The busy function: a new analytic function for describing the integrated 21-cm spectral profile of galaxies

T. Westmeier, R. Jurek, D. Obreschkow, B. S. Koribalski and L. Staveley-Smith

1ICRAR, M468, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
2Australia Telescope National Facility, CSIRO Astronomy and Space Science, PO Box 76, Epping, NSW 1710, Australia

ABSTRACT
Accurate parametrization of galaxies detected in the 21-cm H I emission is of fundamental importance to the measurement of commonly used indicators of galaxy evolution, including the Tully–Fisher relation and the H I mass function. Here, we propose a new analytic function, named the ‘busy function’, that can be used to accurately describe the characteristic double-horn H I profile of many galaxies. The busy function is a continuous, differentiable function that consists of only two basic functions, the error function, \( \text{erf}(x) \), and a polynomial, \( |x|^n \), of degree \( n \geq 2 \). We present the basic properties of the busy function and illustrate its great flexibility in fitting a wide range of H I profiles from the Gaussian profiles of dwarf galaxies to the broad, asymmetric double-horn profiles of spiral galaxies. Applications of the busy function include the accurate and efficient parametrization of observed H I spectra of galaxies and the construction of spectral templates for simulations and matched-filtering algorithms. We demonstrate the busy function’s power by automatically fitting it to the H I spectra of 1000 galaxies from the HI Parkes All-Sky Survey (HIPASS) Bright Galaxy Catalog, using our own c/c++ implementation, and comparing the resulting parameters with the catalogued ones. We also present two methods for determining the uncertainties of observational parameters derived from the fit.

Key words: line: profiles – methods: data analysis – radio lines: galaxies.

1 INTRODUCTION
Observations of the 21-cm emission line of neutral hydrogen provide measurements of several important parameters of galaxies, including their redshift, mass and rotational velocity, as well as evolutionary indicators such as the Tully–Fisher relation (Tully & Fisher 1977) and the H I mass function (Zwaan et al. 1997). While high-resolution H I maps have been obtained for several hundred nearby galaxies using radio interferometers, the vast majority of catalogued H I properties of galaxies have been extracted from integrated spectra obtained with single-dish telescopes. Over the next decade, H I surveys with the Square Kilometre Array (SKA; Dewdney et al. 2009) and some of its precursor and pathfinder instruments, such as Australian Square Kilometre Array Pathfinder (ASKAP; DeBoer et al. 2009) and Apertif (Oosterloo et al. 2009), will probe larger volumes of space to much greater depth than ever before. As a result, the number of H I-detected galaxies is expected to rise from currently \( \geq 30 000 \) (HYPERLEDA; Paturel et al. 2003) to more than half a million galaxies predicted for the planned all-sky surveys Widefield ASKAP L-band Legacy All-sky Blind surveY (WALLABY) and Westerbork Northern Sky HI Survey (WNSHS; Duffy et al. 2012). Yet, even in interferometric surveys like WALLABY, 95 per cent of all expected detections will be less than three beams across (assuming a beam size of 30 arcsec) and hence only marginally resolved, highlighting the importance of accurate parametrization methods based on the integrated H I spectrum.

Integrated H I line profiles encode much physical information. For example, (i) the frequency centroid of the H I line measures the cosmological redshift plus peculiar motion of the galaxy, (ii) the integral of the H I line provides a direct measure of the total H I mass (Roberts 1962), (iii) the linewidth traces the projected circular velocity of the galaxy and hence its dynamical mass (Caser-tano & Shostak 1980) and (iv) the small-scale structure of the line profile encodes information on turbulent motion and warps (Sancisi 1976). Furthermore, the shape of the H I line depends on disc asymmetries (Baldwin, Lynden-Bell & Sancisi 1980), extra-planar gas (Swaters, Sancisi & van der Hulst 1997; Heald et al. 2011), tidal tails and companions. Finally, observational settings, such as spectral resolution and noise level, also affect observed H I lines. The efficient extraction of all this information from thousands of noisy H I spectra requires a quick and accurate parametrization method.
Here, we present the ‘busy function’, $B_B$, a heuristic, analytic function suitable for the parametrization of galaxy H\textsc{i} spectra, as well as two modifications, $B_1$ and $B_2$. We explore their characteristics and potential applications. Furthermore, we apply the function to the 1000 H\textsc{i} spectra from the HI Parkes All-Sky Survey (HIPASS) Bright Galaxy Catalog (HIPASS BGC; Koribalski et al. 2004) and compare the results with the published H\textsc{i} properties of the galaxies. Possible applications of the busy function include the extraction of galaxy properties from spectral profile fitting, the construction of profile shapes for matched-filtering techniques in automated source finding and the possibility of analytically describing complex spectral profiles of galaxies using only a few basic parameters. This paper focuses on the application of spectral fitting for the purpose of accurate galaxy parametrization, an aspect driven by the need for automated parametrization of large samples of galaxies in future H\textsc{i} surveys with the SKA and its precursors.

This paper is structured as follows: in Section 2, we discuss previous approaches to fitting the spectral profiles of galaxies. In Section 3, we introduce the busy function and its basic properties. Two variations of the busy function are described in Section 4, while examples of fitting the function to the integrated spectra of real galaxies are presented in Section 5. In Section 6, we demonstrate the power of the busy function by automatically fitting it to the 1000 spectra of the HIPASS BGC and comparing our results with the catalogued parameters of the galaxies. Section 7 summarizes our results and conclusions. In the appendices, we present some of the more technical aspects of this paper. Appendix A discusses some of the analytic solutions of the busy function, while Appendix B explores the relationship between the busy function and a Gaussian. In Appendix C, we present the details of the c/c+ implementation of the busy-function-fitting algorithm developed for this paper. Finally, in Appendix D, we present and discuss methods for determining the uncertainties of parameters derived from busy function fits.

2 Previous Work

Previous attempts to describe the double-horn profiles of galaxy spectra usually involved Gaussian components. For the purpose of H\textsc{i} source finding in the extragalactic ALFALFA survey (Giovanelli et al. 2005) on the Arecibo telescope, Saintonge (2007) employed a matched-filtering technique using analytic template spectra. As a template she chose a linear combination of two Hermite functions of the form

$$S_1(x) = a_0 \Psi_0(x, \sigma) + c \Psi_2(x, \sigma),$$

(1)

where $\Psi_0(x, \sigma)$ and $\Psi_2(x, \sigma)$ are the first two symmetric Hermite functions, and $a_0$ and $c$ are free parameters. An example of this profile shape is shown in the left-hand panel of Fig. 1.

While the profile bears some resemblance to the general double-peaked profiles seen in the H\textsc{i} spectra of galaxies, it differs from the true shape of most galaxy spectra. First, the overall shape of the profile is close to Gaussian, and hence the flanks are not steep enough to reflect the sharp rise generally seen in the spectra of observed galaxies, in particular large spiral galaxies. Secondly, the central trough does not resemble the generally wide and flat troughs actually seen in many disc galaxies. In addition, the peaks of the profile are much smoother than the sharp peaks often seen in observed H\textsc{i} spectra. However, a great advantage of the approach by Saintonge (2007) is the small number of free parameters required to describe the profile, making it ideal for fulfilling its original purpose of serving as a template for matched filtering.

A more realistic profile was developed by Obreschkow et al. (2009a,b) for the description of H\textsc{i} discs generated by semi-analytic models. They used a combination of two different functions to model the outer flanks and the central trough of the profile separately, namely

$$S_2(x) = \begin{cases} k_1 \exp\left(-\frac{(|x|+\Delta x_p)^2}{k_2}\right) & \text{for } |x| \geq \Delta x_p/2, \\ \frac{k_1}{\sqrt{k_2}} & \text{for } |x| < \Delta x_p/2, \end{cases}$$

(2)

where $\Delta x_p$ is the separation between the two peaks. This heuristic profile shape was motivated by the desire to closely match the line shape derived from the modelling of an edge-on, rotating gas disc using simple analytic functions. An example of this profile is shown in the right-hand panel of Fig. 1.

