

Package name: EPRSIM C – technical information

Program name: GHOSTmaker

Icon: 

OS environment: MS Windows (32 bit)

Windows application

Version & Release: 3.5.1

Purpose:

- Condensation of multirun HEO solutions into GHOSTs

Technical documentation – GHOSTmaker - Table of content

| | | |
|-----|---|----|
| 1 | GHOST concept..... | 1 |
| 1.1 | GHOSTs – 5-dimensional cross-sections | 2 |
| 1.2 | GHOST condensation algorithm..... | 3 |
| 2 | GHOSTmaker usage | 10 |
| 2.1 | Input files – POP and MTP files | 11 |
| 2.2 | Output files – WMF and POP files | 11 |
| 2.3 | Defining relative uncertainties..... | 12 |
| 2.4 | Defining threshold densities by density minimum and density level | 12 |

1 GHOST concept

The concept of GHOST is built on the multidimensional presentation and construction of quasi-continuous distributions from multirun HEO optimization.

Multidimensional presentation: Many component simulations deals with large number of spectral parameters – 6 to 12 parameters per component can result in 23 to 47 spectral parameters per spectrum. It is impossible to present such a large solution at the same time. However, by grouping parameters into logical groups (same kind of spectral parameter together), one can use parameter cross-section space to present 2 such groups together. In addition, if one can code other parameters with shape or color of points one can construct several-dimensional cross-sections and in that way present many parameters at the same time.

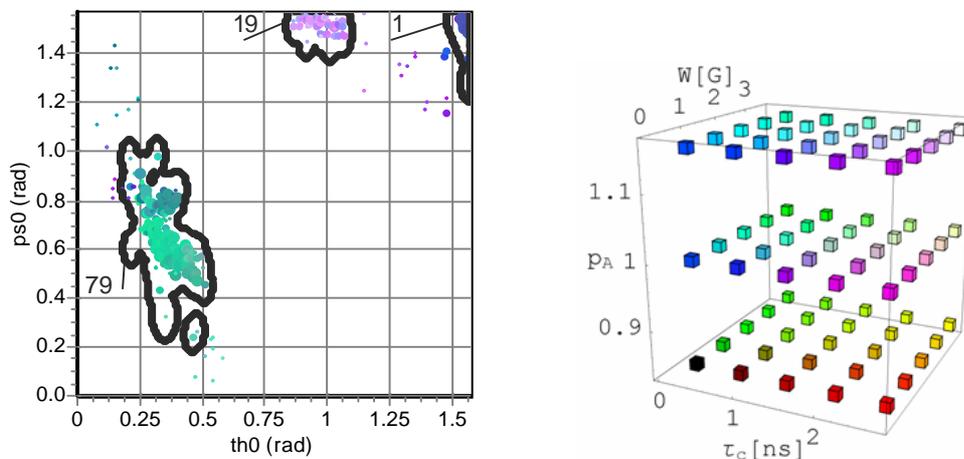
Construction of quasi-continuous distributions: If the proposed complexity is high enough, i.e. the experimental spectrum reports a low number of well-defined groups of solutions (spectral components), the proportion determined directly from the best fit of all the optimization runs seems to provide the most accurate values. If the proposed complexity is too low, i.e. the experimental spectrum includes a quasi-continuous distribution of solutions, the solution gathering can be thought in terms of projections. To get many projections HEO is run several times. In addition, modified HEO can keep solution diversity on much higher level than original HEO algorithm, enabling us to extract more than single solution from single HEO run. Typically, 10 solutions can be extracted from population of 400 solutions from single HEO run. By

special filtering and condensation these solutions can then be grouped for quasi-continuous distribution construction.

[Top of the Document](#)

