With this approach, the systematic error is strongly reduced, while the random error, which has been estimated by using bootstrapping, is increased. This is consistent with the results obtained for simulated spectra in Sect. 9.4. Only the strongest and broadest lines in Fig. 10 might be affected by systematics. However, besides the systematic errors resulting from the computational approach used, also the simplistic instrument model employed in this work might contribute to systematic errors for the broad and strong absorption lines.

### 11.3. Example: H $\alpha$ and H $\beta$ absorption in open clusters

As an example for the application of the method presented in this work to a larger set of sources, we selected open clusters and determined the H $\alpha$  and H $\beta$  equivalent widths. For

the test, we selected the three open clusters NGC 3114, Mel 20 ( $\alpha$  Persei), and NGC 3532, with the cluster members taken from Cantat-Gaudin & Anders (2020). For the test, we applied strict quality filters to the data, including only sources with fractions of blended and contaminated transits below 0.1, a small renormalised unit weight error below 1.4, a small number of multi-peaked transits or transits with complex windows, and an absolute value of the corrected flux excess below  $1.5\sigma_{C^*}$  according to Riello et al. (2021). This filtering resulted in 851, 552, and 820 sources for the three open clusters, respectively. Figure 11 shows the equivalent widths as a function of effective temperature for the three open clusters. The effective temperatures are the field `teff_gspphot` from the DR3 `gaia_source` and astrophysical parameters table (Creevey et al. 2023; Fouesneau et al. 2023). The  $p$ -value of the hydrogen lines is shown as a colour scale, and lines which are local extrema in the XP spectra are shown as circles, the ones which are extrema in the second derivative of the XP spectrum as triangles. The generally expected behaviour can be seen in the results for all three clusters, and for H $\alpha$  and H $\beta$ . Weaker lines tend to be detected in the second derivative, and the statistical significance of the lines, expressed by the  $p$ -value, increases as the lines are getting stronger. Also, the errors on the equivalent widths tend to become smaller if the line strength increases. Also as expected, a maximum hydrogen absorption around an effective temperature of 9000 K can be seen in all three clusters and for H $\alpha$  and H $\beta$ .

The open cluster NGC 3114 shows one prominent outlier from the general trend of H $\alpha$  with effective temperature, which is Gaia DR3 5256686653442846336 (HD 87471). This star with an effective temperature of about 10<sup>4</sup> K has no significant H $\alpha$  line, with  $W = (0.087 \pm 0.100)$  nm, and a  $p$ -value of 0.414. The H $\beta$  absorption for this star is also somewhat lower for this effective temperature, with  $W = (-0.87 \pm 0.23)$  nm, with a  $p$ -value of  $p = 0.9999$ . No spectral peculiarities have been reported for this star. It is classified as a sub-giant (Houk & Cowley 1975), and this classification is consistent with the fact that the star is

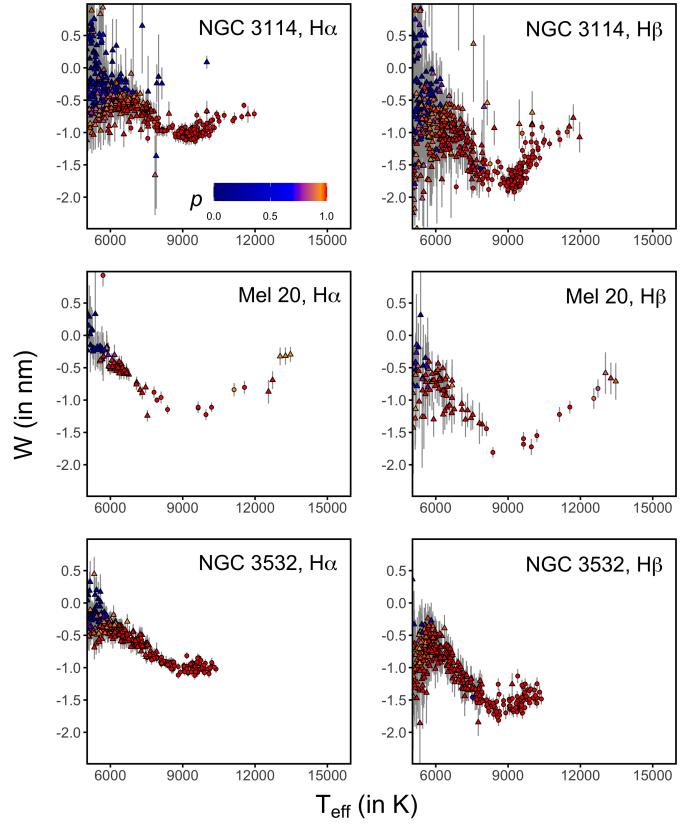


**Fig. 10.** Comparison of the equivalent widths of the H $\alpha$  line for CALSPEC stars, derived from *Gaia* RP spectra,  $W_{\text{RP}}$ , and from CALSPEC spectra,  $W_{\text{CALSPEC}}$ . Top panel: Results obtained with the narrow line approximation. Bottom panel: Results for assuming a Lorentz line profile. The colour scale gives the FWHM of the lines, as measured in CALSPEC spectra, and it is the same for both panels.

