

Gaussian decomposition of the Leiden/Dwingeloo survey

I. Decomposition algorithm

U. Haud

Tartu Observatory, 61602 Tõravere, Tartumaa, Estonia (urmas@aai.ee)

Received 3 July 2000 / Accepted 7 September 2000

Abstract. A new, fully automatic computer program for decomposition into Gaussian components of large 21-cm H I line surveys is described. To solve, at least partly, some problems inherent to such decompositions, special attention is paid on the following features:

1. Several quite different solutions may often fit the observations almost equally well. To choose from these solutions, it is supposed that general properties of the hydrogen distribution are somewhat correlated at neighboring sky positions and therefore the program tries to find for corresponding profiles also similar decompositions.
2. With increasing complexity of the observed profiles the number of Gaussians in decompositions usually grows rapidly and the values of their parameters become mutually dependent. To reduce this problem, special means have been used to keep the number of Gaussians as small as possible.

The recent Leiden/Dwingeloo Survey of the galactic neutral hydrogen has been used as test data for this decomposition program. These tests have demonstrated that the program is able to decompose fully automatically all the profiles in this survey in a reasonable time, spending on an average about 1.03 s per profile on 233 MHz PII PC. The comparison of the obtained results with those by other authors shows that, in general, this program needs less Gaussians per profile than used in comparison sets. The analysis of the stability of the decomposition results seems to indicate that the major source of the uncertainties is the observational noise, but also the precise behavior of the program has certain influence on final results.

Key words: methods: data analysis – surveys – ISM: atoms, ions – radio lines: ISM

1. Introduction

Hydrogen emission in the radio regime is caused by a hyperfine transition in the ground state of the atom with a natural frequency of 1 420.405 752 MHz (Kerr 1968). The natural width of the spin-flip transition is 10^{-16} km s⁻¹, which is negligibly small in any astrophysical context. Although, the radio-astronomical techniques would rather straightforwardly allow a resolution of

a line a fraction of a km s⁻¹ broad, H I emission lines less than a few km s⁻¹ wide are rarely found. The broadening of the 21-cm line occurs through several mechanisms. Broadening due to thermal velocities of atoms within a single concentration of gas at a certain temperature will produce a line of a *Gaussian shape*:

$$T_b = T_{b,0} e^{-\frac{(\nu-\nu_0)^2}{2\sigma_\nu^2}}, \quad (1)$$

with dispersion $\sigma_\nu = 0.09\sqrt{T_k}$ km s⁻¹ ($T_{b,0}$ and ν_0 are the central brightness temperature and central frequency of the line, respectively; T_b is the brightness temperature at frequency ν). For a realistic kinetic temperature of $T_k = 100$ K, $\sigma_\nu = 0.9$ km s⁻¹ (Burton 1992), which corresponds to a full width between half-intensity points of 2.1 km s⁻¹. Usually turbulent motions within a concentration of H I gas may produce additional profile broadening of the order of 5 km s⁻¹ without altering its general Gaussian shape.

What was actually obtained, starting from the first observations of the 21-cm H I line (first detected in 1950 by Ewen & Purcell 1951), were the line-profiles giving intensity as a function of frequency. These H I emission profiles were never simple, single Gaussians; virtually all of them showed multiple components, line asymmetries, or broad wings. At the same time, the assumption of a gas of uniform parameters was certainly not valid under the variety of conditions found in the interstellar environment. As a result, the interpretation of the line profiles usually refers to the collective properties of gas fragments occurring in the volume along the line of sight, sampled by the telescope beam. Moreover, it was possible to suppose that most of this total broadening comes from the global rotation characteristics of the Galaxy. This was of particular importance because, if the galactic H I consists of separate hydrogen clouds, the observed profile can be considered as a sum of corresponding Gaussian cloud components, shifted with respect to each other by the differential rotation.

These considerations probably can explain, why in earlier years of radio-astronomical hydrogen surveys it was rather popular to publish the results in the form of tables of Gaussian components in observed profiles (Heeschen 1954; Matthews 1956; Kaper et al. 1966; Lindblad 1966; Shane 1971 and others). While the first papers employed direct estimation of the Gaussian parameters (Heeschen 1954; Matthews 1956), soon

a computer program for the Gaussian analysis of 21-cm line profiles was developed in Groningen (Kaper 1959; van Woerden 1962; Schwarz & van Woerden 1962; Kaper et al. 1966; Schwarz 1968).

Unfortunately, shortly it became clear that the actual situation is much more complicated. The shape of the emission spectra at low galactic latitudes in the inner Galaxy does not change greatly with angular resolution of observations (Baker & Burton 1979). The relative absence of structure is caused partly by velocity crowding, a term first used by Burton (1966). When a lot of space is squashed into a few kilometers per second, individual interstellar components blend completely, and little structure is revealed by increasing the resolution. The importance of small velocity variations is demonstrated by the fact that essentially any low-latitude HI profile can be reproduced by a simulation, using a smooth gas density field in which the only free parameter is the line-of-sight variation of gas velocities (Burton 1972). The kinematic perturbations necessary to produce the profiles are of the same amplitude, nature, and spatial distribution as those which are, in fact, known to exist. It is not possible to similarly reproduce the observations by a simulation in which only the gas density is varied. Therefore, generally it is not reasonable to try to separate physical components of the interstellar medium by decomposing the emission spectra into Gaussian components. The Gaussian analysis provides unambiguous information only when an emission feature has a very odd velocity or is considerably brighter than its surroundings, so that it is not blended by other emission.

The situation is made even worse by intrinsically non-Gaussian contributions to the emission (due to groups of atoms with asymmetrical or in any other way pronounced non-Gaussian velocity distribution or due to saturation in optically thick regions and self-absorption by very cold foreground gas layers) and by non-uniqueness of the least squares Gaussian analysis (often several quite different solutions may fit the observations almost equally well, and the method of the least squares provides no satisfactory means for choosing between these solutions, while others, equally good or even better ones, may not be found at all). Strictly speaking, the pure method of the least squares is even not valid, when applied to this problem, as neither the form of the components nor their number is known, nor can it be assumed that the residuals are randomly distributed. The solution is often partially determined by the number of components introduced and the initial estimates of their parameters used and only partially by the observed profile. All this makes a rigorous Gaussian analysis somewhat illusive.

These weaknesses of the Gaussian analysis were understood rather early (Kaper et al. 1966; Takakubo & van Woerden 1966). Nevertheless, the method continued to be used up to the present time (e.g. Cappa de Nicolau & Pöppel 1986; Pöppel et al. 1994; Verschuur & Peratt 1999). This indicates that besides the weaknesses the method must have also some benefits. We would like to stress three of them:

1. When the line profiles lack complexity, so that at least some Gaussians are well separated, the derived parameters for

these Gaussians often yield direct information regarding the structure of the interstellar medium (Shane 1971);

2. When more complex profiles are represented in this way, the Gaussian parameters are often of little individual physical significance, but they provide a compact means for representing the observed data – hundreds of channel values in profile are replaced by some tens of Gaussian components, while physically significant information, such as mean velocities and the HI content of complexes, can be easily extracted (Shane 1971);
3. Usually some specific features in observed profiles (not necessarily corresponding to some distinct gas clouds in physical space) are represented by some specific set of Gaussians, which can be found from the overall dataset more readily than un-parameterized spectral features. The Gaussian analysis allows us to characterize general properties of the profiles from region to region in the sky and to draw conclusions, based upon similarities and differences in profile shapes (Verschuur & Peratt 1999).

Proceeding from these considerations, we attempted to create a new Gaussian decomposition program (written in FORTRAN) to extract emission line information from large HI surveys. Of course, in this new program we could not avoid the physical problems such as blending, saturation and absorption, but we tried to reduce the ambiguities inherent to the decomposition procedure. To achieve this, we introduced two modifications in otherwise rather classical decomposition procedure:

1. To reduce the number of free parameters and thus also the multitude of more or less equally good solutions, we tried to keep the number of Gaussians as low as possible. For this we tried not only to be conservative when adding new components to the decomposition, but also to reduce the number of used Gaussians by analyzing the obtained decompositions. Our aim was to find possibilities for removing some Gaussians without reducing too much the accuracy of the representation of the original profile with decomposition.
2. To reduce the ambiguity due to the selection of the initial solution, we did not consider every profile of the given survey separately, but treated the whole survey as a representation of some general, more or less smooth structure, where every observed profile must share some similarities with the ones observed in neighboring sky positions.

In this paper we describe in detail all aspects of this Gaussian decomposition program and some aspects of its application to the newest and the largest HI survey available to date, the Leiden/Dwingeloo Survey of galactic neutral hydrogen by Hartmann & Burton (1997) (hereafter the L/D Survey). This program has been successfully tested under UNIX, VMS and MS-DOS operating systems on different computers and with HI surveys of considerably different characteristics (e.g. broad-beam-width Bell Labs survey by Stark et al. 1992).

In the following chapter we describe the preparation of the survey data for decomposition. Main attention is focused on the determination of the noise level for every profile. In the third

chapter we explain the behavior of the main decomposition program and in the fourth chapter analyze the stability of the results and compare them to those of other investigators. A more detailed description of the application of this program to the L/D Survey data and the analysis of the obtained results will be postponed to the following papers of this series. After completion of these papers the actual catalog of Gaussian representation of the L/D Survey profiles will be made available on the Internet.

It is important to stress that in these papers we do not use the version of the L/D Survey, published by Hartmann & Burton (1997) on a CD-ROM, but the original observed profiles, reduced to T_b by P. M. W. Kalberla at Bonn University exactly as the published ones, but without averaging the repeated observations at identical sky positions and without re-gridding onto a common lattice. This choice was made as averaging and re-gridding smear the differences between neighboring profiles and have undesirable influence on the Gaussian decomposition process (as re-gridding is made only in galactic longitudes, the neighboring profiles become more similar in l direction than in b ; decomposition is rather sensitive to different problems in profiles and may enable some further corrections of decomposed profiles, compared to simple averaging of the results). The profiles, used here, differ from the published versions also by additional corrections for ground reflections, as shortly discussed by Kalberla et al. (1998).

2. Data preparation

As it is described in greater detail in the following chapter, our Gaussian decomposition program may return several times to every given profile in the survey. Therefore, it is reasonable to prepare the observational data for the easiest access by the main program and separate from the main program all computations, which depend only on individual observed profiles and not on the general process of decomposition. These preprocessing tasks include:

1. construction of the “road map” for the decomposition program, so that it can easily find the nearest neighbors of every profile under decomposition,
2. estimation of the noise levels rms_0 in the signal-free regions of the profiles, which are afterwards used as termination criteria of the decomposition process, and
3. estimation of the dependence of noise strength on the level of signal in profiles, described by some growth constant a and used later to assign weights to all channel values in the given profile.

Technically the first data preparation stage involves rewriting of the original profiles in a sorted order from distribution files to FORTRAN unformatted direct access files on a computer hard disk, together with a certain index for high speed access during the decomposition. Moreover, as it was mentioned in the introduction, the purpose of our program is to deal only with the emission part of the profiles and not with absorption features extending below the baseline. Therefore, to eliminate such features from the following processing, during the data copying

stage in every profile the channel values more negative than 5 times the noise level rms_0 were clipped to zero and marked. The clipping influenced mainly the profiles falling into two classes: a) some profiles, obtained from the regions around the Galactic center, where the foreground self-absorption is strong, and b) profiles containing the strongest interferences.

The second case is out of the question. Radio interference is a source of annoyance to all radio astronomers. The authors of the L/D Survey have made considerable efforts to detect and remove such features from the final survey data, but they have not always succeeded. Moreover, as they admit the subtraction of their sinc-interference model from the original profile was not always harmless: usually the rms noise of the spectrum remained higher than average, and in some cases the baseline was affected (Hartmann 1994). Therefore, clipping of high negative channel values, caused by unsuccessfully removed sinc-interferences, cannot harm the useful emission signal in the profile any more.

By clipping off real absorption parts of the profiles, we certainly lose some useful information contained in the original profile, but the approach may be justified by the rare appearance of so strong absorption (we clip only the values considerably below $T_b = 0$ K, all weaker absorption features on top of emission signal remain unaffected) and much greater stability of decomposition program, obtained by fitting only positive Gaussians to the observed profile. Therefore, we hope that in this way we gain more than lose.