While providing a fairly accurate description of the sharp and narrow peaks and broad troughs of typical galaxy spectra, there are a few disadvantages with this approach. First, the profile gets broken up into two separate functions that would have to be fitted separately. Secondly, precise adjustment of the parameters is required to avoid creating large discontinuities at the boundary between the two functions. Another complication arises from the function used.
to create the central trough of the profile, which is undefined for \( x^2 \geq k_2 \); this may pose a problem for fitting algorithms. In contrast to the profile used by Saintonge (2007), the profile of Obreschkow et al. (2009a,b) can reproduce the steep flanks and sharp peaks of observed galaxy spectra by decoupling the width of the Gaussian component from the overall profile width.

### 3 THE BUSY FUNCTION

#### 3.1 Definition

In order to improve on previous attempts to describe the double-horn profile of galaxies, we looked for a function that would allow us to describe the steep flanks often seen in the spectra of galaxies while also recovering the characteristic trough and sharp, narrow peaks of the spectrum. In addition, we require the function to be continuous and differentiable for the purpose of least-squares fitting, for which calculation of the partial derivatives with respect to the function’s free parameters, \( \frac{\partial f(x,p)}{\partial p_i} \), is required. Such a function can indeed be constructed in a relatively simple fashion by multiplying two error functions and a parabola,

\[
B_0(\xi) = \frac{a}{4} \times (\text{erf}[b(w + \xi)] + 1) \times (\text{erf}[b(w - \xi)] + 1) \times (c \xi^2 + 1),
\]

with \( \xi \equiv x - x_0 \). Here, the variable \( x \) represents the spectral axis of the data, e.g. frequency or radial velocity. The two error functions, \( \text{erf}(x) \), generate the profile flanks, while multiplication with a parabola produces the central trough of the profile. Given its characteristics and versatility, we chose to call this function the ‘busy function’. Examples of the busy function with different parameter values are shown in Fig. 2. In this section, we will first discuss the fundamental properties of the basic form of the busy function as specified in equation (3) before introducing a more general, asymmetric form of the function in Section 4.

#### 3.2 Free parameters

The busy function is characterized by five free parameters: the centroid of the profile, \( x_0 \); the half-width of the profile, \( w \); the total amplitude scaling factor, \( a \); and two additional parameters, \( b \) and \( c \). The parameter \( b \) controls the steepness of the two error functions constituting the flanks of the spectrum. In the case of \( b \to \infty \), the flanks will become infinitely steep, whereas for \( b \to 0 \), the slope of the flanks will approach zero. The parameter \( c \) controls the emphasis of the parabola and hence the amplitude of the central trough. Values of \( c > 0 \) imply increasing amplitudes of the trough, whereas for \( c = 0 \), the trough will disappear altogether. Negative values of \( a, b, c \) and \( w \), while mathematically allowed, are not physically meaningful in the case of \( H I \) emission lines, but may be useful in other situations not considered here, such as absorption lines.

#### 3.3 General properties

Some of the analytic solutions to the busy function are elaborated and presented in Appendix A. For profiles with \( bw \gg 1 \) (i.e. flat-topped or double-horn profiles), the value at the centre of the profile is simply given by \( B_0(0) = a \). The profile’s half-width, \( w \), is equal to half the separation of the two error functions, and hence the full width at half-maximum of the profile is equal to \( 2w \) in the case of \( c = 0 \) and \( bw \gg 1 \).

An advantage of the busy function, and the motivation for its name, is its versatility when it comes to fitting spectral profiles of different shape. Examples of the busy function mimicking double-peaked profiles of different shape are presented in the left-hand panel of Fig. 2. By carefully choosing appropriate values for \( b, c \) and \( w \), almost any shape of (symmetric) double-peaked profile can be reproduced by the function. By using error functions to represent the flanks of the spectrum, we can reproduce the characteristic, steep rise of the spectral profile typically observed in the integrated \( H I \) spectra of disc galaxies.

The flexibility of the busy function goes well beyond the fitting of double-peaked profiles, as is illustrated in the right-hand panel of Fig. 2. Here, the parameter \( c \) was set to zero to entirely remove the parabolic component and, in combination with different values of \( b \), produces profiles of varying shape ranging from a steep top-hat function (with large \( b \)) to a Gaussian function with gradual slopes on both sides (using smaller values of \( b \)). In fact, with the right choice of parameters, the product of two error functions will take almost the exact shape of a Gaussian function (see Appendix B). Hence, the busy function is capable of fitting both Gaussian and double-peaked line profiles.
4 MODIFICATIONS OF THE BUSY FUNCTION

4.1 Generalized busy function

In its original form of equation (3), the busy function is symmetric, and the shape of its central trough is determined by the parabolic component. However, at the expense of additional free parameters, the busy function can be generalized to describe a wider range of spectral profiles, e.g. profiles that are not symmetric or have a differently shaped trough. In a more general form, the busy function can be written as

\[ B_1(x) = \frac{a}{4} \times (\text{erf}[b_1(x + x_0)] + 1) \times (\text{erf}[b_2(x - x_0)] + 1) \times (c |x - x_p|^{n} + 1). \]  

(4)

The number of free parameters in this case has increased from 5 to 8, including separate slopes, \( b_1 \) and \( b_2 \), for the two error functions, separate offsets, \( x_0 \) and \( x_p \), for the error functions and the polynomial, and a variable degree, \( n \), of the polynomial. The properties of the generalized busy function are similar to those of the original busy function and discussed in Appendix A.

An example of an asymmetric, generalized busy function with a fourth-degree polynomial trough (\( n = 4 \)) is depicted in the left-hand panel of Fig. 3. The significantly broader trough more closely resembles those typically found in the spectra of many spiral galaxies.

4.2 Simplified busy function

A simplification of the busy function can be achieved by replacing the two error functions, \( \text{erf}(x) \), with just a single error function with the argument, \( \xi \), squared. While the shape and properties of the resulting function are very similar to the original busy function, the expressions for the simplified busy function, and in particular its derivatives, are less complex and significantly shorter:

\[ B_2(\xi) = \frac{a}{2} \times (\text{erf}[b (w^2 - \xi^2)] + 1) \times (e \xi^2 + 1). \]  

(5)

As before, we define \( \xi = x - x_0 \) for simplicity. This function, as depicted in the right-hand panel of Fig. 3, is very similar to the original busy function as specified in equation (3). Like \( B_0 \), the simplified busy function can be generalized by introducing an independent \( x_0 \) for the parabola or by replacing the parabolic trough with a different function, e.g. a fourth-degree polynomial. The slopes of the two flanks, however, will always be the same.

Some of the analytic solutions to the simplified busy function are discussed in Appendix A. Unlike \( B_0 \), there is no combination of parameters for which the profile would closely resemble a Gaussian function. In the best approximation, \( B_2 \) will appear more compact than a Gaussian, having steeper flanks and a slightly smaller amplitude.

5 EXAMPLES

In this section, we present a few examples of fitting the busy function to the H\(_{\alpha}\) spectra of observed galaxies to illustrate its usefulness and flexibility.

5.1 Symmetric profiles

In Fig. 4, we show integrated H\(_{\alpha}\) spectra of the two spiral galaxies NGC 300 (Westmeier, Braun & Koribalski 2011) and NGC 3351/M95 (Walter et al. 2008). All three versions of the busy function, \( B_0 \), \( B_1 \), and \( B_2 \), have been fitted to the spectra using a \( \chi^2 \) minimization algorithm. In the case of the generalized busy function, \( B_1 \), we assumed symmetry (\( b_1 = b_2, x_0 = x_p \)), but used a fourth-degree polynomial (\( n = 4 \)) to generate a broader trough.

NGC 300 is an example of a medium-sized spiral galaxy with a symmetric H\(_{\alpha}\) spectrum and a relatively sharp, almost V-shaped trough that is equally well described by either the original busy function, \( B_0 \), or the simplified busy function, \( B_2 \), both of which use a second-degree polynomial to describe the trough. In contrast to these, the much wider fourth-degree polynomial trough chosen for the generalized busy function, \( B_1 \), does not describe the appearance of NGC 300 very well. This is obvious from the residuals between the data and model, which are significantly larger (\( \sigma_{\text{rms}} = 0.46 \) Jy) for \( B_1 \) as compared to the other two fits (\( \sigma_{\text{rms}} \approx 0.32 \) Jy).

Note that the residuals in this and all following examples are due to real structures and asymmetries in the galaxies themselves and much larger than expected from the baseline noise of the integrated spectra alone.