approximately 1.5 magnitudes brighter than the main sequence stars of comparable colour in the same cluster. A lower hydrogen absorption as compared to the other stars in the cluster might thus result at least partially from a lower surface gravity of a sub-giant.

#### 11.4. Example: Blended narrow lines

As an example for the treatment of blended narrow lines we chose the cataclysmic variable HT Cassiopeiae, *Gaia* DR3 426306363477869696. This star, with a  $G \approx 16.313^{\text{m}}$  and a  $G_{\text{BP}} - G_{\text{RP}} \approx 0.664^{\text{m}}$ , is known to show strong emissions from hydrogen and neutral helium (Marsh 1990). We therefore studied the hydrogen Balmer lines and three strong He I lines within the wavelength range covered by BP and RP, at wavelengths of 447.3 nm, 587.7 nm, and 706.7 nm, which are the vacuum Ritz wavelengths from the National Institute of Standards and Technology Atomic Spectra Data Base (Kramida et al. 2021). The XP spectra for this source as a function of pseudo-wavelength are shown in Fig. 12 for illustration. The derived line parameters



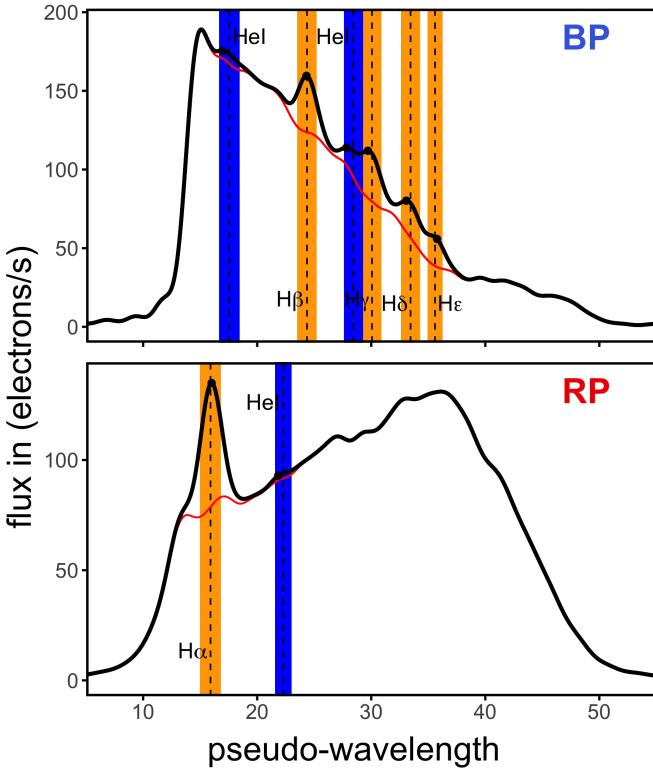
**Fig. 11.** H $\alpha$  and H $\beta$  equivalent widths against the effective temperature for the three open clusters NGC 3114, Mel 20, and NGC 3532. The colour scale shows the  $p$ -value and is common to all panels. Circles correspond to extrema in the zero derivative. Triangles correspond to the extrema in the second derivative.

are presented in Table 2. All hydrogen Balmer lines for H $\alpha$  to H $\epsilon$  are clearly detected as emission lines, with  $p$ -values above 0.95 in all cases. There are local extrema in the XP spectra or its second derivative that are consistent with being He I emissions. However, the statistical significance of these extrema is rather low. The  $p$ -values for the lines at 706.7 nm and 587.7 nm are 0.420 and 0.270, respectively, and the extrema are thus well in agreement with the status of noise. The extremum consistent with the 447.3 nm He I line has a somewhat larger  $p$ -value of 0.700, and might be a tentative detection of He I.

The lines of H $\beta$ , He I at 447.3 nm, and Hy to He $\epsilon$  are blended, and the equivalent widths of these lines are computed using the approach presented in Sect. 8.4. With five blended lines, the continuum model is constructed over a wide range of pseudo-wavelengths, shown as the red line in Fig. 12. The continuum model has a wavy structure, partly because it is affected by the noise. Additionally, the assumptions on the LSF made in this work are simplified and, in particular, the outer parts of the pseudo-LSF are affected by these simplifications. When combining the outer parts of the pseudo-LSF in case of blended lines, these simplifications might therefore additionally affect the resulting solution. An improved LSF model might therefore allow for improvements in the future.