The second and third preparation stages are closely related to each other. In general, the noise level in every observed profile channel is given by the radiometer equation (e.g. Kraus 1966), which may be written here as

$$T_{\text{rms}} = c(T_{\text{sys}} + T), \quad (2)$$

where T_{sys} is the mean signal-free system temperature, T corresponds to the signal part of the observations and $c = 0.00168$ for the L/D Survey (Hartmann 1994). This formula indicates that in the regions of the HI profiles, where the signal is stronger, also the noise level must be higher. Considering that near the galactic plane the highest profile peaks reach the levels of about 150 K, which is several times higher than the mean system temperature of the L/D Survey ($\langle T_{\text{sys}} \rangle = 40.66$ K according to Hartmann 1994), it is clear that there the noise level also exceeds considerably the one in signal-free regions of the same profile. Therefore, it is not very meaningful to fit all parts of the profiles with the sum of Gaussians using the same accuracy. For such case of noise correlation with signal, there are special fitting procedures, based on the maximum likelihood principle, but unfortunately, here the situation is even somewhat more complicated.

The radiometer equation (2) holds for all representations of the profiles as far as they differ from each other by simple scaling relations. Unfortunately, during the reduction of original observations to the profiles of T_b , several steps, such as bandpass corrections, baseline subtractions and stray radiation corrections, are more complicated than simple scaling. As a result, Eq. (2) is not any more strictly valid for T_b and we must

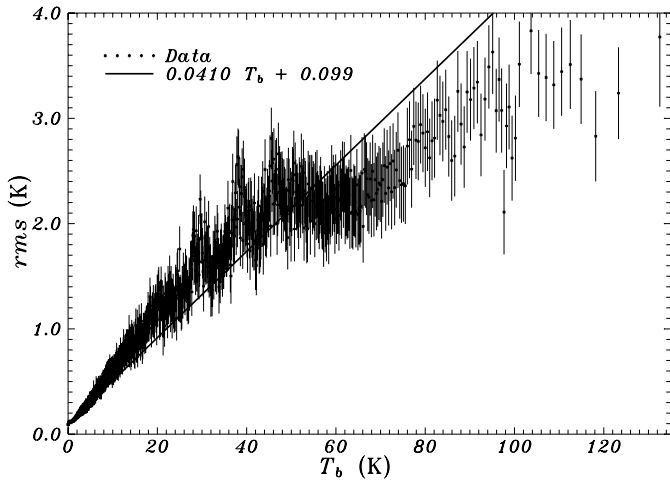


Fig. 1. The dependence of the observational uncertainties on the strength of the signal. The error bars are proportional to $n^{-\frac{1}{2}}$, where n is the number of points over which the data were averaged in a given group. The groups were larger at lower signal levels and the last group near $T_{b,s} = 130$ K contained about 500 points.

look for other possibilities for estimating the precision of final profiles. Due to the lack of exact correlation between signal and noise we also decided not to turn to special fitting methods, but to proceed in a more traditional way by using *the classical method of least squares* and trying to introduce the weights for all profile channels, depending on the expected noise level in the corresponding channel.

In the following sub-sections we examine the dependence of the observational noise and other observational and reductional errors on underlying signal strength in somewhat greater details by using different methods for the noise estimation. Of these results only those, presented in Sect. 2.3 will be directly used in actual decomposition process, but we give here also the descriptions of some other estimation possibilities, as they allow additional insights into properties of the initial data. Thus the results of Sect. 2.1 describe the general precision of the survey (Fig. 1) and the comparison of the results from Sect. 2.2 and 2.3 (Fig. 7) indicates that special actions, taken by observers on some profiles during the data reduction phase, divide all profiles into 2 distinct classes, which must be treated slightly differently during the Gaussian decomposition as well. These conclusions are described in greater detail in Sect. 2.4.

2.1. Comparison of repeated observations

The most direct method to estimate the accuracy of the final survey profiles is to compare the repeated observations at identical sky positions. If we compare the corresponding observed profiles channel by channel, then the mean over different observations must yield the estimate of the signal level $T_{b,s}$ in every channel and the *rms* of the corresponding channel values is an indicator of the precision of the profile at this signal strength.

For the comparison we separated from the survey profiles from sky positions, observed more than once. For these positions

we took all possible pairs of observations at given position and compared the results. For every channel (excluding the ones at extreme negative and positive velocities, not used later in decomposition) of every profile pair we calculated the mean of the channel values of both profiles $T_{b,s} = (T_{b,1,i} + T_{b,2,i})/2$ and the deviation $\Delta T_{b,s} = (T_{b,1,i} - T_{b,2,i})/2$. In this way we obtained for the whole survey 52 114 503 pairs of numbers, which we grouped according to the values of $T_{b,s}$. For all groups of the closest $T_{b,s}$ we computed $\sqrt{2}$ times (to account for the deviations of $T_{b,s}$ from the actual signal level) the square root of the mean of squares of deviations $\Delta T_{b,s}$ as the final estimate of the precision corresponding to the mean signal level in this group.

The results are presented in Fig. 1. It is evident that there is still a more or less linear increase of uncertainties with the signal level, but the growth speed is about 0.041, that is nearly 25 times higher than just for observational noise, accounted for by radiometer equation (constant c in Eq. 2). This high value may be understandable, if we consider that such comparison of re-observed profiles reveals not only the level of physical noise in profiles, but also the general uncertainties in the brightness temperature calibration (e.g. those related to gain instabilities and stray radiation corrections), the presence of interference spikes and other observational and reductional problems.

The use of the obtained relation as a basis for assigning the weights to profile channels may be undesirable, however, as in this way the tips of the line profiles obtain very small weights and we lose a lot of information about the profile structure during the decomposition. At the same time, a considerable amount of the uncertainties at high T_b seems to be caused by gain variations, which preserve the structure of the profile and affect only its scale. This possibility is indicated by the fact that if we correct the errors introduced by gain inaccuracy to first order, by scaling the multiple spectra to the same total integrated intensity, the points, corresponding to the highest T_b in Fig. 1, drop to less than half of their initial level. Therefore, to preserve all the available information about the shape of the emission line profile during decomposition, it seems reasonable not to assign weights on the basis of total uncertainties, but restrict ourselves just to eliminating the observational noise from the results.

Moreover, it turns out that if we try to obtain similar relations for different re-observed sky positions separately, the obtained results will differ considerably from position to position (Fig. 2) and it may be dangerous to use the same mean relation for all profiles. We also cannot determine a similar relation for every sky positions individually, as only a small part of them has been observed repeatedly. Therefore, although the comparison of re-observed profiles gave us some general impression about the accuracy of final HI profiles, we still need a better basis to assign the weights to profile channels.

In Fig. 2 we may also observe that, when the slope of the regression lines varies considerably from sky position to another, the constant terms of all formulae, which should give us the estimates of the noise level in the signal-free regions of corresponding profiles, are relatively close to each other and not much higher than the mean spectral noise level of $\langle \sigma_{rms} \rangle = 0.070$ K,

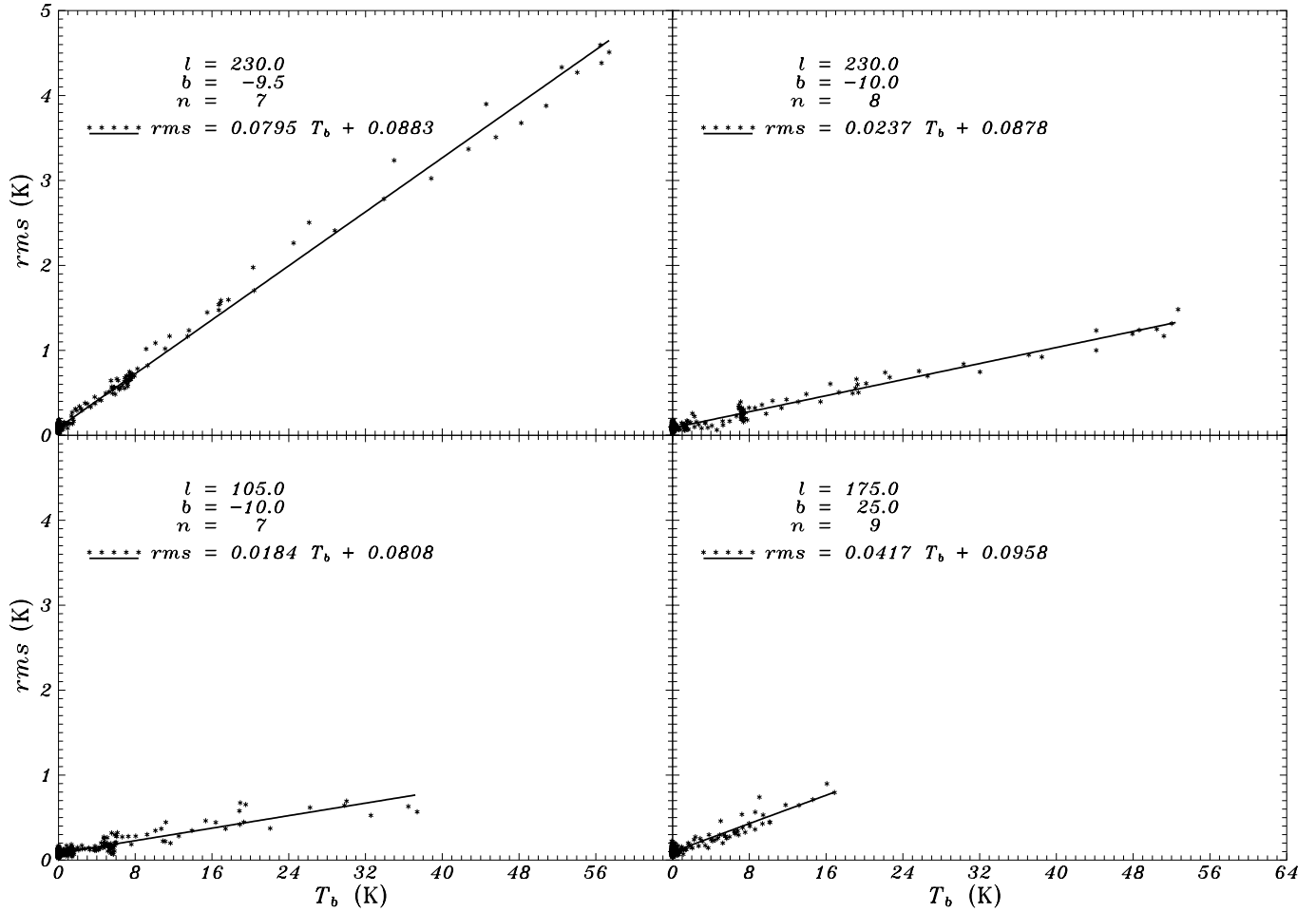


Fig. 2. The dependence of the observational uncertainties on the strength of the signal at some selected sky positions. The parameter n is the number of the observations at particular sky position.

determined by Hartmann (1994) over the entire L/D Survey. Therefore, the ordinate of the points in Fig. 3 near $T_b = 0$ K may be interpreted also as an estimate of $\langle rms_0 \rangle = 0.0909$ K for the survey (a somewhat higher value of the constant term of the regression line in Fig. 1 is caused by the deviations from the linearity of the actual distribution of the points).

2.2. Noise level from the Savitzky-Golay filtering

Examination of repeated observations demonstrated that even the dependence of the level of total uncertainties in final profiles on the signal strength can be relatively well approximated with the linear function

$$rms = aT_b + rms_0, \quad (3)$$

where T_b is the signal strength in the line profile and a , rms_0 are constants. The parameter rms_0 equals the noise level in signal-free parts of the profile and the value of the parameter a may differ considerably from one sky position to another. At the same time, the survey does not contain repeated observations for all different sky positions to estimate the value of a . As a result, we must turn to less straightforward methods of estimating the

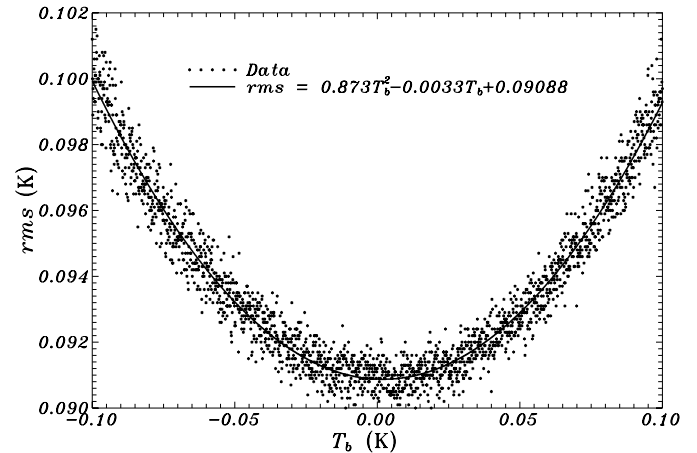


Fig. 3. The same dependence as in Fig. 1, but around $T_b = 0$ K and averaged in 0.0001 K bins in T_b .

dependence of the noise on signal, which can be used to a certain extent with every single profile, however. When doing so, we still assume that the noise dependence on signal strength can be

described by relation (3), and limit us to estimating the values of parameters a and rm_{s_0} for every profile.