NGC 3351 (M95) is an example of a galaxy with a broad trough. In this case, the generalized busy function, \( B_1 \), with its fourth-degree polynomial provides a much better fit (\( \sigma_{\text{rms}} = 8.8 \) mJy) to the integrated H\(_{\alpha}\) spectrum than the busy functions \( B_0 \) and \( B_2 \) (\( \sigma_{\text{rms}} \approx 21.2 \) mJy).
5.2 Asymmetric profiles

As discussed in Section 4, the busy function can be easily generalized to fit asymmetric profiles by simply introducing a separate offset for the polynomial component describing the trough. Two examples of asymmetric busy functions fitted to the integrated H\(_I\) spectra of the two galaxies NGC 55 (Westmeier, Koribalski & Braun 2013) and NGC 4826 (M64; Walter et al. 2008) are presented in Fig. 5. Here, we fitted two versions of the generalized busy function, \(B_1\), to the spectrum, this time including separate offsets, \(x_e\) and \(x_p\), for the error functions and the polynomial trough. This allows the trough to shift with respect to the flanks of the spectrum, producing an asymmetric profile with peaks of different height. The two functions use polynomial degrees of \(n = 2\) and 4, respectively, while adopting a single slope for the two flanks (\(b_1 = b_2\)).

The integrated spectrum of NGC 55 resembles that of NGC 300 with the exception of a noticeable asymmetry that is well fitted by the generalized busy function. As in the case of NGC 300, the parabolic trough (\(n = 2\)) fits better (\(\sigma_{\text{rms}} = 0.41\) Jy) than the fourth-degree polynomial (\(\sigma_{\text{rms}} = 0.62\) Jy). Within their uncertainties, the offsets significantly differ between the overall spectrum described by the error functions, \(x_e = 119.4 \pm 0.4\) km s\(^{-1}\), and the parabolic trough, \(x_p = 110.6 \pm 1.3\) km s\(^{-1}\), thereby quantitatively confirming the intrinsic asymmetry of the spectrum.

The spectrum of NGC 4826 is clearly asymmetric, too. As before, both versions of the generalized busy function provide a good fit to the spectrum, although the fourth-degree polynomial is capable of fitting the broad trough and sharp peaks in the spectrum much better (\(\sigma_{\text{rms}} = 14.0\) mJy) than the parabolic trough (\(\sigma_{\text{rms}} = 17.3\) mJy).
Again, there is a significant difference between the location of the overall spectrum, $x_0 = 409.3 \pm 0.7$ km s$^{-1}$, and that of the fourth-degree polynomial trough, $x_0 = 391.5 \pm 2.4$ km s$^{-1}$, confirming the intrinsic asymmetry of the spectrum.

6 PRACTICAL APPLICATION

In this section, we present the practical application of busy function fitting to the integrated spectra of the HIPASS BGC sources (Koribalski et al. 2004). We present both the methodology and the results. Most of the sources in the BGC are individual galaxies with a unique optical counterpart, although 68 sources are flagged as confused, 44 identified as pairs and 11 identified as compact groups. 91 detections do not have an optical identification, mostly as a result of Galactic foreground extinction.

6.1 Method

We developed our own software to fit the busy function to all 1000 galaxies in the BGC. The technical details of our implementation are presented in Appendix C. Our software attempts to fit six variants of the busy function with a varying number of free parameters, using the Levenberg–Marquardt algorithm (LMA; Levenberg 1944; Marquardt 1963). The software then selects the best of these fits based on the Akaike information criterion (AIC; Akaike 1974) in an attempt to determine the optimal number of free parameters needed to describe the data. A key feature of our method is the use of multiple (several thousand), short duration (tens of iterations) LMA attempts for each model, each attempt starting from a randomly chosen position in parameter space. This implementation is available in the form of C and C++ libraries and as a PYTHON module (see Appendix C for details). The advantages of our implementation are as follows.

(i) Model fits do not suffer from parameter discretization.
(ii) The method ensures rapid and efficient exploration of multidimensional parameter spaces.
(iii) A covariance matrix is produced for the model fit, thus providing parameter uncertainties as well as correlations.
(iv) The likelihood of finding the best global fit is higher.
(v) The code takes advantage of multicore CPUs, distributed systems, and GPUs.
(vi) The use of model variants allows us to set model components to exactly zero (true zeroing). This is otherwise impossible in practice because of the limited numerical precision of computers.
(vii) The algorithm takes into account the uncertainties of individual data points, resulting in data points with large uncertainties to be effectively excluded from fitting.
(viii) The software can easily be made to fit any analytic model or function.

Our implementation of the busy-function-fitting algorithm can be obtained from a dedicated website\(^1\) as a C library, a C++ template library and a PYTHON module.

6.2 Results

We successfully fitted the busy function to all 1000 HIPASS BGC sources. We accomplished this by using an iterative strategy. In each iteration, we fitted the busy function to all BGC sources that were not successfully fitted in the previous iteration, each time using different random LMA starting positions in parameter space. The details of this procedure are explained in Appendix C. Success or failure of a fit was assessed by checking whether the reduced $\chi^2$ of the fit was reasonable. In addition, the quality and accuracy of the fit was visually confirmed. The success rate was about 85 per cent for each iteration, and in the fourth and fifth iterations we were only processing two and a single BGC source, respectively. The success of this iterative approach confirms that the failures in any iteration are purely a result of poor LMA starting positions, because they are randomly chosen in each iteration. This also confirms our expectation that we could improve the success rate of each iteration, by increasing the number of LMA starting positions (at an increased computational cost). Alternatively, initial estimates of the free parameters, e.g. from a preceding source finding run, can be used instead of random starting positions in parameter space, thereby avoiding the need for multiple iterations altogether.

To test the performance of the busy function in the case of spectra with lower signal-to-noise ratio, we injected additional noise into the HIPASS BGC spectra to generate spectra with a peak signal-to-noise ratio of 3 and 5. We then refitted the busy function, using the same iterative procedure based on visual inspection. The success rate for both sets of noiser BGC spectra was approximately 90 per cent. After six iterations, we were able to fit a ‘sensible’ busy function (as qualitatively judged by us) to every noisy BGC spectrum.

In Table 1 and Fig. 6, we compare the catalogued observational properties of BGC sources against those derived from the busy function fits, allowing us to infer the quality of the fits. The catalogued parameters simply serve as a reference and are not necessarily unbiased or more accurate than our fitting results. We compare the integrated H\textsc{i} flux ($F_{\text{int}}$), peak H\textsc{i} flux density ($F_{\text{peak}}$) and the spectral linewidths at 20 and 50 per cent of the peak flux density ($w_{20}$ and $w_{50}$, respectively). The comparison was carried out using both a busy function sampled at the same spectral resolution as the data (SD) and a high-definition (HD) busy function sampled at 100 times the data’s spectral resolution. It turns out that there is no discernible benefit to using the HD busy function over the SD busy function, and hence only the SD results are listed in Table 1. We believe that this is a result of the catalogued properties being derived directly from the data. We also recalculated the observational properties directly from the data, following the same approach as described in section 3.3 of the BGC paper (Koribalski et al. 2004). These ‘direct’ properties are used as a sanity check and are calculated in the same manner as the catalogued properties, but using the channel range within which the fitted busy function is $\geq 1$ per cent of its peak value. Any differences between the direct and catalogued properties should solely be due to differences in the channel range used in the measurement.