#### 11.5. Example: Broad interstellar absorption line

As an example for a broad line, we considered the interstellar band at approximately 770 nm in RP. Maíz Apellániz et al. (2021)

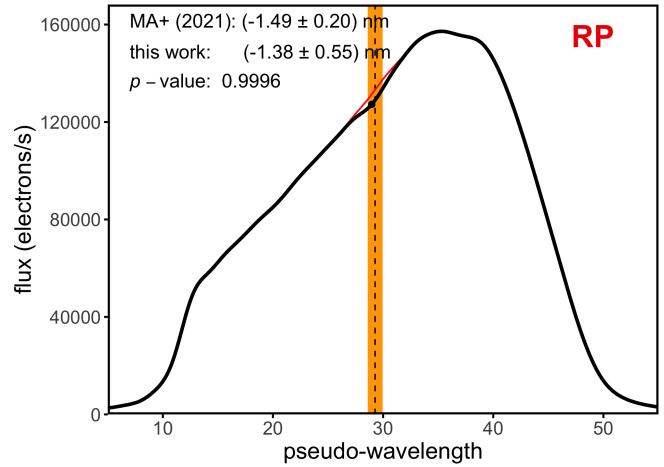


**Fig. 12.** Analysis of hydrogen Balmer and He I lines for Gaia DR3 426306363477869696. As in Fig. 9, the wavelengths increase from right to left for BP (upper panel) and from left to right for RP (bottom panel).

discovered this band in reddened OB stars and provides a number of stars with equivalent widths for this band. The band shape is Gaussian, with a central wavelength of 769.9 nm and a FWHM of 17.66 nm. We use this information on the line shape to apply the approach discussed in Sect. 9.4. The internally calibrated RP spectrum for one example, Gaia DR3 2067888218857234304, with  $G \approx 9.246^m$  and  $G_{\text{BP}} - G_{\text{RP}} \approx 2.008^m$ , is shown in Fig. 13. Maíz Apellániz et al. (2021) reported an equivalent width of the 770nm band of  $(-1.49 \pm 0.20)$  nm for this star. An extremum in the second derivative, consistent with the pseudo-wavelength position of the 770 nm band, is detected, with a high  $p$ -value of 0.9996. The continuum model is shown as a red line in Fig. 13, derived for a Gaussian band profile. The resulting equivalent width of the band of  $W = (-1.38 \pm 0.55)$ nm is well in agreement with the result by Maíz Apellániz et al. (2021). The line width parameter  $\omega$  however is only  $1.15 \pm 0.16$ , and the band is thus still in agreement with being narrow. It is nevertheless consistent with the model for the pseudo-LSF used, which has a FWHM of already 22.7 nm at a wavelength of 770 nm. The expected value of  $\omega$  is 1.21, slightly larger than unity only, and also in good agreement with the measured value of  $\omega$ . The line parameters for the 770 nm band are listed in Table 2 as well.

## 12. Summary and discussion

We derived a methodology for the analysis of spectral emission and absorption lines in *Gaia* DR3 low resolution spectra. These spectra are represented by linear combinations of Hermite functions in their form of internally calibrated spectra and we exploited the useful mathematical properties that come with this form of representation. These properties include that the



**Fig. 13.** Example for the analysis of the broad interstellar band at 770 nm for source Gaia DR3 2067888218857234304. The nominal band position is indicated by dashed lines, the shaded regions indicate the range in pseudo-wavelength within which a local extremum is considered to be in agreement with the status of the band. The black curve shows the XP spectra and the red curve shows the continuum model. The black dot indicates the positions of local extrema in the second derivative.

derivative of a linear combination of Hermite functions is another linear combination of Hermite functions, that can be derived by a matrix-vector multiplication. Furthermore, all roots of a linear combination of Hermite functions can be computed as eigenvalues of a simple non-standard companion matrix. These combined properties allow for a convenient and efficient determination of local extrema and inflection points in *Gaia* DR3 XP spectra. We used this to derive expression for the position, error on position, statistical significance, and equivalent widths of spectral lines, both for narrow and broad lines. The resulting computations are very fast and robust and therefore can be applied to search for spectral lines and extract their parameters even for large numbers of objects with modest computational resources. Neither machine learning nor human interaction is required.