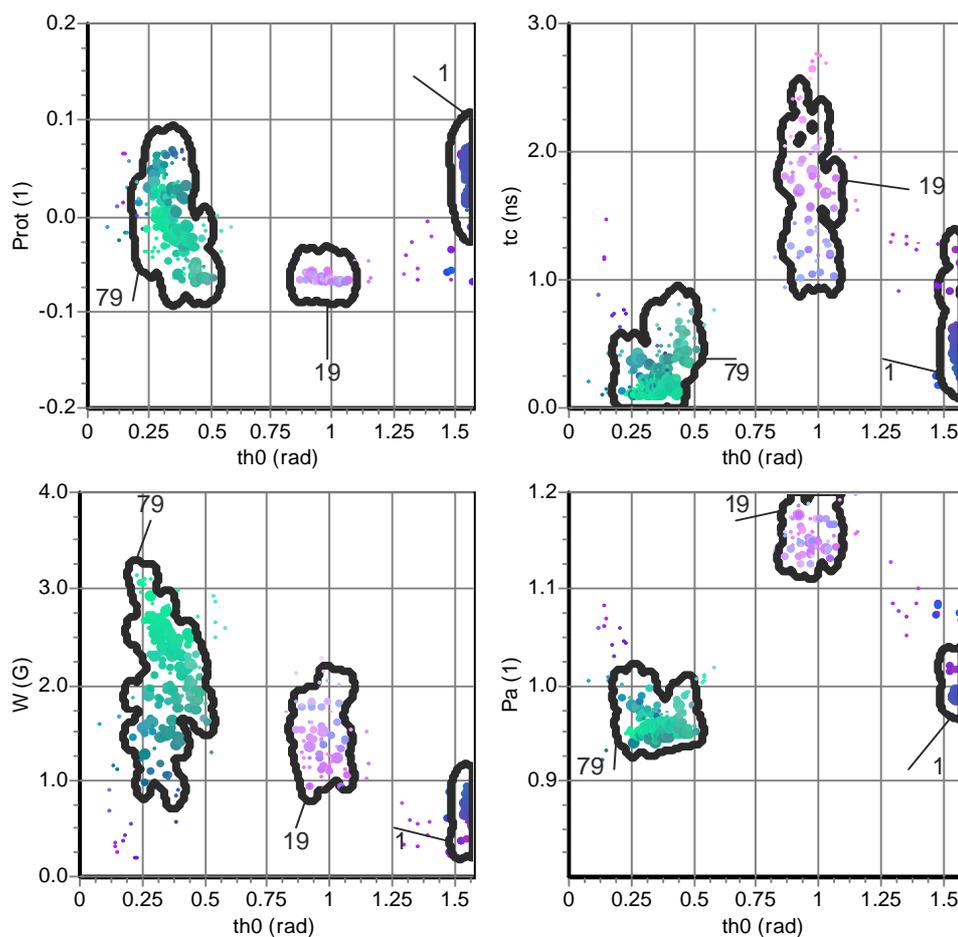
1.1 GHOSTs – 5-dimensional cross-sections

GHOST concept involves 2-dimensional cross-sections with axes representing two most important parameters (ϑ, φ) and RGB color of points that codes 3 additional spectral parameters (τ_c, W, p_A). In this way GHOSTs represent 5-dimensional cross-sections. Note that in this case color legend (cube) should be used to decode the last 3 parameters.



Since GHOSTs tries to represent many solutions or distribution of solutions through such a space, it involves also grouping/condensing these solutions into groups (ghosts) that can be characterized also by their spectral weight. In that way also 6th parameter is represented by GHOST representation (see weight indicator and group border on GHOST).

In addition to main GHOST one can represent any parameter on the second axis. In this way one can simplify determination of other spectral parameter. All the cross-sections use the same color representation enabling easier identification of particular group. Note, that τ_c , W , and p_A are defined both with position and its color code on their particular cross-section.



[Top of the Document](#)

1.2 GHOST condensation algorithm – implementation notes

The following general tools are implemented to make the algorithm able to detect, represent, distinguish and quantify the most important groups of solutions that can be extracted and statistically approved from spin label EPR spectra:

1. Data filtering according to the goodness of fit. Stochastic optimization routines that implement the usage of various random operators cannot guarantee that any fit (overall solution) found has the lowest values of χ^2 even if the proposed complexity of the model is high enough to enable a satisfactory description of an experimental response. On the other hand, if the proposed complexity is too low, the phase space (space of the solutions) in which the optimization will be performed, will inherently include many global minima for χ^2 . Therefore, the algorithm should act in a conservative way, i.e. it should damp the (worse) solutions with the highest χ^2 values and at the same time pass-through those points that belong to different global minima and that do not possess the lowest χ^2 due to the inappropriateness of the proposed model due to a too low complexity.