Let us suppose first that in every profile the strength of the signal varies from channel to channel relatively smoothly and all high frequency fluctuations are caused by noise. In this case, we may use in first approximation the Savitzky-Golay filtering (Savitzky & Golay 1964) to separate the estimates of the underlying signal strength $T_{b,p}$ and corresponding noise level rm_{s_p} . In practice we used a (3, 3, 2) filter, which is equal to estimating the $T_{b,p,i}$ as a value at i of the second order polynomial, fitted to profile in channels $i - 3$, $i + 3$, and $rm_{s_p,i}$ as a rm_{s_p} of actual channel values of the profile around this approximating polynome in the same channel range. Combining the values of $T_{b,p,i}$ and $rm_{s_p,i}$ for all channels in a given profile, we may be able to estimate the dependence of noise level on signal strength for this profile.

As this procedure is rather sensitive to the assumption about smooth variation of the signal, we did not use the obtained results directly for estimating the values of a and rm_{s_0} in Eq. (3), but we tried to add to every pair of ($T_{b,p,i}$, $rm_{s_p,i}$) values also some estimates of their relative accuracy, and then used the linear regression procedure for the case of errors in both variables. We obtained these “error estimates” for $T_{b,p,i}$ and $rm_{s_p,i}$ of every channel by repeating the filtering with a (2, 2, 2) filter, which gives less smoothing than the one used above. The errors were estimated as differences of $T_{b,p,i}$ and $rm_{s_p,i}$ from these two cases. In this way we obtain largest errors for profile regions where the signal is rapidly fluctuating and these regions are then used in approximation (3) with lower weights. Moreover, after obtaining the preliminary values for parameters in approximation (3), we rejected all points ($T_{b,p,i}$, $rm_{s_p,i}$) with rm_{s_p} higher than $5(aT_{b,p} + rm_{s_0})$ (such points appeared only from profile regions, where the profile was extremely fluctuating and our assumptions of smoothness therefore obviously invalid), and repeated the regression estimate. The described rather arbitrary error estimates were used only for obtaining the values of constants a and rm_{s_0} in Eq. (3) and not for actual Gaussian decomposition of observed profiles.

After having actually performed this procedure on all profiles of the L/D Survey, we once again obtained for rm_{s_0} rather reasonable results, which were slightly higher than the $\langle\sigma_{rms}\rangle = 0.070$ K, given by Hartmann (1994). Averaging our results over the whole survey, we obtained in this case $\langle rm_{s_0}\rangle = 0.0907$ K. When turning to the values of a , however, the results were much more disappointing. For most profiles, except the strongest ones, the error estimates of the obtained values considerably exceeded 100% and the values themselves had scattered over an extremely wide range. Nevertheless, we studied the results in somewhat greater detail. First, we averaged the values of a for profiles of close mean signal strength $\langle T_b \rangle$ (defined as a sum of all channel values, divided by a number of profile channels). In the upper panel of Fig. 4 we can see a rather unexpected dependence of $\langle a \rangle$ on $\langle T_b \rangle$ (upper row of points), but, in principle, something like this may be caused by the interference peaks, presented in profiles, or by the unknown contribution of the stray radiation corrections.

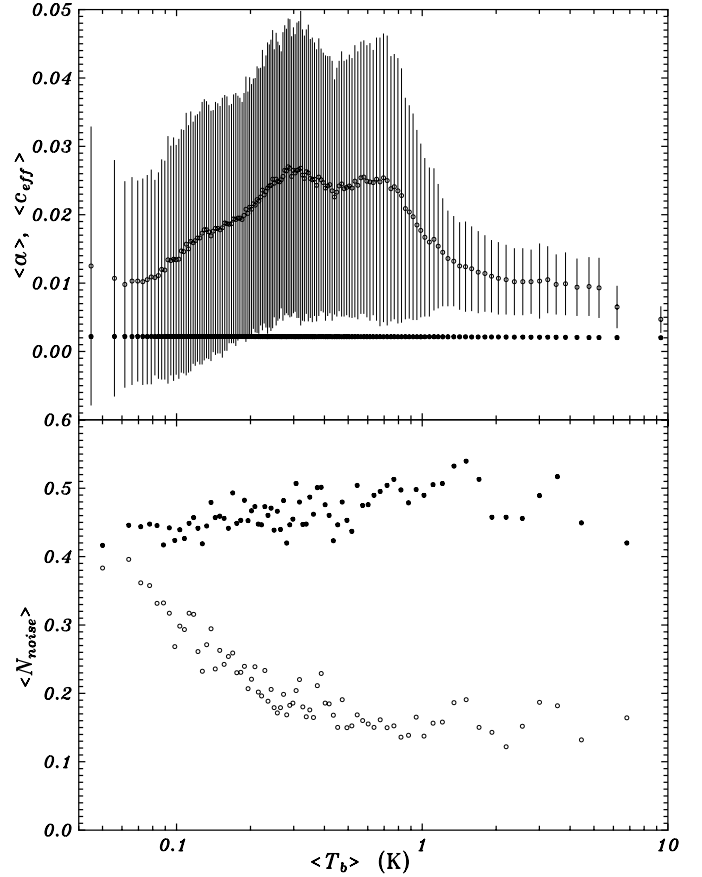


Fig. 4. Upper panel – the dependence of the mean value of a (from Eq. 3), and c_{eff} (from Eq. 4) on the mean signal strength $\langle T_b \rangle$, lower panel – the dependence of the mean number of very weak Gaussians per profile, $\langle N_{noise} \rangle$, on the mean signal strength. Open circles – values from the Savitzky-Golay filtering, filled circles – values, determined from rm_{s_0} and the system temperature. Error bars are proportional to the dispersion of the values of a and c_{eff} around their mean.

To further check the results, we proceeded to Gaussian decomposition, using the obtained values of rm_{s_0} and a (those for individual profiles). As during the decomposition the largest profile features are fitted by Gaussians in the first order and then the fitting process advances to smaller and smaller Gaussians, the number of small Gaussians in the final decomposition may be used as a certain indication of the detailness of the fit (we describe this in greater detail, when analyzing the final results of the decomposition). In the lower panel of Fig. 4 the mean numbers of very small Gaussians are plotted per every decomposed profile (lower row of points), averaged in similar ranges of $\langle T_b \rangle$, as used in the upper panel of Fig. 4. The decomposition can be considered acceptable, if all profiles of the survey are decomposed with the same average detailness. With the values of a and rm_{s_0} , used in the considered fit this is not the case. The behavior of the mean numbers of small Gaussians clearly reflects the changes in the values of a and therefore we must recognize that the described method of estimating a has failed. At the same time, this approach gave us information on the behavior of rm_{s_0} , which we discuss later.

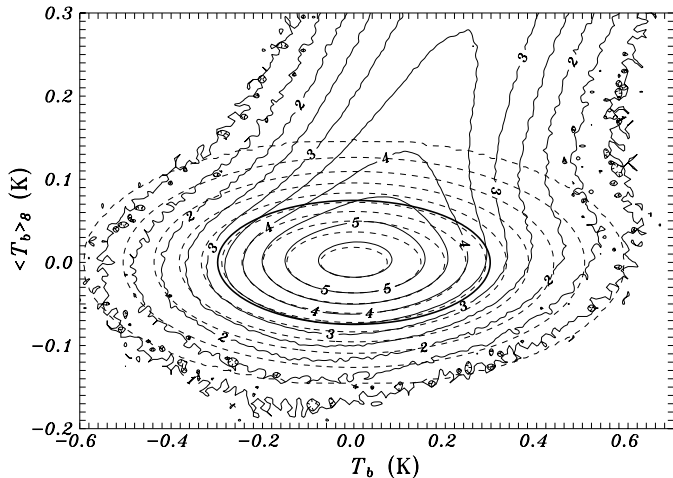


Fig. 5. The frequency distribution of $(T_b, \langle T_b \rangle_8)$ value pairs for all channels in the L/D Survey profiles (thin solid lines), compared to the pure random noise model with $rms_0 = 0.0893$ K (dashed lines). The labelled contours are drawn at density levels of $10^1, 10^2, 10^3, \dots$ points per 10^{-4} K^2 . The thick solid ellipse is drawn at the model density level, such that if all points outside it are expected to correspond to profile channels, containing some signal, only one noise peak per 2 profiles is erroneously considered as a signal.

2.3. Noise level from signal-free regions of the profiles

As our attempt to estimate the coefficient a directly from the profiles failed, and we still need some relation between the signal strength and the corresponding noise level, we introduced an additional assumption: let us suppose that the reduction process just scales the noise level by some factor x , where x is a constant for every profile, which does not depend on signal strength in any particular channel. In this case, we can still use Eq. (2), if we replace the original value of c with some effective value c_{eff} . This approach is partly justified by the fact that for the strongest emission profiles near the galactic plane, where the weighting is most important, one of the largest corrections during the reduction phase is due to stray radiation, which in this case is dominated by the near-side-lobe (NSL) contribution (Hartmann 1994). As the NSL correction largely mimics the shape of the spectrum in question, its general effect is similar to down-scaling the whole T_b profile, which in turn, with respect to the noise, can be described as an increase of the constant in the radiometer equation. Of course, at high latitudes, where the emission is very low, the stray radiation is almost un-correlated with the signal, but still a significant fraction of the emission is removed. It is very hard to estimate how this would affect the noise statistics, but the lower the profile intensities the less important for decomposition are the exact weights, as they become in any case nearly equal for all profile channels.

Therefore, in our next attempt to assign weights to profile channels we used Eq. (2), rewriting it as:

$$rms = c_{\text{eff}}(T_{\text{sys}} + T_b). \quad (4)$$

In this case we do not need any independent estimates of c_{eff} , but having determined the noise level rms_0 in signal-free ($T_b =$

0 K) regions of the profiles, we obtain $c_{\text{eff}} = rms_0/T_{\text{sys}}$, where for T_{sys} we may use the values from the headers of the profile records.

To estimate rms_0 for every profile, let us suppose that all the noise in profiles has normal distribution with mean $\mu = 0$ and dispersion σ^2 . In this case, if the profile contains no signal, the estimation of the noise level reduces to the calculation of rms of all channel values and the mean of channel values over any n channels is a random number with mean $\mu_n = 0$ and $\sigma_n^2 = \sigma^2/n$. However, as every line of sight in the Galaxy contains substantial hydrogen emission, we must first determine the emission-free regions of the profile, which can be used as representative intervals to estimate the rms_0 for the whole profile. To discriminate the profile regions, containing some emission signal, we proceed from the fact that in these regions the mean of channel values must be raised above zero by the H I emission and therefore, we may suppose that in every region of the profile, containing some signal, for at least one channel the signal plus noise value is above zero.

To mark in profiles the ranges of channels, probably containing some H I emission signal, we computed for every channel with $T_b > 0$ the mean $\langle T_b \rangle_n$ for n channels from both sides of the selected one (altogether $2n$ channels). For final decomposition we chose $n = 8$, so that our averaging window was about the size of FWHM (Full Widths of the line at the level of Half Maximum) for most of the well separated emission line components. This was chosen as a compromise between larger values, giving us too much smoothing to detect narrow and weak emission features, and smaller ones, giving too much importance to higher single-channel noise peaks.

In Fig. 5. we have plotted by thin solid lines the frequency distribution of $(T_b, \langle T_b \rangle_8)$ value pairs for all channels in the L/D H I survey profiles (before the clipping of absorption components). To this distribution the above described pure random noise model is fitted (dashed lines) by adjusting the mean survey noise level $\langle rms_0 \rangle$ as a free parameter. We can see that with $rms_0 = 0.0893$ K, in general, the model fits the observed distribution very well. Only in the upper right-hand part of the figure we can see considerable deviations of the observed distribution from the pure noise model. It is clear that these deviations must have been caused by the H I emission contained in the observed profiles. The slight deviations in the lower left-hand part of the figure have been caused by absorption and interference peaks in original profiles.