Fig. 6 presents a comparison of our parametrization of the HIPASS BGC spectra (without additional noise) with the original measurement of each parameter in the BGC. We accurately recover the fluxes and linewidths of the galaxies from the fitted busy function, with differences of usually well under 10 per cent relative to the original, catalogued parameters. The results illustrate the great accuracy with which galaxy parameters can be derived from busy function fits. The few outliers seen in Fig. 6 are mostly due to artefacts in the data, including interference and variations in the spectral baseline. Such effects can have a strong impact on the accuracy of the parametrization and will potentially affect any parametrization method. Some of the outliers in Fig. 6 in particular could be due to the additional baseline subtraction carried out for the BGC, but not

---

\(^1\) http://code.google.com/p/busy-function-fitting/
Table 1. Comparison of the catalogued and calculated HIPASS BGC properties for direct extraction of parameters from the spectrum as well as parametrization based on busy function fitting. For each parameter and signal-to-noise ratio, the table lists two different components of the comparison: the result of a linear regression with the catalogued parameter as the independent variable (best-fitting slope and intercept; see Fig. 6) and the fraction of parameters within a certain percentage of the original, catalogued values.

| Original | S/N = 5 | S/N = 3 |
|----------|---------|---------|
|          | Direct  | BF fit  | Direct  | BF fit  | Direct  | BF fit  |
|          |   BF fit |         |   BF fit |         |   BF fit |         |
|          |          |         |          |         |          |         |
| Integrated flux ($F_{\text{int}}$) | | | | | | |
| Best-fitting slope | 1.02 | 1.03 | 0.96 | 1.05 | 0.92 | 1.07 |
| Best-fitting intercept (Jy km s$^{-1}$) | $-1.08$ | $-1.35$ | $2.01$ | $-1.02$ | $4.49$ | $0.14$ |
| Within 5 per cent of cat. (per cent) | 89.8 | 83.3 | 35.8 | 37.1 | 22.3 | 23.2 |
| Within 10 per cent of cat. (per cent) | 96.9 | 95.5 | 65.6 | 66.9 | 42.5 | 43.2 |
| Within 25 per cent of cat. (per cent) | 98.9 | 99.1 | 93.2 | 91.4 | 77.4 | 76.4 |
| Peak flux density ($F_{\text{peak}}$) | | | | | | |
| Best-fitting slope | 1.00 | 0.99 | 1.14 | 1.04 | 1.27 | 1.09 |
| Best-fitting intercept (Jy) | 0.00 | 0.00 | $-0.01$ | $-0.01$ | 0.01 | 0.00 |
| Within 5 per cent of cat. (per cent) | 99.0 | 68.6 | 21.9 | 28.3 | 5.4 | 18.1 |
| Within 10 per cent of cat. (per cent) | 99.2 | 88.0 | 43.3 | 54.5 | 14.7 | 32.8 |
| Within 25 per cent of cat. (per cent) | 99.7 | 99.1 | 86.3 | 91.9 | 47.1 | 72.2 |
| Linewidth ($w_{50}$) | | | | | | |
| Best-fitting slope | 0.98 | 0.99 | 0.94 | 0.95 | 0.86 | 0.87 |
| Best-fitting intercept (km s$^{-1}$) | 0.64 | 1.47 | 12.9 | 6.39 | 60.2 | 21.7 |
| Within 5 per cent of cat. (per cent) | 99.4 | 86.6 | 46.4 | 50.3 | 25.1 | 29.0 |
| Within 10 per cent of cat. (per cent) | 99.4 | 94.0 | 69.4 | 72.2 | 42.4 | 47.8 |
| Within 25 per cent of cat. (per cent) | 99.6 | 97.7 | 86.0 | 89.3 | 63.2 | 72.1 |
| Linewidth ($w_{20}$) | | | | | | |
| Best-fitting slope | 0.96 | 0.96 | 0.90 | 0.87 | 0.89 | 0.80 |
| Best-fitting intercept (km s$^{-1}$) | 8.70 | 5.60 | 62.2 | 39.4 | 94.9 | 83.7 |
| Within 5 per cent of cat. (per cent) | 94.0 | 86.0 | 34.6 | 47.7 | 20.7 | 26.5 |
| Within 10 per cent of cat. (per cent) | 94.6 | 93.2 | 53.0 | 68.1 | 36.8 | 45.4 |
| Within 25 per cent of cat. (per cent) | 96.7 | 97.9 | 72.2 | 82.0 | 56.7 | 65.0 |

for the busy function fitting, potentially resulting in discrepancies for individual galaxies affected by baseline artefacts.

In Table 1, the calculated and catalogued HIPASS BGC observational properties are compared in two ways. First, a linear regression is carried out to test if a 1:1 correlation exists. The result of the linear regression is shown as the dashed, red curve in Fig. 6 (note that for a non-vanishing intercept the linear fit appears curved in logarithmic space). Secondly, we measure the fraction of calculated values that lie within 5, 10 and 25 per cent of the catalogued values. In doing so, we implicitly ignore the uncertainties in the original BGC parameters, although these would contribute to the measured differences as well. The uncertainties published by Koribalski et al. (2004) are of the order of 10 per cent for $F_{\text{peak}}$, $w_{50}$ and $w_{20}$, and about 15 per cent for $F_{\text{int}}$.

Two main conclusions can be drawn from Table 1. First, as the signal-to-noise ratio decreases the busy function fits recover the catalogued properties more reliably than direct measurement of the properties does. This is most evident for the peak flux density and, to a lesser degree, for the velocity widths. The reason for the busy function to perform better is that, at low signal-to-noise ratio, the direct parametrization method may be strongly affected by noise, whereas the fit is much less affected by individual noise peaks. There is little advantage, however, when measuring the integrated flux. It should be noted that this is to be expected with Gaussian noise, provided that the chosen channel range contains all of the signal, because the emission is integrated over the entire source. Secondly, despite the differences in the catalogued and calculated properties, we find that the majority of properties derived from the busy function fits are within 5, 10 and 25 per cent of the catalogued properties at signal-to-noise ratios of $\geq 10$, 5 and 3, respectively.

Our analysis demonstrates that the catalogued HIPASS BGC properties can be reliably recovered in a fully automated approach for the majority of sources across a range of signal-to-noise ratios. Potential applications include the parametrization of a large sample of galaxies as well as the construction of realistic H$\alpha$ profiles for simulations such as the S$^2$–SAX simulation (Obreschkow et al. 2009b). Additionally, storing the fitted busy function parameters in addition to the full galaxy spectra will be of particular benefit to large future H$\alpha$ surveys such as WALLABY and Deep Investigation of Neutral Gas Origins (DINGO; Meyer 2009; Duffy et al. 2012), allowing them to include a simplified representation of every integrated H$\alpha$ spectrum in their source catalogue using a maximum of just eight parameter values.

Another great advantage of galaxy parametrization through the method of busy function fitting is the possibility to determine the statistical uncertainties of derived observational parameters from the uncertainties of the fit. This will enable a full error analysis based on the fitting results alone and without the need to modify the input data for that purpose. In Appendix D, we present the detailed
The busy function

Figure 6. Comparison of the integrated flux, peak flux density, and $w_{50}/w_{20}$ linewidths of the HIPASS BGC galaxies as derived from the busy function fit (ordinate) with the original values listed in the BGC (abscissa; parameters marked with tilde). The solid, grey line represents the identity, while the dashed, red line is the result of a linear regression carried out in linear space (see Table 1).

description and analysis of two methods to determine uncertainties of observational parameters from the covariance matrix provided by least-squares-fitting algorithms.

7 CONCLUSIONS

We present a continuous and differentiable analytic function, called the busy function, $B_0$, and two modifications, $B_1$ and $B_2$, designed to describe the typical double-horn profile commonly observed in the integrated H I spectra of spiral galaxies. With a set of five to eight free parameters, the busy function accurately describes a wide range of spectral profiles of galaxies, including symmetric and asymmetric double-horn profiles, simple Gaussian profiles and flat-topped profiles with steep flanks.

The most promising application of the busy function, and the main focus of this paper, is the possibility to automatically fit the integrated H I spectra of a large sample of galaxies. This will allow common observational parameters of galaxies, such as the linewidth or the integrated flux, to be measured with great accuracy and in an automated fashion. In addition, a simple functional representation of each galaxy’s integrated spectrum with a maximum of just eight parameters can be stored in a source catalogue. Another potential application of the busy function, although not investigated in this paper, is the generation of a sample of analytic mock profiles of galaxy spectra to be used as templates, either for the purpose of modelling or in matched-filtering algorithms of source-finding pipelines.

A great advantage of parametrizing galaxies by fitting a busy function to the integrated spectrum is the possibility to determine statistical uncertainties of the derived observational parameters. In Appendix D, we present two methods that allow the uncertainties of observational parameters to be determined without the need to
modify the input data. This will enable a proper error analysis even in situations where classical Monte Carlo or bootstrap methods cannot be applied, e.g. when a single spectrum of low signal-to-noise ratio needs to be parametrized.