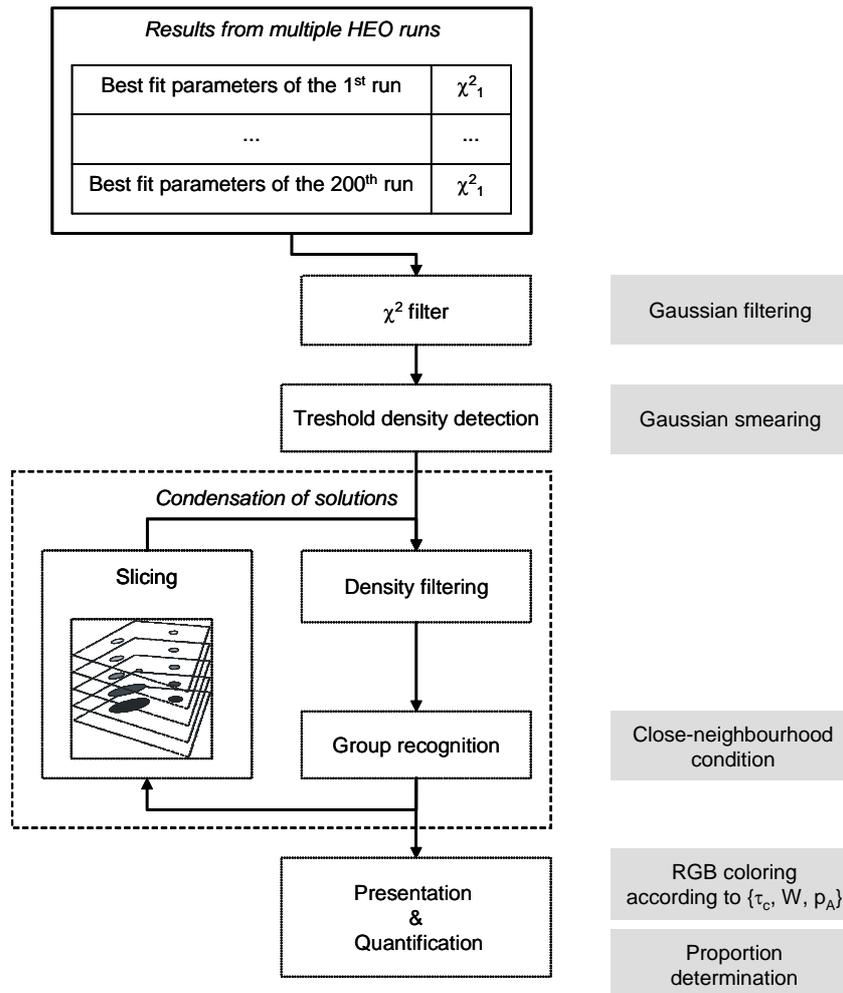
2. Data filtering according to local density. The method of determining the local density of the solutions is implemented to provide a second tool to classify the quality of the solutions. In contrast to χ^2 filtering, density filtering does not account for the property of a single solution (spectral component), but for a group of solutions, i.e. the

solution is significant only if it can be found with higher probability, so the local density must be high enough (many solutions appears in the neighborhood). An important property implemented in the density calculation is the continuity of the density function across the phase space. Instead of calculating densities within a grid, or at discrete sections of the phase space, the density at a certain point in the phase space is calculated by a summation of the contributions of the tails of points spread by a Gaussian probability function, corresponding to different, not perfectly defined solutions.

3. Grouping of solutions according to their neighborhood. In this part the solutions are grouped according to the neighborhood principle, defined in Equation (8) in the subsection *Condensation of solutions*. The grouping principle reduces the information coming out from the multiple runs of the optimization routine.

4. Quantification of the group proportions. The main goal of the GHOST condensation algorithm is to quantify the contributions of the groups of spectral parameters, which found to be a non-trivial task for the following reasons. If the proposed complexity is too high, i.e. the experimental spectrum reports a low number of well-defined groups of solutions (spectral components), the proportion determined directly from the best fit of all the optimization runs seems to provide the most accurate values. If the proposed complexity is too low, i.e. the experimental spectrum includes a quasi-continuous distribution of solutions, the extraction of the proportions from best fits that can be described as projections, is not appropriate anymore. To solve this problem, we propose that the proportions should be biased with the local solution density, i.e. the accumulated proportion of the appropriate group is obtained by summing the proportions of solutions divided by the local solution densities. In the case of a group with a high density, which one would obtain when finding a discrete group of closely spaced solutions, the calculated proportion will be close to the best-fit value. However, in the case of a low-density group, which one would obtain when finding a quasi-continuous group of distributed solutions, the proportion of the group will increase due to the accumulation of the many solutions that have not high local densities. In this way, this enables also the characterization of the quasi-continuous groups of spectral parameters. Note, that at the end the GHOSTmaker describes the solution distribution with up to 5 discrete solutions and send this description to EPRSIM BBW for spectral correction of the group proportion. However, one should switch this option of (“Automatic WC”), if the complexity of the solution is too high since then description with up to 5 discrete solutions is no longer valid.