This figure indicates the way for discriminating the noise and signal regions in the observed profiles: if for some channel the point $(T_b, \langle T_b \rangle_8)$ lies near the center of the figure or outside its first quadrant, the channel probably contains only noise, but if the corresponding point lies in the first quadrant at a considerable distance from the center ($T_b = 0, \langle T_b \rangle_8 = 0$), it probably contains some emission signal and it, together with its neighbors must be excluded from the calculation of the channel-by-channel rms for this profile. As a criterion for treating the point as located far from the distribution center, we used the requirement that only about one strong noise peak per two profiles

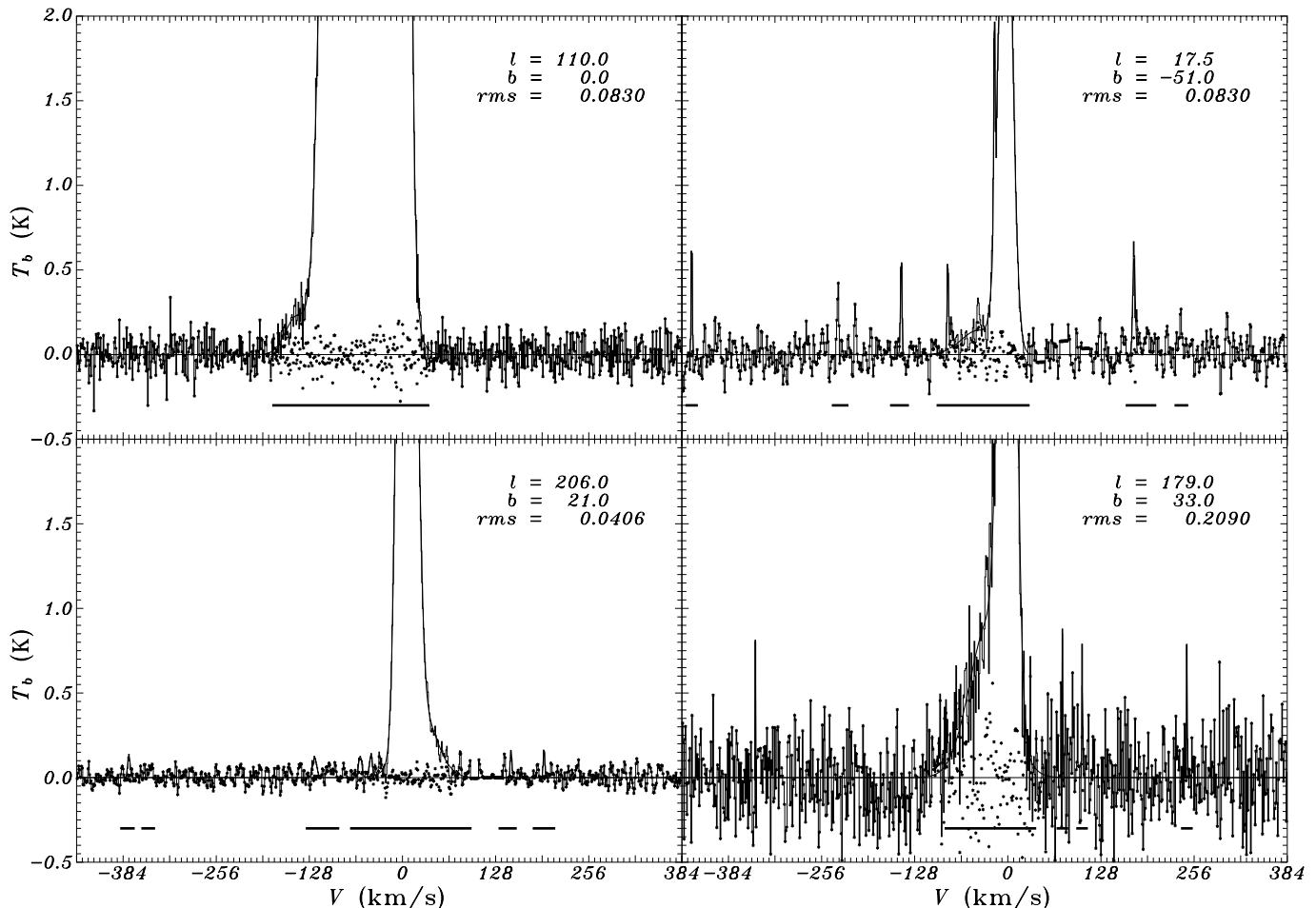


Fig. 6. Examples of the results of choosing signal-free regions of the profiles. The observations are given by a thin stepped line, the Gaussian model with a smooth thicker line and residuals are represented by stars. The channels, excluded as possibly containing some signal are underlined by a thick solid line.

may be considered as a signal. The border of the corresponding region is marked in Fig. 5 by a thick solid ellipse.

In practice, to process the profiles with different noise levels, the signal-free regions of the profiles were determined iteratively following the Expectation-Maximization (EM) algorithm (e.g. Little & Rubin 1987). We started with signal and noise discrimination criteria as described above, but after obtaining the first guess for the signal-free region, we computed the rms_0 for this region and modified the selection criteria (the bold ellipse in Fig. 5) according to this first estimate, and repeated the procedure until the estimated signal-free region did not change any more, or became periodically fluctuating. In the latter case, we accepted the estimate, corresponding to the smallest signal-free region among the periodically repeated ones. In general, there was no need for more than 5 iteration steps.

Fig. 6 gives some examples of the final results of this procedure for some profiles with a different noise level. To demonstrate, which features in the profiles were considered by the above-described procedure as possible signal peaks, the profiles in this figure were chosen from those, in which relatively many signal regions were detected. The only exception is the

one at $l = 110^\circ$, $b = 0^\circ$, which has been chosen from the most typical profiles of the survey. Actually, for about 43% of the profiles, only one signal feature was detected, and for about 36% two signal features. The division of profile peaks to the signal and noise ones is made here only for estimating the rms_0 of individual profiles, and this information is not handed over to Gaussian decomposition program. The latter treats all channels of every profile as probably containing a signal.

After obtaining estimates for rms_0 , we calculated the corresponding c_{eff} as described above. The obtained dependence of mean c_{eff} values on $\langle T_b \rangle$ is given as the lower row of points on the upper panel of Fig. 4. We can see that these results have very low dispersion around the general mean value of $\langle c_{\text{eff}} \rangle = 0.00216$ and practically no systematic dependence on $\langle T_b \rangle$. The corresponding mean numbers of very small Gaussians in decomposition are given as the upper row of points on the lower panel of Fig. 4. We can see that there is no considerable dependence on $\langle T_b \rangle$ any more, which indicates that all profiles have been decomposed with approximately the same level of details. Moreover, for all values of $\langle T_b \rangle$ we have about 0.5 small Gaussians per profile, or in other words, approximately

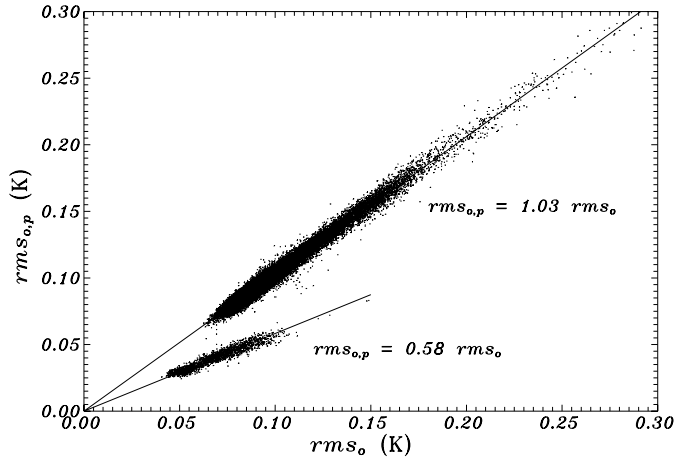


Fig. 7. The comparison of the estimates of the noise level in the signal-free regions of the profiles. $rm_{s_{0,p}}$ – estimates, obtained by the Savitzky-Golay filtering (polynomial fit), rm_{s_0} – estimates, directly calculated for presumably signal-free regions.

one noise peak is fitted with a Gaussian per 2 profiles, as it was planned. All this indicates that regardless of several rather arbitrary assumptions, this approach may give us the best description of the noise level on the profile signal strength.

2.4. Comparison of the noise estimates and the weights

Let us now turn back to the estimates of rm_{s_0} . If everything is correct, the results, obtained in Sect. 2.3, must agree with those from the polynomial fit method in Sect. 2.2, as both of them are the estimates of the noise level in signal-free regions of the profiles. First of all, we may calculate the general mean of the obtained rm_{s_0} estimates from Sect. 2.3 over the L/D Survey $\langle rm_{s_0} \rangle = 0.0882$ K, which in this case is about 3% smaller than the estimate obtained in Sect. 2.2. A more detailed comparison is given in Fig. 7. We can see that for most profiles the agreement is good, but on the average, the polynomial fit method gives the results, which are about 3% higher than those obtained by identifying the signal free regions. This is evidently caused by the fact that the second method considers some higher noise peaks as a signal – the behavior that was known in advance. To some extent, this may have even a positive effect, as it guarantees that during decomposition we do not miss some weak signal features, but for the sake of safety we fit a small number of the strongest noise peaks with Gaussians.

However, in Fig. 7 we can see also a group of profiles for which the first approach gives systematically 42% lower values for the noise level. A closer examination reveals that about 1.22% of all survey profiles contain some kind of periodic fluctuations with a period of 32 channels and an amplitude of the order of some milli-kelvins. Finally, it turned out that these are the profiles, originally containing strong sinc-pattern interferences, which were mathematically removed (that caused the appearance of the above-mentioned fluctuations) during the reduction and which were thereafter hanning-smoothed to formally reduce the remaining high noise level. As such smoothing smears out

and broadens all noise features, the polynomial fit method for estimation of the noise level interprets a substantial part of this smoothed noise as a possible signal and gives underestimated values of rm_{s_0} .

As the second method of determining the noise level is less sensitive to different smoothings performed on the profiles, we preferred for the following this one, but even so we must consider an effect introduced by such smoothing. When the values of rm_{s_0} , obtained by the second method, can be directly used in the decomposition process, we cannot apply them to calculate c_{eff} , as such smoothing changes rm_{s_0} , but not the degree of dependence of the noise on the signal. Therefore, for all such smoothed profiles we did not calculate c_{eff} from rm_{s_0} and T_{sys} , but used just the mean value $\langle c_{\text{eff}} \rangle = 0.00216$. The actual weights for every channel i of every profile were finally calculated according to the formula

$$w_i = \left(\frac{rm_{s_0}}{rm_{s_0} + c_{\text{eff}} T_{b,i}} \right)^2. \quad (5)$$

As we do not know $T_{b,i}$ in advance, we had to turn to iteratively re-weighted least squares as a sub-case of EM algorithm by replacing for the first iteration step $T_{b,i}$ with the observed channel values $T_{b,o,i}$ of the profile, and with the best Gaussian decomposed model profile $T_{b,m,i}$, obtained so far, for all following iteration steps.

3. Gaussian decomposition

The decomposition of each profile to Gaussians proceeds through two stages: adding new Gaussians to the fit until the residuals rm_{s_r} (defined by Eq. 6) is below rm_{s_0} , and then looking for possibilities to reduce the number of the used Gaussians without increasing the rm_{s_0} substantially above rm_{s_0} . In the following, we describe both stages in detail and give also the rules how the decomposition process proceeds from one survey profile to another.

3.1. Adding new Gaussians

The decomposition of every profile starts with reading the observed channel values $T_{b,o,i}$ for this particular profile into the memory and computing for it the array of weights w_i , as described above. Then the initial trial Gaussian approximation is chosen (more about this later) and the values of its parameters are adjusted to get the best fit with this number of Gaussians. For adjustment of Gaussian parameters by least squares fitting we use the Levenberg-Marquard method of nonlinear modeling. In every iteration, made by this method, we use for solving the system of linear equations the standard TFSPDM Fortran subroutine. When this method fails due to the system singularity, the program switches over to a much slower, but a more stable singular value decomposition technique (based on the Fortran subroutine SVDCMP from Press et al. 1992) and keeps using it until the next call to the Levenberg-Marquard routine.

After finishing the iterations with the given number of Gaussians, the array of residuals $T_{b,r,i} = T_{b,o,i} - T_{b,m,i}$ is filled in and the residual rms_r computed as

$$rms_r = \sqrt{\frac{\sum_{i=1}^{n_c} w_i T_{b,r,i}^2}{n_c - 1}}, \quad (6)$$

where n_c is the number of profile channels, used for Gaussian decomposition and $T_{b,m,i}$ are the channel values of the Gaussian decomposed model profile.