In order to test the suitability of the busy function for automated spectral-line fitting, we implemented an LMA in c/c++ (Appendix C) to fit busy functions to the 1000 galaxies of the HIPASS BGC. Our results demonstrate that it is possible to fit the busy function to a large number of galaxy spectra in a fully automated way without any human intervention. A comparison of several measured galaxy parameters (integrated flux, peak flux density, and $w_{20}$ and $w_{50}$ linewidths) with those listed in the HIPASS BGC reveals that we accurately recover the observational parameters of the galaxies from the fit.

For the original spectra, almost all of our measured parameters lie within 25 per cent of the catalogueued ones. Even when reducing the peak signal-to-noise ratio of the spectra to 5 and 3, that fraction still remains at about 90 and 70 per cent, respectively. In addition, our measurement based on the busy function fit is often more accurate than the direct parameter measurement carried out on the spectrum. This result illustrates another great strength of the busy function: parametrisation based on fitting a busy function to the spectrum is less strongly affected by the noise in individual channels and thus produces more accurate results than any direct measurement. As a consequence, the number of galaxies in an observational sample that can be successfully parametrized would potentially increase compared to conventional parametrisation methods, thereby improving the accuracy of scientific studies such as the measurement of the Tully–Fisher relation.

While originally designed to describe the integrated H I emission spectra of galaxies, the busy function’s versatility will allow it to be used in many other areas, for instance in the fitting of H I absorption spectra (see Allison et al. 2013 for an actual example), stacked H I spectra of galaxies (e.g. Fabello et al. 2011; Delhaize et al. 2013) and integrated CO spectra of galaxies (e.g. Saintonge et al. 2011; Young et al. 2011; Tacconi et al. 2013).

ACKNOWLEDGEMENTS

We thank the members of the H I source-finding collaboration for stimulating discussions that led to the development of the busy function. The Australia Telescope is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO. This work made use of THINGS, ‘The H I Nearby Galaxy Survey’ (Walter et al. 2008).

REFERENCES

Akaike H., 1974, IEEE Trans. Autom. Control, 19, 716
Allison J. R., Curran S. J., Sadler E. M., Reeves S. N., 2013, MNRAS, 430, 157
Baldwin J. E., Lynden-Bell D., Sancisi R., 1980, MNRAS, 193, 313
Box G. E. P., Muller M. E., 1958, Ann. Math. Stat., 29, 610
Casertano S. P. R., Shostak G. S., 1980, A&A, 81, 371
Deboer D. R. et al., 2009, Proc. IEEE, 97, 1507
Delhaize J., Meyer M., Staveley-Smith L., Boyle B., 2013, MNRAS, 433, 1398
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, Proc. IEEE, 97, 1482
Donley J. L. et al., 2005, AJ, 129, 220
Duffy A. R., Meyer M. J., Staveley-Smith L., Bernyk M., Croton D. J., Koribalski B. S., Gerstmann D., Westerlund S., 2012, MNRAS, 426, 3385
Fabello S., Catinella B., Giovannelli R., Kauffmann G., Haynes M. P., Heckman T. M., Schiminovich D., 2011, MNRAS, 411, 993
Giovanelli R. et al., 2005, AJ, 130, 2598
Heald G. et al., 2011, A&A, 526, A118
Hong T. et al., 2013, MNRAS, 432, 1178
Koribalski B. S. et al., 2004, AJ, 128, 16
Levenberg K., 1944, Q. Appl. Math., 2, 164
Marquardt D., 1963, SIAM J. Appl. Math., 11, 431
Meyer M., 2009, in Heald G., Serra P., eds, Proc. Panoramic Radio Astronomy: Wide-field 1–2 GHz Research on Galaxy Evolution. PoS(PRA2009)015
Obreschkow D., Croton D., De Lucia G., Khochfar S., Rawlings S., 2009a, ApJ, 698, 1467
Obreschkow D., Klückner H.-R., Heywood I., Levrier F., Rawlings S., 2009b, AAp, 703, 1890
Oosterloo T., Verheijen M. A. W., van Cappellen W., Bakker L., Heald G., Ivalishina M., 2009, in Torchinsky S. A., van Ardeen A., van den Brink-Havinga T., van Es A. J. J., Faulkner A. J., eds, Proc. Wide Field Astronomy & Technology for the Square Kilometre Array, PoS(SKADS 2009)070
Paturel G., Petit C., Prugniel P., Theureau G., Rousseau J., Brouty M., Dubois P., Cambrésy L., 2003, A&A, 412, 45
Roberts M. S., 1962, AJ, 67, 437
Saintonge A., 2007, AJ, 133, 2087
Saintonge A. et al., 2011, MNRAS, 415, 32
Sancisi R., 1976, A&A, 53, 159
Swaters R. A., Sancisi R., van der Hulst J. M., 1997, ApJ, 491, 140
Tacconi L. J. et al., 2013, ApJ, 768, 74
Tully R. B., Fisher J. R., 1977, A&A, 54, 661
Walter F., Brinks E., de Blok W. J. G., Bigiel F., Kennicutt R. C., Jr, Thornley M. D., Leroy A., 2008, AJ, 136, 2563
Westmeier T., Braun R., Koribalski B. S., 2011, MNRAS, 410, 2217
Westmeier T., Koribalski B. S., Braun R., 2013, MNRAS, 434, 3511
Young L. M. et al., 2011, MNRAS, 414, 940
Zwaan M. A., Briggs F. H., Sprayberry D., Sorar E., 1997, MNRAS, 490, 173

APPENDIX A: ANALYTIC SOLUTIONS OF THE BUSY FUNCTION

In this section, we describe some of the analytic solutions for the evaluation of the three busy functions, $B_n$, with $n \in \{0, 1, 2\}$, as defined in equations (3)–(5). For simplicity, we will only consider symmetric functions here (i.e. $b_1 = b_2$ and $x_0 = x_0$ for $B_1$).

It is straightforward to calculate the derivatives of the busy function with respect to $x$ or any of the free parameters, and we will refrain from presenting the rather lengthy analytic expressions here. By setting $db_n/dx = 0$, we would in principle be able to calculate the positions of the extrema of $B_n$, but we have been unable to find a solution to this equation for any of the variants of the busy function, and there may not be an analytic solution other than the trivial solution of $x = x_0$ (i.e. $\xi = 0$). Hence, numerical methods will need to be used to determine the positions of the peaks of $B_n$ as well as the resulting galaxy parameters, such as peak flux density or profile width.

For all three versions of the busy function, we can calculate the value at the position of the trivial extremum, $\xi = 0$, which corresponds to the centre of the profile. The value at the centre of the original busy function, $B_0$, and the generalized busy function, $B_1$, is given by

$$B_0(0) = B_1(0) = \frac{a}{4} (\text{erf}[bw] + 1)^2,$$

which will converge to $a$ for $bw \gg 1$ (i.e. broad profiles with steep flanks). Note that this only applies to the symmetric version of $B_1$. 

1184 T. Westmeier et al. 


where \( b_1 = b_2 \) and \( x_c = x_0 \). The value at the centre of the simplified busy function, \( B_2 \), is given by
\[
B_2(0) = \frac{a}{2} \left( \text{erf} \left( bw^2 \right) + 1 \right),
\] (A2)
which will again converge to \( a \) for \( bw^2 \gg 1 \).

While we cannot produce general solutions for the profile width, there is a simple solution for cases where the polynomial component disappears, i.e. \( c = 0 \), and the flanks of the spectrum are sufficiently steep, i.e. large values of \( bw \) (for \( B_0 \) and \( B_1 \)) or \( bw^2 \) (for \( B_2 \)). In this case of "boxy" spectra, the separation between the two error functions of \( 2w \) is equal to the full width at half-maximum (or \( w_{50} \)) of the profile. In other cases of double-horn profiles with steep flanks, although not exact, this solution may still provide a first-order estimate of \( w_{50} \).