The GHOST algorithm that includes the filtering, grouping and quantification is schematically presented in Figure below:



Firstly, the best-fit solutions (200 best-fits from 200 optimization runs of normal HEO algorithm or 20 runs of modified dHEO that preserves solution diversity) are collected and filtered by χ^2 filtering to get a representative ensemble. Then the local density is calculated at the position of each solution in order to discard about the least significant solutions. In the following slicing procedure grouping is performed at different slice densities according to the neighborhood condition. By slicing from the highest to the lowest densities the recognition and discrimination between discrete and quasi-continuous groups of solutions is enabled. Finally, the quantification as well as parameterization of each group of solution is done.

To present the resulting solutions and groups of solutions in a compact way, four different spectral parameters are used in the description of lipid domains in biomembranes: order parameter S , rotational correlation time τ_c , additional broadening W , and polarity correction factor p_A . This results in three different two-dimensional cross-section plots (S - τ_c , S - W , S - p_A) for which RGB coloring according to the normalized vector (τ_c, W, p_A) is used. In the case of the study of the low-resolution structure of membrane proteins, the order parameter is replaced by a cone angle \mathcal{G} and the angular amplitude of motion within a cone φ , respectively. In this situation an additional cross-section plot (\mathcal{G} - φ) can be presented.

Each phase of the GHOST algorithm is described in detail below:

χ^2 filtering

Data filtering is performed according to the χ^2 of the appropriate fit using a Gaussian damping factor, which works as a low-pass filter:

$$F_{\chi_i^2} = \text{Exp} \left[-\frac{1}{2} \left(\frac{\chi_i^2 - \chi_{\min}^2}{\Delta\chi^2} \right)^2 \right], \quad (1)$$

where χ_i^2 is the value for χ^2 for i -th HEO run, χ_{\min}^2 is the minimum of all runs, and $\Delta\chi^2$ the filter width. The filter width is determined such that the filter should pass through 40% of the better solutions. This number is empirically found and is based on the success rate of HEO applied on discrete problem, which was 80%, divided by 2 to make the procedure even more restrict.

Density calculation

The local solution density ρ_j at the position of the j -th point in the parameter space is calculated by the summation of contributions of tails of Gaussian-spread points, corresponding to different solutions positioned at the i -th points:

$$\rho_j = \sum_i \text{Exp} \left[-\frac{1}{2} \sum_k \left(\frac{p_{k,i} - p_{k,j}}{\sigma_k(d_i)} \right)^2 \right]. \quad (2)$$

The summation over k represents the summation over different parameters p_k weighted by the uncertainties $\sigma_k(d_i)$ calculated as

$$\sigma_k(d_i) = \sigma_k f(d_i), \quad (3)$$

where σ_k is the uncertainty for each parameter type k (a fraction of the appropriate definition intervals, as presented in EPRSIM library, fractions are defined in GHOSTmaker) and $f(d_i) = 1 + d_i \text{Exp}(-d_i^2/2)$ is an empirical weighting function, which tries to correct uncertainty according to the weighting factor d_i of the particular spectral component. The form of the function $f(d_i)$ implements the idea that spectral components with smaller contributions possess larger errors in the parameters, however spectral components with very small contributions do not represent statistically significant spectral components at all.

Density filtering

Density filtering is implemented in two ways: within the presentation of the χ^2 -equivalent solutions, as well as within the condensation of solutions.

In order to perform density filtering and to prevent rare spectral components to influence the final characterization, a threshold density was defined to damp the contribution of the solutions with local density lower than the threshold density. The threshold density is determined with the majority of the solutions that do not possess

high values of goodness of fit, e.g. 90% of all solutions (this is called *density level* in GHOSTmaker):

$$\frac{\sum_{i, \rho_i > \rho_t} F_{\chi_i^2} d_i}{\sum_i F_{\chi_i^2} d_i} = 90\%, \quad (4)$$

where $F_{\chi_i^2}$ is the χ^2 filter according to Equation (1), d_i is the contribution of i -th solution, and ρ_i is the local solution density according to Equation (2). The value of the threshold density is down-limited to the lowest possible threshold density ρ_t^{min} . This guarantees that at least a number, corresponding to ρ_t^{min} , of very good points (or equivalent number of points with a slightly lower χ^2) can be found in the neighborhood of each “significant” point. Usually $\rho_t^{min} = 5$ is used as the lowest possible average density in a single continuous distribution constructed with a predefined number of four spectral components and a number of HEO runs $M = 200$ resulting in 800 solution points.