If the obtained rms_r is greater than the initial estimate rms_0 for signal-free regions of the given profile, we try to improve the decomposition by adding an additional Gaussian. To get the initial trial values of the parameters of this Gaussian, we examine the residuals $T_{b,r,i}$, looking for all channels, corresponding to local maxima of $T_{b,r,i}$ with $T_{b,r,k} > rms_0$. At every such channel k , we smooth the residuals with a (1, 1, 2) Savitzky-Golay filter and record the parameters a , b and c of the fitting parabola $aj^2 + bj + c$ of this filter. If $a < 0$ and $c > 0$ for the parabola, we consider it as the first 3 terms of the power series approximation of the Gaussian (1) and calculate the Gaussian parameters as

$$\sigma_\nu = \frac{c}{\sqrt{b^2 - 2ac}}, \quad (7)$$

$$\nu_0 = \frac{bc}{b^2 - 2ac}, \quad (8)$$

$$T_{b,0} = ce^{\frac{b^2}{2(b^2 - 2ac)}}. \quad (9)$$

Next we estimate the decrease of rms_r , caused by adding this Gaussian to the decomposition. If the result is positive, we will study similarly also somewhat wider surroundings of the same channel k by applying a (4, 4, 2) filter. If the corresponding Gaussian gives greater decrease of rms_r than the one, corresponding to the (1, 1, 2) case, we proceed to an even wider filter (16, 16, 2) and try to use the parameters of the Gaussian, corresponding to this window. Finally, after going through the whole profile of residuals, we choose for adding into decomposition the Gaussian, giving the largest decrease in rms_r . This Gaussian is added to the decomposition and the above-described fitting procedure is performed with an increased number of Gaussians. All this is repeated until rms_r becomes less than rms_0 . The obtained value of rms_r is recorded as the best one, $rms_{r,b}$, for future use.

3.2. Removing Gaussians from decomposition

The number of Gaussians in decomposition may be reduced in three ways. First of all, at any iteration of the Levenberg-Marquard procedure the solution, obtained so far, is checked for the presence of Gaussians with incredible values of parameters and, if found, these Gaussians are immediately rejected from further calculations. At this stage, we remove the Gaussians with the height greater than 400 K or smaller than 0.0001 K. We also remove the Gaussians with a center far outside the decomposable region of the profile so that in any channel of the

profile its contribution is below 0.0001 K, and reject approximations, containing components with the width greater than the whole length of the profile under decomposition. These components are actually not needed for the fit and their removal does not influence the rms_r of the fit (those with $T_{b,0} > 400$ K have an extremely narrow width and are located between channels). Two other possibilities to reduce the number of Gaussians in the final solution are examined only after finding for the given profile some decomposition with $rms_r < rms_0$ and applied only, if this does not increase rms_r considerably above rms_0 .

After obtaining acceptable decomposition with $rms_r \leq rms_0$ we check, whether it contains some Gaussians with relatively similar locations and widths. In this context we consider Gaussians similar, if the parameter

$$S' = \frac{\int_{-\infty}^{\infty} (T_{b,i} - T_{b,j})^2 dV}{\int_{-\infty}^{\infty} T_{b,i}^2 dV + \int_{-\infty}^{\infty} T_{b,j}^2 dV}, \quad (10)$$

where $T_{b,i}$ and $T_{b,j}$ are two Gaussians, given by Eq. (1), is small (as $T_{b,0}$ enters (1) only linearly, we regard at this stage all Gaussians as having equal heights).

We compute the similarity parameter S' for all pairs of Gaussians in the given decomposition and try to replace the most similar pair by a Gaussian, having parameters approximated as:

$$\sigma_{\nu,0} = \frac{T_{b,1}\sigma_{\nu,1}^2 + T_{b,2}\sigma_{\nu,2}^2}{T_{b,1}\sigma_{\nu,1} + T_{b,2}\sigma_{\nu,2}} \left[1 + \frac{T_{b,1}T_{b,2}}{(T_{b,1} + T_{b,2})^2} \cdot \left(\frac{T_{b,1}\sigma_{\nu,1}^2 + T_{b,2}\sigma_{\nu,2}^2}{T_{b,1}\sigma_{\nu,1} + T_{b,2}\sigma_{\nu,2}} \right)^2 (\nu_1 - \nu_2)^2 \right], \quad (11)$$

$$\nu_0 = \frac{\frac{T_{b,1}\sigma_{\nu,1}\nu_1}{\sqrt{(\sigma_{\nu,1}^2 + \sigma_{\nu,0}^2)^3}} + \frac{T_{b,2}\sigma_{\nu,2}\nu_2}{\sqrt{(\sigma_{\nu,2}^2 + \sigma_{\nu,0}^2)^3}}}{\frac{T_{b,1}\sigma_{\nu,1}}{\sqrt{(\sigma_{\nu,1}^2 + \sigma_{\nu,0}^2)^3}} + \frac{T_{b,2}\sigma_{\nu,2}}{\sqrt{(\sigma_{\nu,2}^2 + \sigma_{\nu,0}^2)^3}}}, \quad (12)$$

$$T_{b,0} = \sqrt{2} \left[\frac{T_{b,1}\sigma_{\nu,1}}{\sqrt{\sigma_{\nu,1}^2 + \sigma_{\nu,0}^2}} e^{-\frac{(\nu_1 - \nu_0)^2}{2(\sigma_{\nu,1}^2 + \sigma_{\nu,0}^2)}} + \frac{T_{b,2}\sigma_{\nu,2}}{\sqrt{\sigma_{\nu,2}^2 + \sigma_{\nu,0}^2}} e^{-\frac{(\nu_2 - \nu_0)^2}{2(\sigma_{\nu,2}^2 + \sigma_{\nu,0}^2)}} \right]. \quad (13)$$

After this replacement we repeat the least squares adjustment of parameters of all Gaussians in decomposition and if the resulting $rms_r < rms_0 + |rms_0 - rms_{r,b}|$, we accept this new version of decomposition and record corresponding rms_r as $rms_{r,b}$. In this way we try to tune $rms_{r,b}$ as close to (not necessarily smaller than) rms_0 as possible.

This procedure is repeated until the replacement of the pair of most similar Gaussians by a single component increases the rms_r over the above described limit. When this situation is reached, we still try to remove from decomposition some smallest Gaussians. For this we rearrange all components in decomposition into the order of the decreasing area under the Gaussians (as a result of this step, the Gaussians in the final solution are presented in the same order) and try to reject the last one. For this we once again repeat the least squares adjustment of parameters

of all Gaussians, except the rejected one, in decomposition, and if the resulting rms_r is still less than $rms_0 + |rms_0 - rms_{r,b}|$, we accept this new version of decomposition and record the new value for $rms_{r,b}$. If the step was successful, we return to the examination of the similarity of Gaussians in decomposition, if not, the obtained solution is saved as the best one for this profile at the given stage of decomposition and the program moves to the next profile.

3.3. Selecting the next profile

After completing the decomposition of one profile, the program must advance to another, but this is not a simple monotonous move in our program. At this stage we try to take into account the possible similarity of hydrogen profiles at neighboring sky positions. We consider every sky position P_0 to have 4 neighboring positions, 2 of which are in forward direction (P_1 at the same galactic latitude b_0 and P_2 at the latitude $b_0 + 0.5^\circ$) and 2 in backward direction (P_3 at the same galactic latitude b_0 and P_4 at the latitude $b_0 - 0.5^\circ$). To find the data, corresponding to the positions P_{1-4} , we use the ‘‘road map’’, constructed before the actual decomposition process.

Now we suppose that the results of the decomposition of similar profiles must be similar as well. Therefore, after obtaining the results at position P_0 , first of all we compare these with the ones obtained earlier at positions P_3 and P_4 . If the decomposition at one of these positions is worse than the one obtained at P_0 , the program shifts back to this position $P_?$ ($P_? = P_3$ or $P_? = P_4$). If both previous decompositions are worse than the current one, the program selects the position, corresponding to the worst decomposition. The decomposition is considered worse than the other one, if it contains more Gaussians, or, in the case of an equal number of Gaussians, if its $rms_{r,b}$ differs from the corresponding rms_0 more than in the case of the other profile.

In such backward motion the re-decomposition of the profile at position $P_?$ is always started with the current best solution for position P_0 as an initial trial Gaussian approximation. If the re-decomposition at $P_?$ yields better results than those obtained there so far, the earlier results will be replaced by the new ones and this position $P_?$ becomes a new current position P_0 ($P_0 := P_?$). If the re-decomposition does not succeed, only this fact is recorded to prevent multiple trials with the same initial approximation. The observation of the actual behavior of the program with real data demonstrates that sometimes it takes 4–5 or even more steps backward, improving the decompositions obtained earlier.

If the decompositions at the backward neighborhood positions P_3 and P_4 are better than the one obtained at P_0 , or the re-decomposition attempts did not succeed, the program moves on in forward direction. If at least one of the forward neighborhood positions P_1 or P_2 has already been visited earlier, the program moves to the one with the worst results obtained so far, and selects this position as a new current position P_0 (regardless of the results of re-decomposition, to avoid closed traps). For re-decomposition at this new current position the best results

for backward neighbors are used as initial approximation. If the re-decomposition in this new position is more successful than the previous ones, the earlier results are replaced by new ones, if not, the results of the re-decomposition are rejected.

If none of the forward neighborhood positions P_1 and P_2 have not been visited earlier, the program moves to the position P_1 , or, if this position does not exist (all positions at this galactic latitude have already been decomposed), to position P_2 . As an initial approximation for this new position is chosen the best decomposition from its backward neighborhood positions. We stress that by the best decomposition we mean the one which gives $rms_{r,b}$ close to rms_0 with the smallest number of Gaussians, or, in the case of an equal number of Gaussians, the one with $rms_{r,b}$ closest to rms_0 .

4. Stability of the obtained decompositions

As mentioned in the Introduction, a serious problem with the Gaussian decompositions of complex profiles is that we can never be sure that we have obtained the absolutely best possible decomposition for the given profile, in the sense explained at the end of the previous section. We have tried to reduce this uncertainty by considering the probable similarity of hydrogen profiles at neighboring sky positions and by taking special measures to keep the number of Gaussians low. Nevertheless, for complicated profiles near the Galactic plane the number of Gaussians, used to decompose a single profile, is usually around 20 or even more, which brings the number of free parameters to about 60. With such degree of freedom it is still hard to believe that the obtained results are absolutely the best ones and do not depend on the applied algorithms of the decomposition, the applied procedures of weighting the channel values, the methods for determining the noise level of the observed profiles and also on the uncertainties at these steps. Therefore, we must admit that the decompositions obtained by us are still just one possibility and by no means the final word in this field. However, for most applications of such decompositions these are probably not the key questions and more important is the internal consistency of the obtained results. This means that it would be interesting to know to what extent our particular decompositions are determined by the underlying signal in profiles and how much they are influenced by noise and other disturbing factors.

4.1. Comparison with other authors

We can estimate the value of decomposition in a most direct way, if we can compare our results to those of other authors. This became possible due to a recent paper by Verschuur & Peratt (1999) (VP). Their Table 3 contains parameters of Gaussians fitted to 8 different L/D Survey profiles. This may be an ideal material for comparison, as VP have used a completely different approach to Gaussian decomposition – their work is based on visual inspection of the observed profiles and the computer is used only ‘‘for honing the final fit’’. However, before actually comparing the results, we must give some comments.

First of all, VP and we have used different versions of the L/D Survey. VP have used the official published version with averaged and re-gridded profiles, but we have used the real observed profiles, reduced to T_b . Moreover, as mentioned by G. L. Verschuur, for them “the perfection of fit is not the issue, but the statistics of the results” (private communication). These differences introduce some additional dissimilarities into our decomposition results, but we hope that all these problems are relatively minor ones and the comparison is still informative.

For comparison we have used all profiles whose sky positions are given in Table 3 of VP, except the case, corresponding to their Fig. 2(d). In this case the data, used by VP, and the obtained decomposition actually correspond to the profile at $l = 47^\circ 5$, $b = -20^\circ 5$. The results of the comparison are given in Table 1 and Fig. 8(A–H), where we have rearranged the profiles into the order, where the first ones have been decomposed by VP with less Gaussians than we had and the last ones have less Gaussians in our decomposition. From these profiles, the real disappointment for us is only the first one (A), where VP have managed to fit the main peak with one Gaussian, but our program needed for it two Gaussians. At the same time, we must draw attention to the fact that the secondary peak of the same profile is remarkably mis-fitted by VP and this may be easily the key, opening the possibility to use only one Gaussian for the main peak. The second additional Gaussian in our decomposition of profile (A), is a rather weak, but wide one at 45.2 km s^{-1} , which may be due to a baseline problem.

Our decomposition program yielded more Gaussians than used by VP, also for profile (B), but here both of the additional Gaussians are rather weak. The first of them at -13.5 km s^{-1} is added to fit the feature on the wing of the profile, which VP have neglected, and the other one slightly rises the model profile on the other side of the line peak. Both components may have been caused by slight differences in our quality criteria, used during the fitting procedure.