**APPENDIX B: RELATION BETWEEN THE BUSY FUNCTION AND THE GAUSSIAN FUNCTION**

The Gaussian function is commonly used for the purpose of fitting spectral lines, including the HI spectra of galaxies and gas clouds. In this appendix, we demonstrate that the original busy function, \( B_0 \), without the polynomial trough (i.e. \( c = 0 \)) possesses the flexibility to closely approximate a Gaussian function and hence can describe the generally simple line profiles of dwarf galaxies and face-on galaxies just as well. In order to achieve this, let us first define a normalized Gaussian function of the form
\[
G(\xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\xi^2}{2\sigma^2} \right),
\] (B1)
and a busy function similar to equation (3), but in a simplified form, assuming a width of \( w = 0 \) and no polynomial component (\( c = 0 \)), hence
\[
B(\xi) = a(\text{erf}[b\xi] + 1)(\text{erf}[-b\xi] + 1). \quad \text{(B2)}
\]
For simplicity, we define \( \xi \equiv x - x_0 \), as before. Both functions can be expanded into a Taylor series at the origin, \( \xi = 0 \):
\[
G(\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\xi^2}{2\sigma^2}}, \quad \text{and} \quad B(\xi) = a\xi^0 - \frac{4ab^2}{\pi} \xi^2 + \frac{8ab^4}{3\pi} \xi^4 + \cdots.
\]
(B3)
(B4)

 Apparently, the Taylor series of the busy function and the Gaussian function are very similar. Both only contain terms of even order, and a comparison of the respective coefficients in each order should thus allow us to derive approximate relations between the parameters of the busy function and the standard deviation, \( \sigma \), of the Gaussian function.

From a comparison of the zeroth-order and second-order terms of the two Taylor series in equations (B3) and (B4), we obtain the following solutions for the parameters \( a \) and \( b \) of the busy function as defined in equation (B2):
\[
a = \frac{1}{\sqrt{2\pi}\sigma}, \quad \text{and} \quad b = \frac{\pi}{\sqrt{8\sigma}} = \frac{\pi}{2} a.
\]
(B5)
(B6)

The two parameters above describe a family of busy functions that closely resemble a Gaussian function of width \( \sigma \), although they are approximations only.

**APPENDIX C: IMPLEMENTATION OF A BUSY FUNCTION FITTING ALGORITHM**

We implemented a busy-function-fitting program using C++ and the c libraries Cfitsio and CPlot. Our fitting program is based on a combination of the LMA (Levenberg 1944; Marquardt 1963) and the AIC (Akaike 1974). We use the LMA to carry out a \( \chi^2 \) minimization for six variants of the busy function as listed in Table C1. We use those variants to fix various parameters of the busy function. Note that we use a slight reformulation of the busy function, because we found it to be faster to fit with the LMA. We use the AIC to penalize each busy function variant’s \( \chi^2 \) value for the number of free parameters and then choose the model with the best resultant \( \chi^2 \).

The LMA was used to carry out the \( \chi^2 \) minimization for three reasons. First, the LMA produces a \( \chi^2 \) covariance matrix that can be used to calculate both the uncertainties and correlations of the model parameters. Secondly, the LMA does not rely upon a grid of model parameter values, and the model parameters therefore do not suffer discretization (such as in brute-force \( \chi^2 \) minimization). Finally, the LMA uses the \( \chi^2 \) covariance matrix to ensure that it simultaneously adjusts the model parameter values in a way that achieves the maximum decrement in \( \chi^2 \). As model dimensionality increases, this aspect of the LMA becomes increasingly powerful.

We used variable remapping, singular value decomposition and a modified power law to fit the busy function with the LMA. Variable remapping is necessary because the LMA cannot impose any range limits on model parameters. We used the variable remappings in Table C2 to limit the model parameters to sensible ranges and to avoid unphysical parameter values. This includes positive definite values for scaling/normalization parameters and a user-specified...
range for the error function and power-law origins. We used singular value decomposition to ensure that degenerate model parameter values and extreme noise do not cause the LMA to fail.

Using the LMA does not solve the problem common to all model-fitting methods based on $\chi^2$ minimization: there is no guarantee that the LMA will find the global $\chi^2$ minimum when starting from an arbitrary position in parameter space. We solve this problem by using the LMA to find the nearest $\chi^2$ minimum for 1000 randomly chosen starting positions. These LMA starting positions are created in two steps. First, $\alpha$ is fixed at $10^{-3}$ and every other parameter is randomly chosen (with replacement) from the values in Table C3. The corresponding curve is then used to calculate an $\alpha$ and adjusted $\phi$ that ensure that the model peak is equal to a randomly chosen (with replacement) multiple of the peak data value (also listed in Table C3). We then assume that the model-fitting values of the smallest of these resultant $\chi^2$ values is a good proxy for the model-fitting values of the global $\chi^2$ minimum. An additional benefit of this approach is that we do not have to use too many iterations for each LMA process. The number of LMA iterations and starting positions can be traded against each other. This is because as the number of LMA starting positions is increased, the probability of choosing an LMA starting position close to the global $\chi^2$ minimum increases. For this application, we are using 30 iterations for each LMA process compared to the O(1000) iterations typically used.

To avoid local $\chi^2$ minima, we have also added an additional criterion to the typical definition of a $\chi^2$ minimum. A $\chi^2$ minimum is typically defined as a negligible decrease in $\chi^2$ for consecutive iterations. In addition to this, we require that the $\chi^2$ value must not have increased for five consecutive iterations. This avoids $\chi^2$ minima that are ‘noise troughs’ in unstable regions of parameter space.

It should be noted that there is also a powerful, inherent advantage to model fitting via $\chi^2$ minimization. Model fitting with $\chi^2$ minimization takes into account the uncertainty of each individual data point. Our implementation includes this capability, although it was not required for fitting the HIPASS BGC spectra (which we assume to have a constant noise level of 13 mJy). We expect that this feature will be useful for data sets with channel/frequency-dependent noise. Alternatively, this capability can also be exploited by assigning large uncertainties to channels/frequencies affected by radio frequency interference. This will down-weight the significance of such channels/frequencies when parametrizing affected galaxies.

Table C1. The six busy function variants used in our C/C++/PYTHON implementation, their dimensionality/complexity and the associated $\chi^2$ penalty imposed by the AIC.

| No. of free parameters | AIC $\chi^2$ penalty | Busy function variant |
|------------------------|----------------------|-----------------------|
| 4                      | +8                   | $(\alpha/4) \times (1 + \text{erf}[\beta \{x - \gamma\}]) \times (1 + \text{erf}[\beta \{y2 - x\}])$ |
| 5                      | +10                  | $(\alpha/4) \times (1 + \text{erf}[\beta_1 \{x - \gamma_1\}] \times (1 + \text{erf}[\beta_2 \{y2 - x\}])$ |
| 5                      | +10                  | $(\alpha/4) \times (1 + \text{erf}[\beta_1 \{x - \gamma_1\}] \times (1 + \text{erf}[\beta \{y2 - x\}]) \times (1 + \phi \{x - x\})$ |
| 6                      | +12                  | $(\alpha/4) \times (1 + \text{erf}[\beta_1 \{x - \gamma_1\}] \times (1 + \text{erf}[\beta \{y2 - x\}]) \times (1 + \phi \{x - x\})^2$ |
| 7                      | +14                  | $(\alpha/4) \times (1 + \text{erf}[\beta_1 \{x - \gamma_1\}] \times (1 + \text{erf}[\beta \{y2 - x\}]) \times (1 + \phi \{x - x\})^3$ |
| 8                      | +16                  | $(\alpha/4) \times (1 + \text{erf}[\beta_1 \{x - \gamma_1\}] \times (1 + \text{erf}[\beta_2 \{y2 - x\}]) \times (1 + \phi \{x - x\})^4$ |

Table C2. The variable mappings used to impose parameter range constraints during fitting. For a given user-specified range, the values mid and amp are the mid-point and distance from either end of the range to the mid-point, respectively.

| Variable(s) Mapping | Range | Reason |
|---------------------|-------|--------|
| $\alpha$ exp($\alpha$) | >0 | Meaningful normalization; avoids normalization trade offs. |
| $\beta_1$, $\beta_2$ exp($\beta$) | >0 | Prevents inversion of error function origins ($\gamma_1$, $\gamma_2$). |
| $\gamma_1$, $\gamma_2$ mid + amp sin($\gamma$) | User specified | Avoids irrelevant solutions. |
| $\phi$ exp($\phi$) | >0 | Meaningful normalization; avoids normalization trade offs. |
| $\theta$ mid + amp sin($\theta$) | User specified | Avoids irrelevant solutions. |
| $n$ 5 + 3 sin($n$) | $2 \leq n \leq 8$ | Avoids unphysical solutions. |

Table C3. The parameter list used to generate LMA starting positions, using random selection with replacement.