Secondly, using the threshold density, the algorithm calculates the density filter for every point in an ensemble

$$F_{\rho_i} = \text{Exp} \left[-\frac{1}{2} \left(\frac{\rho_t}{\rho_i} \right)^2 \right] \quad (5)$$

to modify the radii for points corresponding to individual solutions in the GHOST diagram. The radius of a solution point is defined as the product of the contribution of the solution, its χ^2 -filter value and the ρ -filter value. Equation (5) indicates that if the local density is much higher than the threshold density $F_{\rho_i} \rightarrow 1$, the ρ -filter will effectively pass unperturbed solutions.

Condensation of solutions

After the solutions are damped with the χ^2 filter and solutions with a local density lower than the threshold density are removed, the remaining set of solutions is scanned for groups of solution points.

Determination of groups of similar solution points is based on the detection of neighboring points, i.e. two points i and j are close neighbors if the following condition is fulfilled:

$$\frac{1}{N_p} \sum_k \left(\frac{p_{k,i} - p_{k,j}}{\sigma_k(d_i)} \right)^2 < 1. \quad (6)$$

Here the summation over k represents the summation over different spectral parameters and N_p is the number of spectral parameters determining a single solution. Since the relative values of the parameters are divided with their relative uncertainties, this condition represents the condition when two solutions can be statistically indistinguishable.

To find the groups of solutions independently of their type, a special slicing procedure at different density values is implemented in the algorithm. Starting at the slice level s approaching the highest solution density ρ_i , the algorithm gives only groups of solutions that are found with high probability (high density). These are likely to be discrete-type groups of solutions, an example of which is group A in Figure below. Continuing with slices on lower densities, different groups of points are found. To confirm particular group, it has to be found on at least two subsequent slices, i.e. the center of mass of the group on the particular slice has to fulfill the neighboring condition with the center of mass of the group on the next slice. For example, both groups A and B are found on slice levels s and $s+1$ and the average parameters on the two slices fulfill the neighboring condition. So they are both confirmed on the slice level $s+1$. Groups found on the slice level $s+2$ enable reconfirmation of the group A in addition to the detection of new broad group. However, this broad group does not correspond to either group B (confirmed already) or group C (not confirmed), since the neighboring condition is not fulfilled neither for group B nor for group C. Because the group C was not found on at least two subsequent slices it was not confirmed; the solutions comprising group C in fact belong to broader group included in group D. On the other hand, the solutions comprising group B that was confirmed on the previous slice, represent a distinguishable part of group D. The later is considered as a typical case where group B would be regarded as a discrete group superimposed on a continuous group D. Finally, on the slice level $s+3$ group A and group D are confirmed. All the confirmed groups can be presented together on a single slice as a result of the slicing procedure.

Because much more solutions are needed to find continuous groups, the probability of finding them increases with decreasing the density. However, significant translations of center of mass of a continuous group between two subsequent slices can be encountered, causing a continuous group not to be confirmed at higher density slices but at the lowest one simply because such a group usually represents large fractions of all solutions. Note that a group is rejected after slicing if its contribution falls below a predefined value σ_d in Equation (3).