For all other profiles in the comparison set our program has used the same number (cases C and D) or less Gaussians (profiles E, F, G, H) than introduced by VP, but the fits never match exactly with each other. Although it is easy to compare Gaussians, describing some main features of the decomposed profiles, there may be considerable differences in the usage of weaker components. This seems to be due to differences in the aims of the decomposition. While VP have used Gaussians just to describe the structure of the main emission feature of the profile, we have tried to fit the whole profile. As a result, VP have used 4 Gaussians for both profiles (C) and (D), placing them in the region, where the signal is considerably above the zero level and leaving wider surroundings unfitted. We have managed to fit the main peak with only 3 Gaussians, but used the fourth one to describe the slight deviations of the profile from the zero level in much wider surroundings of the line.

The differences in the decompositions of profiles (E) and (F) are rather similar to that of profile (B), but when in the case of (B) our program used 2 more components, if compared to the VP solution, in the case of (E) and (F) VP used 2 additional Gaussians per profile, in comparison to our solution. All these

extra Gaussians are rather weak and insignificant, however. At the same time, the decompositions of profiles (G) and (H) by VP are rather surprising, because all used components are of the same order of magnitude and they have still used more of them than actually needed.

The number of Gaussians is not the only criterion to evaluate the fit. It is also important to know the level of residuals rms . Corresponding values are given in Table 1, but we must stress that they are rather incomparable numbers. First of all, for all averaged (case F: the published profile is an average of two observed ones) or interpolated (cases C, D, E, F) profiles, used by VP, the noise level of the data is reduced below the level of the directly observed profiles used by us. This explains why for practically all these profiles the rms of the fit, obtained by VP is considerably below the level reached by us. On the other hand, for cases, where we have used equal data (profiles A, B, G, H), the rms of the fits by VP is increased by the fact that we could use their values of Gaussian parameters only with the precision published in their paper, but our own values we used with full computer precision. This effect of the roundup errors is most visible for the weakest profiles (G) and (H), but can be identified also in some other cases, if to compare the plots of Gaussian components, as presented by us and by VP.

In general, we may state that our program needed only 32 Gaussians to fit 8 profiles in the comparison set, whereas VP used 39 Gaussians. It is somewhat hard to compare the residual levels of the obtained fits, but there seem to be no unintelligible differences between the values, reached by us and by VP. The only exception may be profile (A), for which the residuals rms , obtained by VP is nearly twice as high as the value for our fit, but for this profile the region around the secondary maximum of the line is clearly mis-fitted by VP. When comparing the individual Gaussians, used in the fits of corresponding profiles, we can see that in most cases the strongest components in fits of different authors are in reasonable agreement, but the components with $T_{b,0} < 0.3 \text{ K}$ are fitted often in a completely different manner. Therefore, the uniqueness of the fit is still a considerable problem even for relatively simple profiles, but this also seems to indicate some advantages of our program, as in comparison with an experienced human researcher, it manages to find the solutions with a smaller number of free parameters.

4.2. Internal stability of results

The comparison with the Gaussian decomposition, made by VP, is based on a rather small number of profiles and includes also the incompatibility, introduced by a different method of decomposition, used by these authors. Therefore, a somewhat different question is, what the stability of obtainable decompositions is within our program. To study this, we decided to decompose many times one and the same typical signal profile of the L/D Survey and to compare the resulting Gaussian parameters.

To select such typical profile, we first counted how frequently the survey profiles have been decomposed with different numbers of Gaussian components. It turns out that most frequently the program uses 6 components per profile (in 20004

Table 1. Comparison of VP and our decompositions of some L/D Survey profiles

	VP decomposition			Our decomposition		
	$T_{b,0}$ (K)	V_0 (km s ⁻¹)	FWHM (km s ⁻¹)	$T_{b,0}$ (K)	V_0 (km s ⁻¹)	FWHM (km s ⁻¹)
A	$l = 47^\circ 5,$	$b = -20^\circ 5,$	$rms = 0.146$ K	$l = 47^\circ 5,$	$b = -20^\circ 5,$	$rms = 0.082$ K
	3.1	15.3	34.6	4.1	13.1	32.0
	6.4	9.0	17.4	5.0	10.1	8.5
	20.7	1.5	3.5	14.6	1.5	2.9
	0.6	22.2	6.1	1.9	18.2	9.9
				9.2	1.8	5.9
				0.1	45.2	65.3
B	$l = 211^\circ 0,$	$b = 28^\circ 0,$	$rms = 0.117$ K	$l = 211^\circ 0,$	$b = 28^\circ 0,$	$rms = 0.097$ K
	2.8	0.4	32.5	2.1	0.1	35.1
	3.6	5.5	13.4	3.9	4.9	16.0
	14.4	8.0	5.1	14.3	8.0	5.3
	4.4	12.3	3.9	4.0	12.4	3.7
				0.4	-13.5	5.0
			0.1	49.8	16.2	
C	$l = 56^\circ 0,$	$b = 66^\circ 0,$	$rms = 0.083$ K	$l = 56^\circ 1,$	$b = 66^\circ 0,$	$rms = 0.080$ K
	0.4	-19.4	50.6	0.7	-18.3	38.8
	0.5	-10.4	28.5			
	1.2	-1.3	12.1	1.2	-2.3	14.1
	0.6	-1.3	4.3	0.8	-1.7	4.7
				0.1	41.6	85.8
D	$l = 119^\circ 0,$	$b = -78^\circ 0,$	$rms = 0.075$ K	$l = 119^\circ 7,$	$b = -78^\circ 0,$	$rms = 0.088$ K
	1.5	-9.5	29.6	2.5	-8.6	22.9
	1.9	-7.4	13.7			
	1.5	-6.8	6.5	2.6	-7.0	7.5
	0.1	-44.6	10.2			
				0.1	-21.6	147.9
E	$l = 83^\circ 0,$	$b = 52^\circ 0,$	$rms = 0.059$ K	$l = 83^\circ 3,$	$b = 52^\circ 0,$	$rms = 0.078$ K
	0.8	-5.2	36.5	0.7	-7.2	38.8
	1.8	-2.6	14.5	1.8	-2.5	14.7
	0.9	-5.1	4.3	0.8	-5.4	4.1
	0.9	-39.0	14.5	1.1	-40.8	14.1
	0.1	-50.6	20.5			
	0.1	31.4	9.4			
F	$l = 233^\circ 0,$	$b = 50^\circ 0,$	$rms = 0.065$ K	$l = 233^\circ 1,$	$b = 50^\circ 0,$	$rms = 0.090$ K
	2.1	-9.0	32.8	1.6	-8.8	37.0
	3.2	-10.1	12.8	3.5	-9.6	14.7
	4.8	-13.2	3.7	5.0	-13.2	3.7
	0.6	-3.4	3.6	0.5	-3.7	2.3
	0.2	2.5	5.2			
	0.1	-46.0	9.1			
G	$l = 85^\circ 5,$	$b = 76^\circ 0,$	$rms = 0.081$ K	$l = 85^\circ 5,$	$b = 76^\circ 0,$	$rms = 0.080$ K
	0.3	-3.3	50.7	0.4	-7.0	47.6
	0.1	-19.6	28.0			
	1.1	0.6	12.4	1.3	-0.2	11.1
	0.3	-0.6	6.4			
H	$l = 183^\circ 0,$	$b = 62^\circ 0,$	$rms = 0.090$ K	$l = 183^\circ 0,$	$b = 62^\circ 0,$	$rms = 0.073$ K
	0.6	-3.8	34.0	0.7	-6.5	38.8
	0.8	-14.4	14.0	0.8	-13.5	14.0
	0.1	-50.8	30.8			
	1.0	-54.2	14.9	1.1	-54.9	21.5
	0.2	-64.9	11.5			
	0.2	-39.5	10.2			
	0.3	-7.8	5.0			

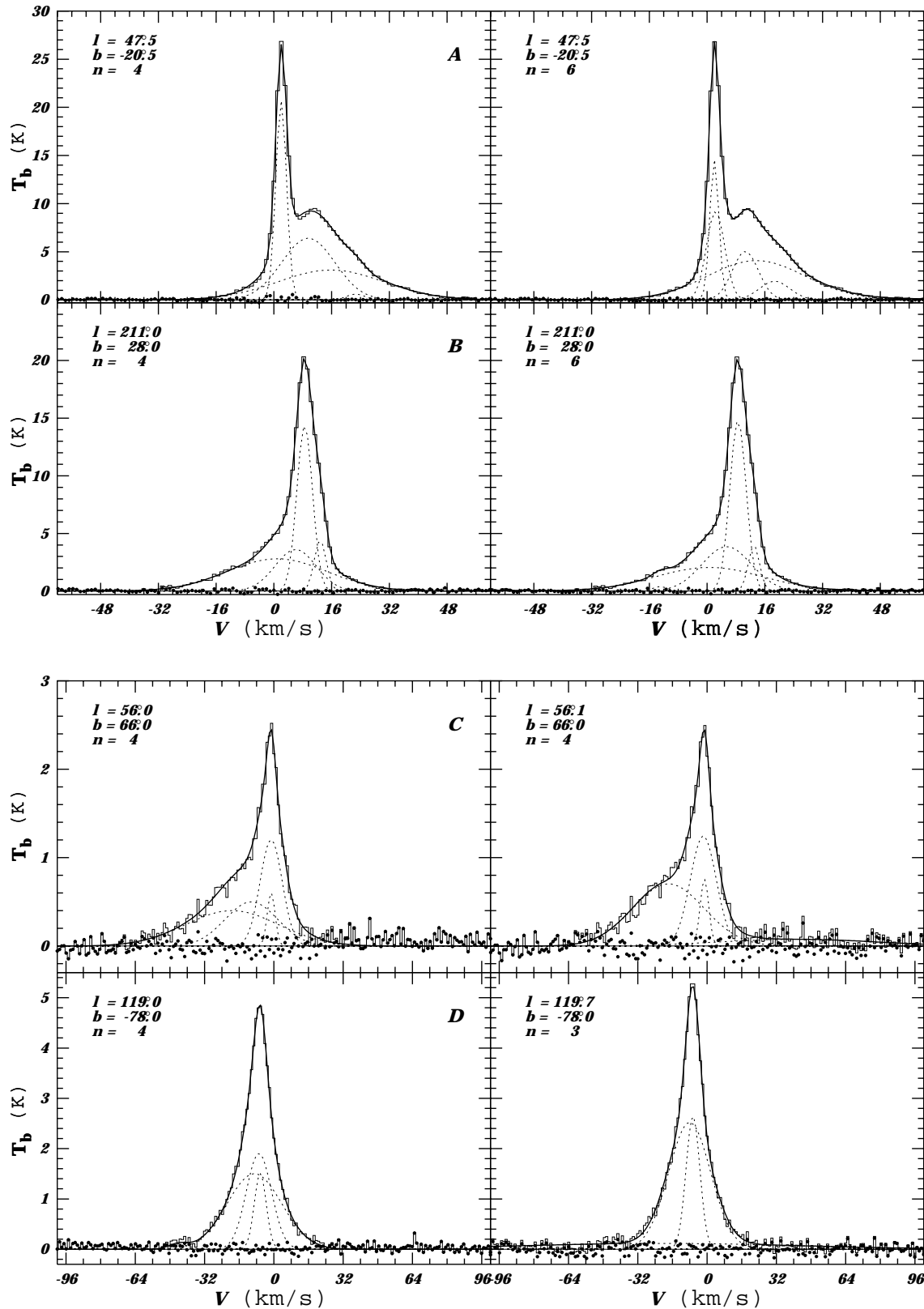


Fig. 8a–d. The comparison of the Gaussian decompositions, obtained by VP (left panels) and by us (right panels). The observations are given by the thin stepped line, the Gaussian model with the smooth thicker line, individual components with dashed lines and residuals are represented by stars. n equals to the number of Gaussians used.