| Property | LMA seed values |
|----------|-----------------|
| Model peak | 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8, 1 × data maximum |
| $\beta_1^{(1)}$, $\beta_2^{(1)}$, $\beta_2^{(2)}$ | 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 10, g/1.25, g/2.5, g/5, g/10, g/15, g/20 |
| $\gamma_1^{(2)}$ | min, $+r/8$, $+r/4$, $+3r/8$, $+r/2$, $+5r/8$, $+3r/4$, $+7r/8$, max |
| $\gamma_2^{(2)}$ | min, $+r/8$, $+r/4$, $+3r/8$, $+r/2$, $+5r/8$, $+3r/4$, $+7r/8$, max provided $\gamma_2 \geq \gamma_1$ |
| $\phi$ for $\alpha = 10^{-3}$ | $10^{-9}$, $10^{-8}$, $10^{-7}$, $10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$, 0.01, 0.1, 1, 10, 100, 1000, 10000 |
| $\theta$ | $\gamma_1$, $(\gamma_2 - \gamma_1)/6$, $(\gamma_2 - \gamma_1)/3$, $(\gamma_2 - \gamma_1)/2$, $(\gamma_2 - \gamma_1)/3$, $(\gamma_2 - \gamma_1)$, $\gamma_2$ |
| $n$ | 2, 4, 6 |

Notes. (1) The constant $g = 0.747806$ is used to calculate $\beta$ values corresponding to Gaussian roll-offs of the form, $\beta = g/\sigma$. (2) min and max are the user-specified roll-off range; $+r$ denotes min plus the additional quantity, with $r \equiv max - min$. (3) $\gamma_1$ must be chosen first so that it can be used to limit the number of possible $\gamma_2$ values.
Our implementation of the busy-function-fitting algorithm can be obtained from the project’s website\(^2\) or by sending an e-mail to one of the authors, RJ (russell.jurek@gmail.com). The implementation is available as a C library, a C++ template library and a PYTHON module, all of which use OPENMP to take advantage of systems with multicore CPUs.

APPENDIX D: DETERMINATION OF UNCERTAINTIES

A crucial aspect in the parametrization of spectral profiles of galaxies is the determination of uncertainties. Several different methods have been used in the parameterization of \(\text{H} \text{I}\) lines in the past, including Monte Carlo methods (e.g. Donley et al. 2005) and bootstrap or jackknife methods (e.g. Hong et al. 2013). These methods generally attempt to emulate repeated measurements by either adding artificial noise to the data or by looking at different subsets of a data set. While providing the most accurate assessment of uncertainties (apart from actually repeating the measurement), Monte Carlo and bootstrap methods can be relatively slow due to the need to repeatedly alter the original data. In addition, not all data sets lend themselves to this type of procedure.

In this section, we introduce two different methods that rely entirely on the covariance matrix of the busy function fit and can therefore be applied in situations where Monte Carlo and bootstrap methods will fail, e.g. when only a single spectrum of low signal-to-noise ratio is available. In addition, the introduced methods are much faster than the former because they do not require the original data to be altered in any way.

D1 The classical approach: Monte Carlo and bootstrap methods

As Monte Carlo and bootstrap methods operate on the input data themselves, they are independent of the actual parametrization method used and can therefore be applied in combination with busy function fitting as well. As a demonstration of a Monte Carlo approach, we created 10 000 realizations of the integrated spectrum of the galaxy NGC 3351 (Walter et al. 2008), as shown in Fig. 4, by adding random Gaussian noise with a standard deviation of 50 mJy to the original spectrum. Next, we fitted the generalized busy function, \(B_1\), with the parameter \(n\) fixed to 4, to each of the 10 000 spectra, and extracted and analysed the resulting parameters. The results are summarized in Fig. D1 and Table D1. The resulting \(\chi^2\) of the fits follows the expected probability density function with a mean of \(\langle \chi^2 \rangle = 54.6\) and \(\langle \chi^2_{\text{red}} \rangle = 1.03\). The mean \(\chi^2\) is slightly larger than the number of degrees of freedom of 53 (60 spectral channels less 7 free parameters), but this small discrepancy can be readily explained by intrinsic structure and noise in the original spectrum of NGC 3351.

As shown in Fig. D1, most parameters obey an approximately normal distribution, although some are significantly skewed, in particular the two profile slope parameters, \(b_1\) and \(b_2\). Such non-Gaussian distributions imply that, strictly speaking, the parameter uncertainties cannot be expressed in terms of a single number, such as the standard deviation usually reported by least-squares-fitting algorithms. However, as a first-order approximation, the parameters of the busy function can be assumed to follow a normal distribution. The same applies to be true for numerically derived observational parameters (line centroid, \(w_{50}\) and \(w_{20}\) linewidths, peak flux density and integrated flux) as presented in the lower sections of Fig. D1 and Table D1.

D2 A different approach: variation of busy function parameters

While Monte Carlo and bootstrap methods provide a robust way of determining parameter uncertainties, it might not be possible to apply them in certain situations, e.g. when only a single spectrum with low signal-to-noise ratio is available. In such situations, one of the great advantages of the busy function over direct \(\text{H} \text{I}\) parametrization methods comes into play: under the assumption of a Gaussian statistic of the busy function parameters, we can determine the uncertainties of derived parameters, including linewidths and fluxes, from the covariance matrix provided by least-squares-fitting algorithms such as the LMA (Levenberg 1944; Marquardt 1963).

As concluded in Appendix A, the lack of analytical solutions for derived parameters implies that we need to determine the uncertainties of derived parameters numerically. The simplest approach would be to randomly vary the busy function parameters coming out of a fit and recalculate derived parameters such as \(w_{50}\) or \(F_{\text{int}}\). This can be repeated many times, and the uncertainties of the derived parameters can then be determined by either simply taking the standard deviation across all iterations or by fitting a Gaussian to the resulting parameter distribution. This approach implicitly assumes that all busy function parameters are normally distributed (see Appendix D1) and that the individual parameters are entirely uncorrelated.

Most parameters, however, have some degree of correlation (as shown in Fig. D2 for the integrated spectrum of NGC 3351), and it is necessary to take this effect into consideration when randomly varying the fit parameters. This can be achieved with the following method based on the parameter covariance matrix provided by the LMA.

(i) Using the LMA, the busy function is fitted to the integrated spectrum of the galaxy to be parametrized, providing us with values for the free parameters of the busy function, \(p\), as well as the full covariance matrix, \(C_p\).

(ii) Using the method of Box & Muller (1958), we can then generate \(M\) sets of independent, random busy function parameters, \(p_m\), each following a normal distribution centred on \(p\) with a standard deviation as derived from the square root of the diagonal elements of \(C_p\).

(iii) Next, we need to transform the vectors of random, uncorrelated parameters, \(p_m\), into vectors of random parameters with the correct correlations, \(p_m'\), as described by the non-diagonal elements of \(C_p\). This can be achieved by performing a Cholesky decomposition of the parameter correlation matrix,

\[
K_p = LL^T, (D1)
\]

where \(K_p^{ij} = C_p^{ij}/(\sigma_i\sigma_j)\) and \(\sigma_i^2 = C_p^{ii}\), and then multiplying each parameter vector with the lower triangular form of the resulting matrix,

\[
p_m' = Lp_m. \quad (D2)
\]

In this step, we need to take into account that each parameter in \(p_m\) must be of zero mean and unit variance for the transform to work. Hence, each parameter may need to be scaled and translated before and after the transform.