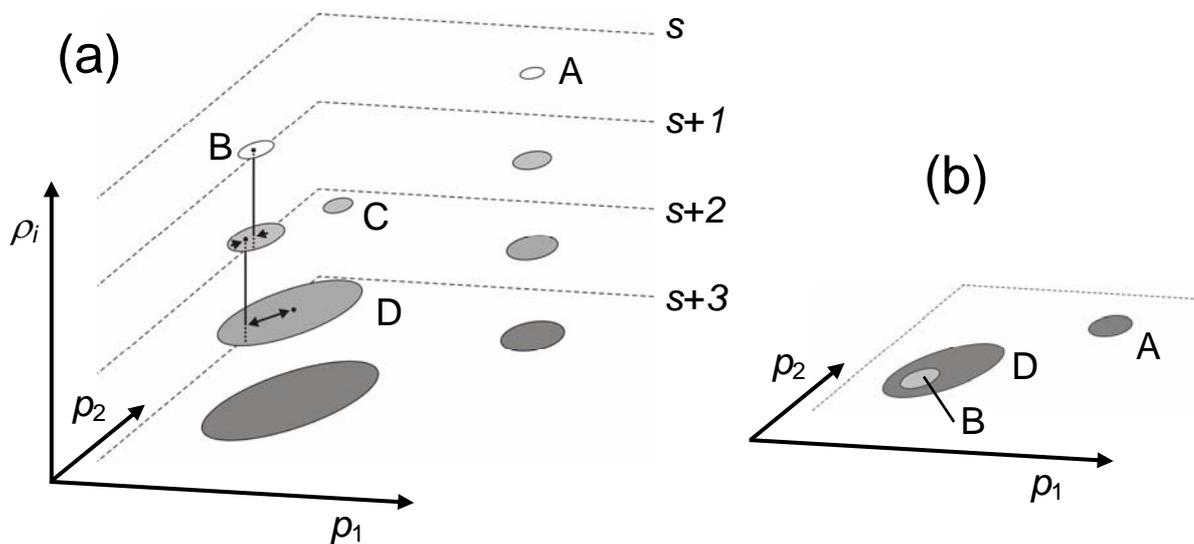


Figure - Slicing procedure. (a) Graphical presentation of detection procedure of discrete or quasi-continuous groups of solutions through slicing. Each slice ($s, s+1$, etc.) represents a two dimensional cross-section plot of spectral parameter p_1 and spectral parameter p_2 at particular local solution density ρ_i . Capital letters A–D represent different groups of solutions that are found through slicing. Center of mass of a group of solutions is indicated with a dot and application of neighboring condition test on the center of mass of the groups of solutions on the two subsequent slices is indicated with arrows. (b) Single slice, with all the confirmed groups (A, B, D), that represents the result of the GHOST condensation.

Proportion determination

After condensation of solutions into groups, the information (all parameters of all solutions) is further reduced into average properties of the groups (average parameters and relative contributions). Accordingly, the contribution $d^{(n)}$ of the n -th group is determined by the following equation:

$$d^{(n)} = \frac{\sum_{i \in \text{group } n} F \chi_i^2 \frac{d_i^{(n)}}{\rho_i}}{\sum_{m: \text{ all groups on the lowest slice}} \left(\sum_{i \in \text{group } m} F \chi_i^2 \frac{d_i^{(m)}}{\rho_i} \right)}. \quad (7)$$

Each other k -th parameter of the n -th group is averaged into $p_k^{(n)*}$ according to the following equation

$$p_k^{(n)*} = \frac{\sum_{i \in \text{group } n} F \chi_i^2 \frac{d_i^{(n)}}{\rho_i} p_{k,i}^{(n)}}{\sum_{i \in \text{group } n} F \chi_i^2 \frac{d_i^{(n)}}{\rho_i}}. \quad (8)$$

Here $p_{k,i}^{(n)}$ is the k -th parameter of the solution i of n -th group.

In order not to quantitative distinguish between different types of groups, discrete or continuous group, the second moment $M_2^{(n)}$ of n -th group of solutions is calculated according to the following equation

$$M_2^{(n)} = \sqrt{\frac{\sum_{i \in \text{group } n} F \chi_i^2 \frac{d_i^{(n)}}{\rho_i} \left(\frac{1}{N_p} \sum_k \left(\frac{p_k^{(n)*} - p_{k,i}^{(n)}}{\sigma_k} \right)^2 \right)}{\sum_{i \in \text{group } n} F \chi_i^2 \frac{d_i^{(n)}}{\rho_i}}}. \quad (9)$$

It measures the width of the distribution of spectral parameters in a group of spectral solutions. For discrete groups of spectral solutions it is expected to be below or close to 1, since the distribution of the EPR spectral parameters should be in the