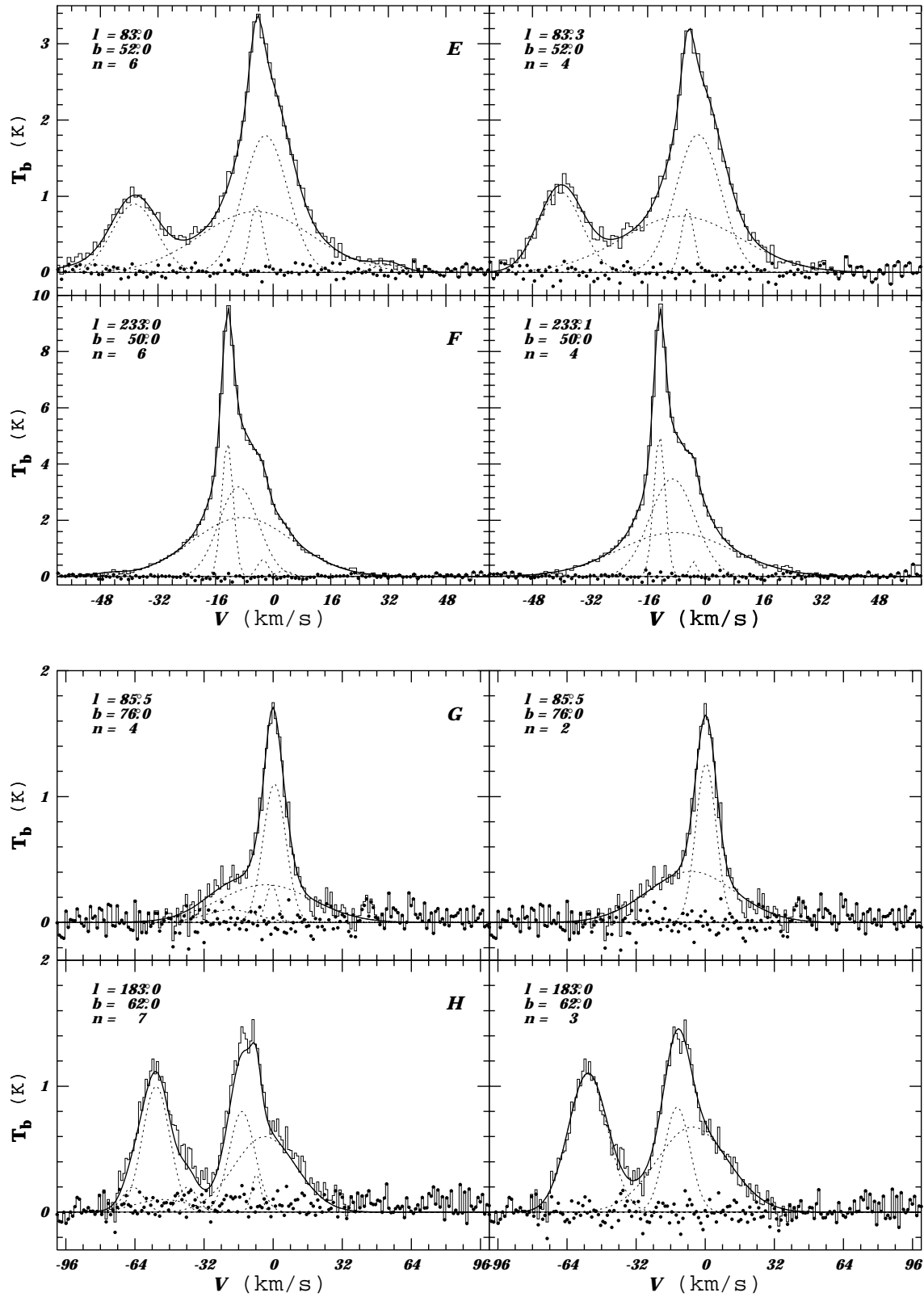


Fig. 8e–h.

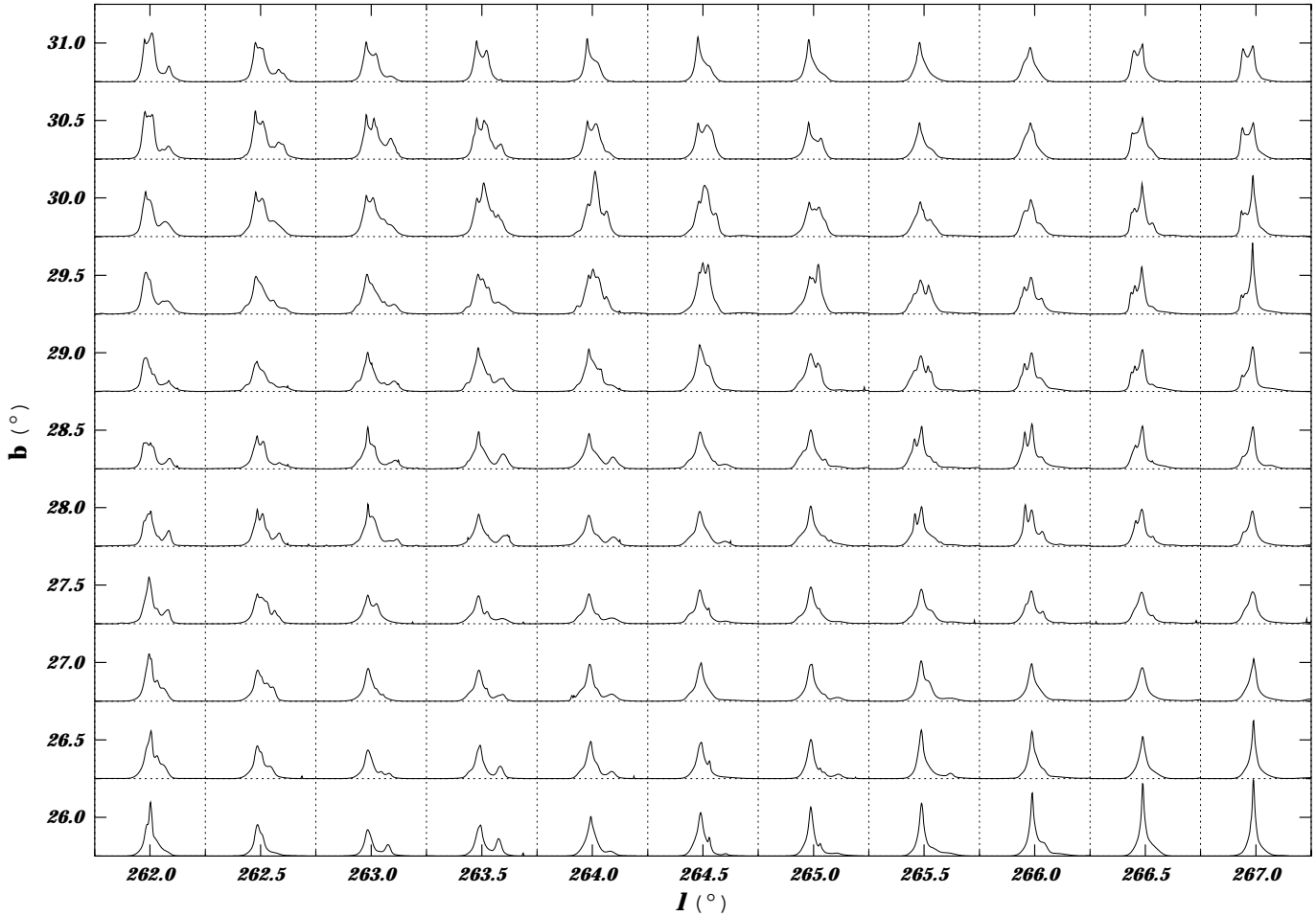


Fig. 9. Hydrogen profiles in the test field.

sky positions out of 138 832). Therefore, for stability analysis we decided to choose the profile from the ones, originally decomposed with 6 components. For each such profile we computed the similarities (as defined by Eq. 10) of all 15 possible combinations of the Gaussians in the given decomposition. In this way we filled in a table of 20004 rows by 15 columns, where in each row we sorted the entries into ascending order of similarities. After this we found the median for each column in this table and selected from the survey a profile, for which all similarity values of the Gaussians were the closest to corresponding medians. In this way the profile at $l = 264^{\circ}5$, $b = 28^{\circ}5$ was selected.

Around this position we took a 5 by 5 degree field (Fig. 9), whose size was chosen on the basis of the fact that often the decomposition program returns to profiles, already decomposed earlier, but these returns usually did not exceed 5 steps (2^5) backwards. Next we generated from the original decompositions of the profiles in this field the synthetic observed profiles by adding to the sum of Gaussians the random noise with parameters, characteristic to the original L/D Survey. In this way we obtained 10000 independent synthetic versions of our test field, adding in all cases a different random noise to all profiles in the field.

When decomposing the obtained test data, to estimate the influence of the predetermined order (starting at $b = -90^{\circ}$ towards $b = +90^{\circ}$) of decomposing the profiles in the survey, we also made all possible 90° rotations of the sky coordinates of the field to force the decomposition program to go through the central position in different possible directions. In this way we obtained 10000 independent decompositions of the profile, corresponding to the center of the field. All these decompositions must describe the same underlying signal spectrum, but may be otherwise as different from each other as permitted by the observational noise level of the L/D Survey and uncertainties in the decomposition process.

The results are characterized in Fig. 10. As we can see, in most cases the original Gaussian structure of the test profile is very well reproduced (as the contour intervals are logarithmic, the concentration around the original parameters is actually very high), but the lowest level contours demonstrate also that sometimes there are considerable deviations from the initial situation. It is also clear from these distributions that the weak and very broad components like D, are relatively ill-defined. The comparison of panels of Fig. 10 demonstrates that the differences in the order of the decomposition of the profile do not introduce radical differences in the results, but there are certain

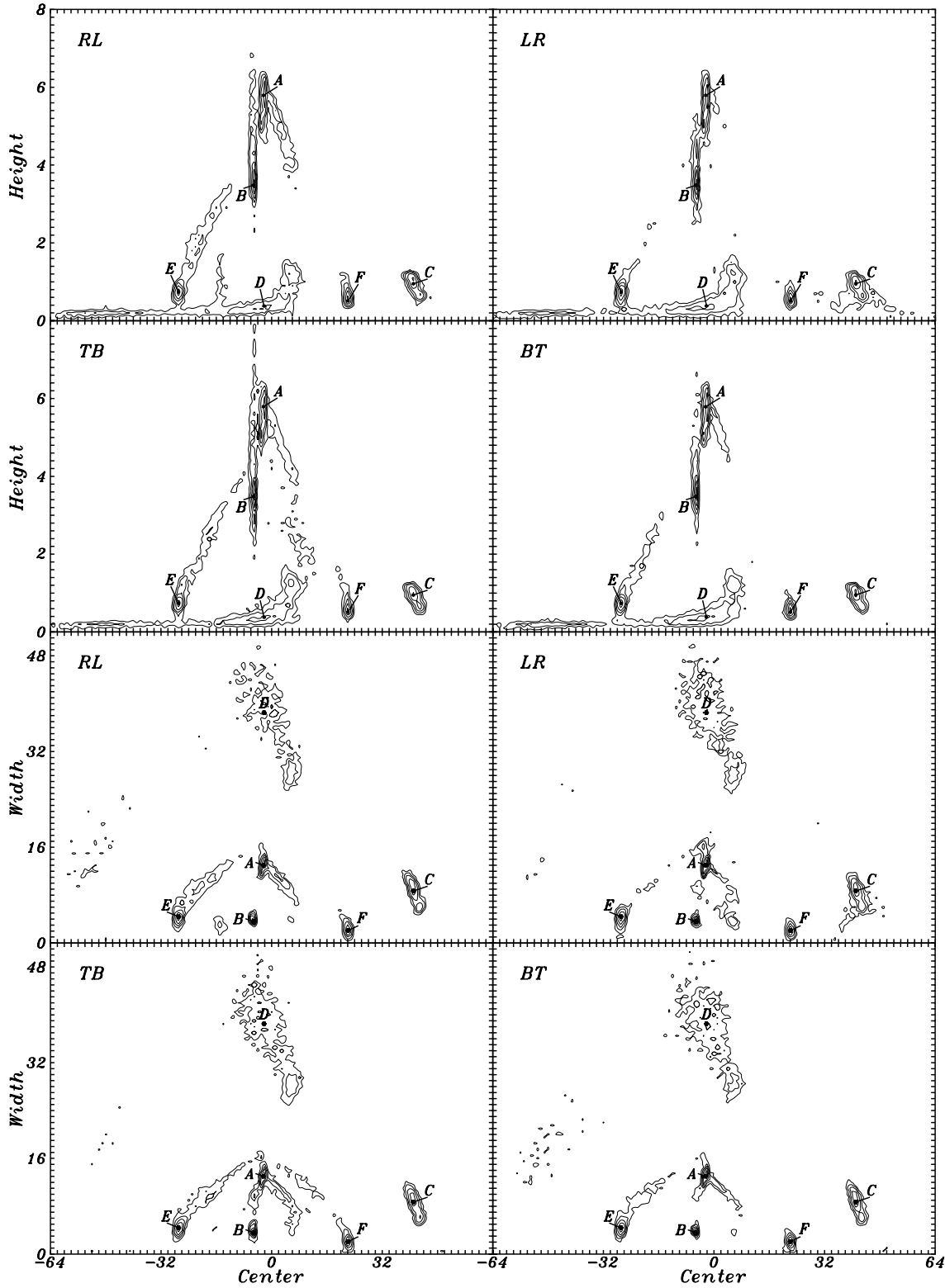


Fig. 10. The density distribution of the parameters of the Gaussian components in central velocity – Gaussian height (upper quadruplet) and central velocity – Gaussian width (lower quadruplet) planes. The positions, corresponding to the parameters of the original decomposition, are marked by stars and labelled in the order of decreasing area under the Gaussians. The iso-density lines are drawn in logarithmic scale at levels $\lg(n + 1) = 0.5, 1.0, 1.5, \dots$. The panels, labeled with RL, LR, TB and BT, correspond to the decompositions of the test profile, obtained when moving through its position from right to left, left to right, top to bottom and bottom to top in Fig. 9, respectively.