\(^2\) http://code.google.com/p/busy-function-fitting/
Figure D1. Distribution of busy function parameters and derived observational parameters after fitting the generalized busy function (with $n = 4$ fixed) to 10 000 realizations of the integrated spectrum of NGC 3351 with artificial noise added. The red curve shows the result of a Gaussian fit to the parameter distribution, with the exception of the $\chi^2_{red}$ distribution where we fitted the $\chi^2$ probability density function instead of a Gaussian. The unit of flux density is Jy, while spectral parameters are specified in channels.
Table D1. Busy function parameters (upper section) and derived observational parameters (lower section) after fitting the generalized busy function (with \( n = 4 \) fixed) to 10 000 realizations of the integrated spectrum of NGC 3351 with artificial noise added. The first two columns list the parameter name and its value derived from fitting the original, high signal-to-noise spectrum. The uncertainties in the upper section were taken from the covariance matrix of the fit, while those in the lower section have been calculated with the new method introduced in Appendix D2. Columns 3 and 4 show the mean, \( \bar{P} \), and median, \( \tilde{P} \), of all 10 000 realizations, columns 5 and 6 list the centroid, \( P_0 \), and standard deviation, \( \sigma_P \), of a Gaussian function fitted to the parameter distribution (see Fig. D1) and column 7 shows the skewness, \( \gamma_1 \), of the distribution. The unit of flux density is Jy, while spectral parameters are specified in channels.

| Parameter | Original value | \( \bar{P} \) | \( \tilde{P} \) | \( P_0 \) | \( \sigma_P \) | \( \gamma_1 \) |
|-----------|----------------|----------------|----------------|-----------|-----------|-----------|
| \( a \)   | 0.121 ± 0.002  | 0.121          | 0.120          | 0.009     | 0.00      |
| \( b_1 \) | 0.604 ± 0.036  | 0.705          | 0.585          | 0.164     | 7.63      |
| \( b_2 \) | 0.572 ± 0.034  | 0.651          | 0.585          | 0.152     | 9.32      |
| \( c \times 10^6 \) | 6.660 ± 0.235 | 6.860          | 6.760          | 6.711     | 1.151     | 0.63      |
| \( x_e \) | 30.183 ± 0.040 | 30.178         | 30.176         | 30.194    | −0.14     |
| \( x_p \) | 30.206 ± 0.098 | 30.198         | 30.200         | 30.200    | 0.470     | −0.02     |
| \( w \)   | 25.387 ± 0.045 | 25.365         | 25.369         | 25.375    | 0.215     | −0.39     |
| Centroid  | 30.164 ± 0.122 | 30.164         | 30.168         | 30.161    | 0.614     | 0.00      |
| \( w_{50} \) | 51.612 ± 0.085 | 51.515         | 51.510         | 51.501    | 0.462     | 0.17      |
| \( w_{20} \) | 53.485 ± 0.130 | 53.432         | 53.399         | 53.377    | 0.736     | 0.32      |
| \( F_{\text{peak}} \) | 0.348 ± 0.005  | 0.364          | 0.363          | 0.363     | 0.023     | 0.25      |
| \( F_{\text{int}} \) | 9.618 ± 0.075  | 9.633          | 9.634          | 9.632     | 0.382     | 0.00      |

(iv) Lastly, we can numerically derive the desired observational parameters (including centroid, \( w_{50} \), \( w_{20} \), \( F_{\text{peak}} \) and \( F_{\text{int}} \)) for each of the \( M \) correlated busy function parameter sets, \( p_i^n \). We can then calculate the mean and standard deviation for each parameter across all \( M \) iterations.

We tested this parameter variation method on the integrated spectrum of NGC 3351 (Fig. 4), fitting the generalized busy function, \( B_i \), with the parameter \( n = 4 \) fixed. The results, listed in the lower section of Table D1, are in good agreement with our expectations, suggesting that the method produces accurate uncertainties. For example, when numerically measuring the flux density at the position of the central trough of the fitted busy function, we derive a value of 0.1210 ± 0.0018 839. Both the value and the uncertainty are identical (within the numerical accuracy) with those of the busy function parameter \( a \) as derived from the least-squares fit. A similar comparison can be made between the measured linewidth, \( w_{50}/2 = 25.81 ± 0.042 \) 26, and the parameter \( w = 25.39 ± 0.045 \) 01 of the busy function. Again, the values and uncertainties agree very well, even though we do not expect an exact identity of \( w_{50}/2 \) and \( w \) (see Appendix A). The results suggest that the error analysis method based on varying the initial busy function parameters produces accurate measurements of the uncertainty of derived parameters.

D3 A faster approach: linear propagation of the covariance matrix

While the parameter variation method presented in Appendix D2 provides an accurate way of estimating uncertainties, it is relatively slow and inefficient due to the large number of iterations required to achieve sufficient numerical accuracy. However, under the assumption of a linear approximation of the function that translates between the free parameters of the busy function and the derived observational parameters of the spectral profile, we can instead use the error propagation law to determine not just the uncertainties of the derived parameters, but in fact the full covariance matrix.

Let us assume that \( p \) is the parameter vector and \( \mathbf{C}_p \) the covariance matrix of the busy function’s free parameters, as before. Let us further assume that there is a differentiable function, \( f \), that translates between the busy function parameters, \( p \), and the derived observational parameters, \( \mathbf{q} = f(p) \). We can then numerically approximate the Jacobian matrix of \( f \) by varying each input parameter, \( p_i \), by a small amount, \( \epsilon_i \), such that

\[
J_{ji} \equiv \frac{\partial f_j}{\partial p_i} \approx \frac{f_j(p + \epsilon_i \hat{e}_i) - f_j(p)}{\epsilon_i}, \tag{D3}
\]

where \( \hat{e}_i \) is the unit vector in the direction of the \( i \)-th component of \( p \). With the Jacobian matrix determined, we can now use the error

![Figure D2. Correlation coefficients of generalized busy function parameters as derived from the covariance matrix of a fit to the integrated spectrum of NGC 3351.](https://academic.oup.com/mnras/article-abstract/438/2/1176/1007156)
propagation law to calculate the full covariance matrix, $C_q$, of the derived observational parameters:

$$ C_q = J C_p J^T. \quad (D4) $$

The uncertainties of the individual parameters, $q_i$, are then given by the square root of the diagonal elements of the covariance matrix. Note that, for this method to work, it is not necessary to know the analytic expression of $f(p)$, as the function can be evaluated numerically.

As before, we tested the method of propagating the covariance matrix on the integrated spectrum of NGC 3351 (Fig. 4), using relative offsets of $\epsilon_j = |p_j| \times 10^{-5}$. The results, including a comparison with the method introduced in Appendix D2, are presented in Table D2. Both the method of parameter variation as well as the method of linear propagation of the covariance matrix yield comparable estimates of the uncertainties of the derived observational parameters of NGC 3351. The largest discrepancy is observed for the peak flux density, $F_{\text{peak}}$, for which the error propagation method yields an uncertainty that is by about 10 per cent higher than that of the parameter variation method. These small discrepancies are likely due to the linear approximation made in the error propagation method.

The error propagation method is generally much faster than the parameter variation method described in Appendix D2, as it does not require a large number of numerical iterations. Another advantage of the error propagation method is that it will produce a full parameter covariance matrix ‘for free’ (i.e. without the need for computationally expensive numerical iterations), thus providing information about correlations between observational parameters. This is illustrated in Fig. D3, where the correlation coefficients derived from the covariance matrix, $C_q$, of the fit to the spectrum of NGC 3351 are presented.

### D4 Conclusions

The possibility of deriving uncertainties from the covariance matrix of a busy function fit, either by randomly varying the initial fit parameters or by linearly propagating the covariance matrix, turns the busy function into a powerful tool for measuring the observational parameters of galaxies from their integrated spectra. The great advantage of the presented methods is that, unlike classical Monte Carlo or bootstrap methods, they are computationally inexpensive and do not require alterations to the data being fitted. Hence, both methods can be applied in situations where the input data are not suitable for classical error analysis methods, e.g. when only a single spectrum of low signal-to-noise ratio is available.

There are several important limitations to both methods, including the implicit assumption of Gaussian errors for all parameters as well as the requirement to have a sufficiently large number of samples across the width of the profile, as otherwise the fit will be underdetermined, resulting in unrealistic uncertainty estimates. These limitations, however, are not specific to the busy function but generally apply whenever analytic functions are fitted to data. In the case of the error propagation method, we need to make the additional assumption of a linear propagation of uncertainties. This assumption will break down in certain situations where highly non-linear relations between parameters exist, potentially leading to unrealistic uncertainty estimates, in particular when the uncertainties of the fitted busy function parameters are large compared to their values. Our tests of the two methods on the integrated spectrum of NGC 3351 demonstrate that both yield accurate uncertainty estimates when applied to well-resolved spectral profiles. However, it is important to keep the limitations discussed above in mind when using either method in the determination of parameter uncertainties.

This paper has been typeset from a \texttt{T\LaTeX/XeLaTeX} file prepared by the author.