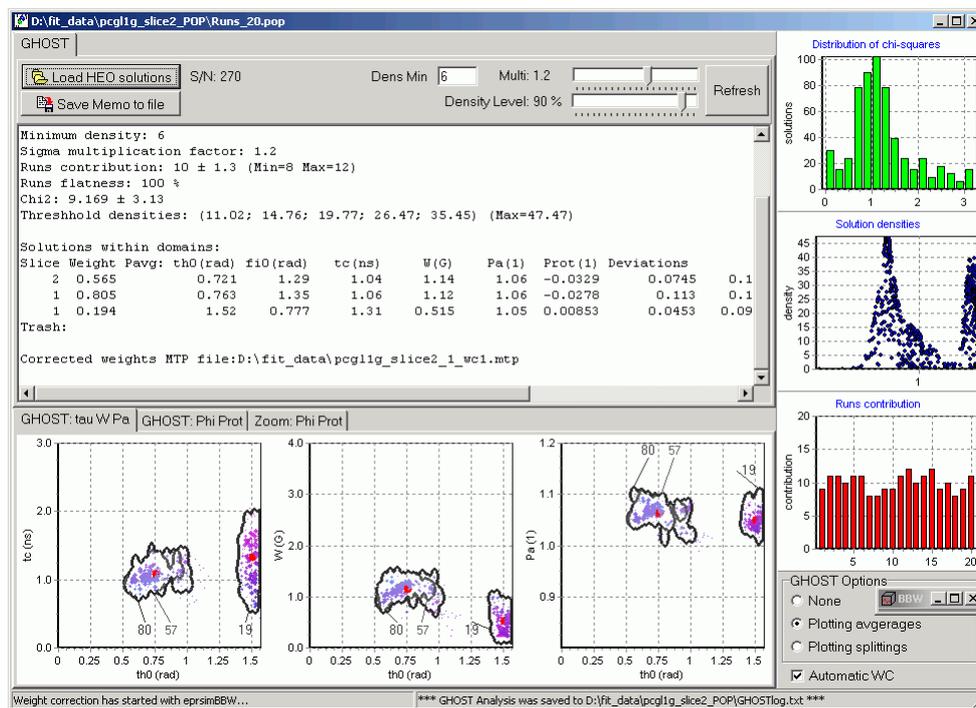
range of the errors of the spectral parameters. For large continuous distribution of parameters it will be significantly greater than 1.

[Top of the Document](#)

2 GHOSTmaker usage

To construct GHOST pattern, one must provide the solutions and define the uncertainties as well as density level.

To follow notes below see the GHOSTmaker application window below:



When using GHOSTmaker for GHOST creation please check the following general rules:

- Runs histogram (red) should be quasi uniform – Run contribution min should not be (much) lower than 8 and max (much) higher than 12; Run flatness should not be lower than 60%; if these quantities fall out of these intervals make additional runs or change HEO settings
- Hi2 histogram should have maximum around small numbers and decrease significantly for larger values; note that the units on ordinate axis represent the Hi2 values divided with Hi2 filter value;
- Solutions should be condensed into solid regions – borders should not encircled individual solution – if this happens reduce Density level;
- Vary Density level to see the effect on the solution condensation; the proper value of Density level is right below the value at which the regions' borders become smooth; the proper Density level should also be low enough that the minimum Treshold density is significantly above the noise density detected at the Solution density plot;

- Final condensed solutions (significant spectral groups that can be extracted from GHOST condensation procedure) are represented with its relative proportion, center of mass parameters and their second moment (representing the distribution widths in each dimension) and can be found in Memo box in the centre of GHOSTmaker application window.
- Weight correction is written in the Memo box additionally
- The following description files are produced automatically in the appropriate POP subdirectory:
 - "bestF200_XX.pop" file that includes the information about best 200 solutions from XX runs;
 - GHOST pictures: "runs_X.wmf" for the GHOST crosssections (X: 1 for τ_c , 2 for W, 3 for p_A , 4 for ϕ if appropriate and 5 for prot if appropriate);
 - "GHOSTlog.txt" that includes the Memo box content (all the GHOST analysis information);
 - "runs_XX_chi.txt" that includes the χ^2 histogram numbers from XX runs;
 - "runs_XX_runs.txt" that includes the runs histogram numbers from XX runs;

[Top of the Document](#)

2.1 Input files – POP and MTP files

If one uses basic HEO algorithm with 200 runs, the solutions are saved in MTP files. Load this file by choosing GHOST meni (in File Collection window) and "Berivse" submeni.

If one uses modified HEO algorithm with 20 runs and population files saved, the solutions are saved in POP files in subdirectories with name that corresponds to spectral filename. Load population data by choosing any of the 20 POP file in this directory through File menu and "Open Save N best" submenu.

[Top of the Document](#)

2.2 Output files – WMF and POP files

When GHOST are created within GHOSTmaker all 5 cross-section figures are saved as WMF files.

If MEM model is used in simulations "runs_1.wmf", "runs_2.wmf" and "runs_3.wmf" are created corresponding to ($S-\tau_c$, $S-W$, $S-p_A$) cross-sections, respectively.

If MES model is applied "runs_1.wmf", "runs_2.wmf", "runs_3.wmf", "runs_4.wmf" and "runs_5.wmf" files appears corresponding to ($\mathcal{G}-\tau_c$, $\mathcal{G}-W$, $\mathcal{G}-p_A$, $\mathcal{G}-\phi$, $\mathcal{G}-prot$) cross-sections, respectively.

In addition, GHOSTmaker creates "bestF200_20.pop" file with all the solution that were used to construct GHOSTs.