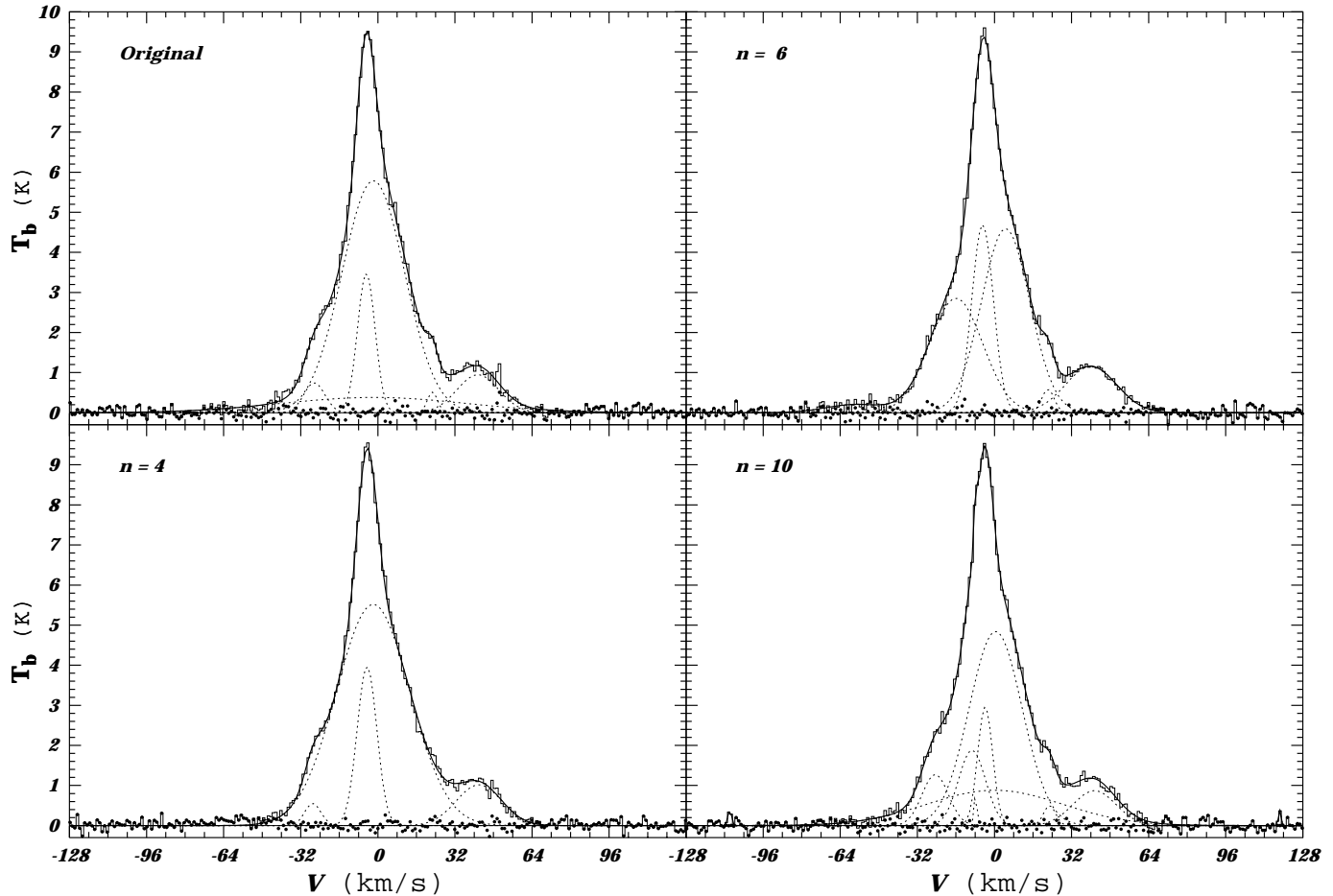


Fig. 11. Some extreme cases of the decomposition of the test profile together with the original one. n is the number of Gaussians used in decomposition.

differences in probabilities of obtaining the results from one or another family.

To understand better the behavior of the decomposition of the test profile, in Fig. 11 are given some extreme cases together with the original one. The upper right-hand panel illustrates the origin of the bridges in Fig. 10 from the component E towards A and from the component A towards F. These are obviously caused by the fact that sometimes the noise distribution smoothes out the small jerk on the left wing of the profile so that the component E becomes rather ill-defined. The rightward shift and growth of the component E causes also the rightward shift and decrease of the component A and increase of the component B. Mutual dependence of the parameters of the components A and B also explains the very broad range of possible values of their height in Fig. 10, but here we must consider that the heights of stronger Gaussians are less determined also due to the weights applied during the decomposition.

The two lower panels in Fig. 11 represent the best (left: the smallest residual rms among the solutions with the smallest number of Gaussians) and the worst (right: the greatest residual rms among the solutions with the largest number of Gaussians) decompositions of the test profile. The solution with 4 Gaussians has obviously arisen due to particular noise distribution, which

has smoothed out the small jerk on the right wing of the main peak of the profile, so that the component F has dropped out. Moreover, the strong negative noise peaks at about -50 km s^{-1} have rejected the component D. Out of 10000 test profiles 73 are decomposed with 4 components. The case with $n = 10$ is probably caused by a slight underestimate of the initial noise level rms_0 of the profile, so that the decomposition program has tried to fit with Gaussians even very weak features of the profile (components at -110 km s^{-1} and 118 km s^{-1}). Among test profiles only one is decomposed with 10 components.

5. Conclusions

The Gaussian decomposition of the complex HI line profiles is not an easy task and its results depend on many factors, including not only the original data. Therefore in recent years the Gaussian analysis has not belonged to the most popular tools for working with large hydrogen surveys. Nevertheless, this method has also some strong sides, and it is from time to time still used.

Several computer programs have been written to perform Gaussian decomposition of line profiles, but most, if not all of them are devoted to decomposition of a relatively small number of independent profiles. In this paper we described a new pro-

gram, which was specially written for decomposition of large hydrogen surveys in a self-consistent manner. While working out this program, the main attention was turned to keeping the number of resulting Gaussians as small as possible. To achieve this, we have analyzed the precision of the original profiles and introduced the weights for all channel values, based on the expected noise level in these channels. After preliminary decomposition of each profile we have applied special analysis to find additional possibilities for removing some less important components from the final decomposition. To reduce the ambiguities in choosing between different, but nearly equally acceptable solutions, we have used the assumption that in survey observations the profiles from neighboring sky positions must share some common properties.

As a result, we have managed to create a fully automatic computer program, which permits in reasonable time (using about 1.03 s per L/D Survey profile, or 0.13 s per every fitted Gaussian on 233 MHz PII PC) to decompose large hydrogen surveys into Gaussian components. The comparison of our results with the small sample, published by Verschuur & Peratt (1999), indicates that on average the program needs for decomposition even less Gaussians than introduced by an experienced human researcher. At the same time, we cannot argue that the solutions, obtained by our program, are absolutely the best ones. Often the detectability of the features in profile critically depends on the amount of noise in some channels. The obtained results depend also on the exact decomposition history of the survey profiles, but this dependence seems to be less important than the uncertainties, introduced by the random noise. Therefore, we believe that the values of Gaussian parameters, obtained by our program, may have somewhat higher weight than the ones from many earlier decomposition algorithms.

As a by-product of the tests of the new Gaussian decomposition program, we have determined the mean noise level of the original L/D Survey, which is somewhat higher than the value given by the observers for the published data. In Table 3 of Atlas (Hartmann & Burton 1997) the authors state that for the final data cube of the L/D Survey the mean spectral noise level $\langle\sigma_{\text{rms}}\rangle = 0.070$ K. When testing different possible ways of prescribing the weights to profile channels, we have determined the mean noise level of the profiles in signal-free regions by 4 more or less independent methods and have obtained the results $\langle rms_0 \rangle = 0.0882, 0.0893, 0.0907, 0.0909$ K, all pointing to the mean noise level close to $\langle rms_0 \rangle = 0.09$ K. A similar result $\langle\sigma_{\text{rms}}\rangle = 0.090 \pm 0.008$ K was earlier obtained also by G. Westphalen (1997). Most likely this difference is caused by the effective smoothing of profiles, performed when re-gridding and averaging the original profiles for publication. Moreover, about 1% of the published profiles are smoothed even more considerably in the process of removal of the strong sinc-pattern interferences (Fig. 7).

Acknowledgements. We would like to thank Prof. W. B. Burton for providing the preliminary data from the L/D Survey for program testing prior the publication of the survey. Considerable part of the work on creating the decomposition program was done at the Radioastronomical Institute of Bonn University. The hospitality of the staff members of the Institute is greatly appreciated with special thanks to Prof. U. Mebold, Dr. P. M. W. Kalberla and L. Wennmacher. We would like to thank Dr. J. Pelt from Tartu Observatory for the help in clarifying some problems in mathematical statistics and the referee Dr. W. G. L. Pöppel for very clear and detailed comments on our manuscript. The project was supported by the Estonian Science Foundation grants no. 180 and no. 2627, EC grant ERB-CIPA-CT-93-0662 and DARA grant WE 2-50 OR 9203-ZA.

References

- Baker P.L., Burton W.B., 1979, *A&AS* 35, 129
 Burton W.B., 1966, *BAN* 18, 247
 Burton W.B., 1972, *A&A* 19, 51
 Burton W.B., 1992, In: Pfenninger D., Bartholdi P. (eds.) *The Galactic Interstellar Medium*. Saas-Fee Advanced Course 21, Springer-Verlag, p. 1
 Cappa de Nicolau C.E., Pöppel W.G.L., 1986, *A&A* 164, 274
 Ewen H.I., Purcell E.M., 1951, *Nat* 168, 350
 Hartmann L., 1994, Ph.D. Thesis, Leiden Univ.
 Hartmann L., Burton W.B., 1997, *Atlas of Galactic Neutral Hydrogen*. Cambridge Univ. Press, 10+236 pp
 Heeschen D.S., 1954, Ph.D.-Thesis, Harvard Univ.
 Kalberla P.M.W., Westphalen G., Mebold U., Hartmann L., Burton W.B., 1998, *A&A* 332, L61
 Kaper H.G., 1959, Report TW1, Math. Inst. Groningen Univ.
 Kaper H.G., Smits D.W., Schwarz U., Takakubo K., van Woerden H., 1966, *BAN* 18, 465
 Kerr F.J., 1968, In: Middlehurst B.M., Aller L.A. (eds.) *Stars and Stellar Systems VII. Nebulae and Interstellar Matter*. Univ. Chicago Press, p. 575
 Kraus J.D., 1966, *Radio Astronomy*. McGraw-Hill, New York
 Lindblad O., 1966, *BANS* 1, 177
 Little R.J.A., Rubin D.B., 1987, *Statistical Analysis with Missing Data*. John Wiley & Sons, 16+265 pp.
 Matthews T.A., 1956, Ph.D.-Thesis, Harvard Univ.
 Pöppel W.G.L., Marronetti P., Benaglia P., 1994, *A&A* 287, 601
 Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., 1992, *Numerical Recipes in Fortran 77. The Art of Scientific Computing*. 2nd. Edition, 31+1486 pp.
 Savitzky A., Golay M.J.E., 1964, *Analytical Chemistry* 36, 1627
 Schwarz U., 1968, *BAN* 19, 405
 Schwarz U., van Woerden H., 1962, *Sitzberg. Akad. Heidelberg* 1962/63, 107
 Shane W.W., 1971, Ph.D.-Thesis, Leiden Univ.
 Stark A.A., Gammie C.F., Wilson R.W., et al., 1992, *ApJS* 79, 77
 Takakubo K., van Woerden H., 1966, *BAN* 18, 488
 van Woerden H., 1962, Ph.D.-Thesis, Groningen Univ.
 Verschuur G.L., Peratt A.L., 1999, *AJ* 118, 1252
 Westphalen G., 1997, Ph.D.-Thesis, Bonn Univ.