We applied the derived methods to *Gaia* DR3 low resolution spectra with hydrogen Balmer lines, He I lines, and a broad interstellar band. The line positions, widths, and equivalent widths, as well as the statistical significance of the corresponding local extrema in the spectra were computed as a demonstration. The presented methodology based on the derivatives of Hermite functions might help to make use of the interesting and in its form unique new *Gaia* data product that are the low-resolution spectra. Despite their low spectral resolution, *Gaia* XP spectra provide opportunities for the analysis of absorption and emission lines from the near-ultraviolet to the near-infrared part of the electromagnetic spectrum. Exploiting the mathematical properties of the unique representation of the spectra as linear combinations of Hermite functions allows for a detailed analysis of the spectral lines.

*Acknowledgements.* This work was funded by the Spanish MICIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe” by the European Union through grant RTI2018-095076-B-C21, and the Institute of Cosmos Sciences University of Barcelona (ICCUB, Unidad de Excelencia ‘María de Maeztu’) through grant CEX2019-000918-M. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the

DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

## References

- Bohlin, R. C., Gordon, K. D., & Tremblay, P. E. 2014, *PASP*, **126**, 711  
 Bohlin, R. C., Hubeny, I., & Rauch, T. 2020, *AJ*, **160**, 21  
 Brychkov, Y. A., Marichev, O. I., & Prudnikov, A. P. 1989, *Tables of Indefinite Integrals* (Gordon & Breach Science Publishers)  
 Cantat-Gaudin, T., & Anders, F. 2020, *A&A*, **633**, A99  
 Carrasco, J. M., Weiler, M., Jordi, C., et al. 2021, *A&A*, **652**, A86  
 Carroll, B. W., & Ostlie, D. A. 2017, *An Introduction to Modern Astrophysics*, 2nd edn. (Cambridge University Press)  
 Creevey, O. L., Sordo, R., Pailler, F., et al. 2023, *A&A*, in press, <https://doi.org/10.1051/0004-6361/202243688>  
 Day, D., & Romero, L. 2005, *SIAM J. Numer. Anal.*, **43**, 1969  
 De Angeli, F., Weiler, M., Montegriffo, P., et al. 2023, *A&A*, in press, <https://doi.org/10.1051/0004-6361/202243680>  
 Fouesneau, M., Fremat, Y., Andrae, R., et al. 2023, *A&A*, in press, <https://doi.org/10.1051/0004-6361/202243919>  
 Gaia Collaboration (Prusti, T., et al.) 2016, *A&A*, **595**, A1  
 Gaia Collaboration (Vallenari, A., et al.) 2023, *A&A*, in press, <https://doi.org/10.1051/0004-6361/202243940>  
 Golub, G. H., & Van Loan, C. F. 2013, *Matrix Computations*, 4th edn. (John Hopkins University Press)  
 Goodman, J. W. 1996, *Introduction to Fourier Optics*, 2nd edn. (McGraw-Hill)  
 Houk, N., & Cowley, A. P. 1975, *University of Michigan Catalogue of Two-dimensional Spectral Types for the HD Stars. I. Declinations -90\_ to -53\_f0*  
 Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, *A&A*, **523**, A48  
 Kariya, T., & Kurata, H. 2004, *Generalized Least Squares* (John Wiley & Sons)  
 Kramida, A., Yu. Ralchenko, Reader, J., & NIST ASD Team. 2021, *NIST Atomic Spectra Database (ver. 5.9)*, [Online]. Available: <https://physics.nist.gov/asd> [2022, July 20], National Institute of Standards and Technology, Gaithersburg, MD  
 Magnasco, V. 2013, *Elementary Molecular Quantum Mechanics: Mathematical Methods and Applications*, 2nd edn. (Elsevier)  
 Maíz Apellániz, J., Barbá, R. H., Caballero, J. A., Bohlin, R. C., & Fariña, C. 2021, *MNRAS*, **501**, 2487  
 Marsh, T. R. 1990, *ApJ*, **357**, 621  
 Monguió, M., Greimel, R., Drew, J. E., et al. 2020, *A&A*, **638**, A18  
 Montegriffo, P., De Angeli, F., Andrae, R., et al. 2023, *A&A*, in press, <https://doi.org/10.1051/0004-6361/202243880>  
 Olivero, J. J., & Longbothum, R. L. 1977, *JQSRT*, **17**, 233  
 Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007a, *Numerical Recipes*, 3rd edn. (Cambridge University Press)  
 Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007b, *Numerical Recipes*, Webnote No. 17, Rev. 1  
 Riello, M., De Angeli, F., Evans, D. W., et al. 2021, *A&A*, **649**, A3  
 Stetter, H. J. 2004, *Numerical Polynomial Algebra* (SIAM)  
 Weiler, M., Carrasco, J. M., Fabricius, C., & Jordi, C. 2020, *A&A*, **637**, A85