2.3 Defining relative uncertainties

As defined in density filtering subsection of previous section, one must use relative parameter uncertainties (called “sigma”-s in GHOSTmaker). These are:

- S-sigma (corresponds to σ_S , σ_g and σ_ϕ)
- Tau-sigma (corresponds to $\sigma_{\tau c}$)
- W-sigma (corresponds to σ_W)
- Pa-sigma (corresponds to σ_{pA} , and σ_{prot})
- d-sigma (corresponds to σ_d)
- Multi (corresponds to multiplication factor that multiplies all specific sigmas)

According to the numerical tests performed, the most appropriate values for nitroxide spectra characterization are:

- S-sigma : 2%
- Tau-sigma : 4%
- W-sigma : 3.5%
- Pa-sigma 3.5%
- d-sigma 2%

These numbers are stored in “ghostmaker.ini” file.

Value of multiplicator “multi” depends on Signal-to-noise value. The following empirical formula is used to define this value directly in GHOSTmaker (S/N value is readed from appropriate MTP file):

$$\sigma \cong 0.8 + 0.4 \frac{230}{S / N}$$

2.4 Defining threshold densities by density minimum and density level

Density filtering is obviously one of the most important tools of the GHOST condensation algorithm as it prevents rare solutions to affect the final characterization. Threshold density is used to define the width of the appropriate Gaussian filter. However, at least two important issues have an effect on threshold density value: signal-to-noise as well as complexity of the final solution. As the last is not known, the threshold density determination becomes one of the most difficult tasks for GHOSTmaker.

As described in GHOST condensation algorithm, a kind of overall solution proportion that are passed through density filter is used as a criteria in threshold density determination. In GHOSTmaker it is called *density level*. Usually it is set between 60 % and 90 % and represents the proportion of all solution weights that are passed

through threshold density filter. Higher slice densities are used in slicing procedure to determine and discriminate between groups of solutions.

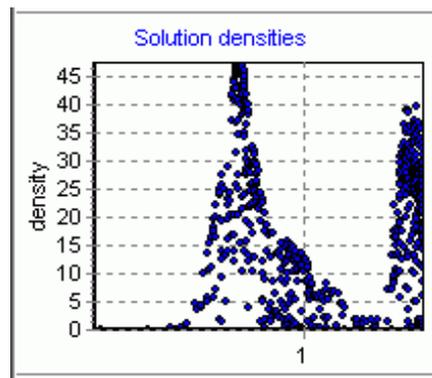
Use the Refresh button to repeat GHOST condensation analysis after you modify the Density level, Multi or DensMin.

Threshold density should in any case exceed minimum threshold density in order that GHOST condensation algorithm makes sense. To ensure that *density minimum* is defined according to the following empirical equation

$$\rho_{\min} \cong 1.7 + 4.3 \frac{S/N}{230}$$

To check whether threshold density determination was successful please check the following:

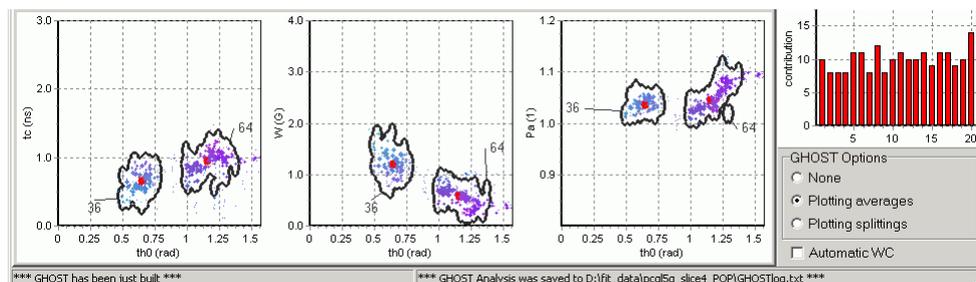
- neighbouring regions of solutions should be connected and not divided into several smaller regions
- threshold density should be slightly above the density noise (plateau-like background distribution on the graph showing density distribution – see blue Solution density graph on the figure below)



[Top of the Document](#)

2.5 Presenting the center of mass of groups or their splittings

To see the center of mass of each group of solutions select the “Plotting averages” in GHOST Options in the lower right corner of the GHOSTmaker application window.



To see how the group of solutions is spitted into 5 discrete solutions for weight correction analysis submitted to EPRSIM BBW, select the “Plotting splittings” in GHOST Options in the lower right corner of the GHOSTmaker application window.

