SIMFIT



1 The Simfit package

SIMFIT is a comprehensive free OpenSource package for simulation, model fitting, graphics and data analysis in Windows, Linux-Wine, and Mac-Crossover, available from the website at https://simfit.org.uk. Macros are provided to interface with Excel and results tables can be extracted from the log files in tab-separated, or html format for incorporation into word processors, or in LATEX format for technical documents.

SIMFIT can be used in statistical analysis (power as a function of sample size), epidemiology (survival analysis), biology (growth curves), pharmacology (dose response curves), pharmacy (pharmacokinetics), physiology (membrane transport), biochemistry (enzyme kinetics), biophysics (ligand binding), chemistry (chemical kinetics), and physics (dynamical systems).

The forty individual programs are run from a driver which provides access to the reference manual, the help program, tutorials with each program, and a mechanism for using test files to demonstrate every procedure. SIMFIT is installed by a setup program into a SIMFIT folder, then desktop short cuts to the driver w_simfit.exe in 32-bit operating systems, or x64_simfit.exe in 64-bit operating systems, can be made. Programs are selected, then test files are provided to demonstrate analysis of correctly formatted files before using your own data. You can scan the results (which are saved to log files), or you can browse current data, edit and re-run interactively, copy results to the clipboard, or create a results archive. SIMFIT is very easy to use for simple procedures like statistical testing, or elementary curve fitting, and there are tutorials to teach students how to use it to explore data, but it is the advanced features, like the extensive constrained nonlinear regression facilities, that will be greatly appreciated by more experienced analysts. All users, however, will benefit from the professional quality graph plotting procedures, which can create industry standard eps files and collages. The eps files can be used to create pdf, svg, png, jpg, tif, or bmp retrospectively, and there is a utility to edit eps files, or create inlays, or collages from them. Educators will appreciate the fact that, when students come to try a procedure for the first time, there is always a correctly formatted example data set to demonstrate that procedure.

1.1 Data format

SIMFIT data files are just ASCII text tables of values, and there are SIMFIT editors to format data. These editors have extensive features to facilitate data preparation, e.g. checking for nondecreasing order or nonnegative values; cutting/pasting rows, columns, blocks; plotting to identify errors; highlighting extreme signal/noise ratios; assigning weights for fitting; performing row/column arithmetic, including trig and hyperbolic functions, variance stabilizing transformations, etc. Data copied to the clipboard from other applications can be analyzed, and there are special macros supplied with the package to interface SIMFIT with MS Excel. SIMFIT will also accept data in spreadsheet export formats such as *csv*, *xml*, *html*, *mht*, *txt*, *prn*, or *mhtml*.

1.2 Curve fitting

There are programs for straightforward fitting, such as multi-linear, or polynomial regression, but also simple programs for nonlinear regression that automatically choose starting estimates, then select models on the basis of goodness of fit. However, there are more advanced programs for fitting general linear models, survival models, time series, differential equations, or sets of models, where users can control parameter limits and set starting estimates. Models can be used from a library or can be supplied by users, and there are many options for calibration and bioassay, as follows.

- □ Sums of exponentials (and estimation of AUC)
- □ Sums of Michaelis-Menten functions (and estimation of half saturation points and final asymptotes)
- □ Sums of High/Low affinity binding sites
- □ Cooperative order n saturation functions (and cooperativity analysis)
- Desitive n:n rational functions
- Growth curves (derivative plots, comparison of models and estimation of min/max growth rates, half-times and final sizes)
- □ Survival curves using several models
- Polynomials (all degrees up to 6 then statistical tests for the best model and prediction of x with confidence limits given y)
- □ Cubic splines (user-placed knots, automatic knots with variable tension, or crosss-validation data smoothing)
- □ Systems of differential equations (Adams, Gear, phase portrait, orbits) using defined starting estimates and limits or random cycles to search for a global minimum
- □ Calibration curves (polynomials, splines or user selected models)
- Area under curve (AUC by choice from several methods)
- Initial rates
- □ Lag times
- □ Horizontal and inclined asymptotes
- Numerical deconvolution of sums of exponentials, Michaelis-Mentens, trigonometric functions and Gaussian densities
- □ Fitting user supplied models
- □ Analyzing flow cytometry profiles

- □ After fitting functions of 1, 2 or 3 variables, parameters and objective functions can be stored for F, Akaike, Schwarz and Mallows Cp tests, and the wssq/ndof contours and 3D surface can be viewed as functions of any two chosen parameters. With all functions of 1 variable, calibration, evaluation, extrapolation, area calculations, derivative estimations and interactive error bar plots can be done.
- Multi-function mode: simultaneous fitting of several functions of the same independent variables, linked by common model parameters
- □ Generalized Linear Models (GLM) can be fitted interactively with either normal, binomial, Poisson or Gamma errors. Appropriate links can be either identity, power, square root, reciprocal, log, logistic, probit or complementary log-log with canonical links as defaults and facility to supply fixed offsets. A simplified interface is included for logistic, binary logistic or polynomial logistic regression, bioassy, log-lin contingency analysis or survival analysis.
- Stratified data sets can be analyzed by Cox regression and conditional logistic regression
- □ Autoregressive integrated moving average models (ARIMA) to time series with forecasting
- □ Facility to store parameter estimates and covariance matrices in order to compute Mahalanobis distances between fits of the same model to different data sets and test for significant differences in parameter estimates.

1.3 Statistics

All the usual descriptive statistics (bar charts, histograms, best-fit distributions on sample cdfs, dendrograms, box and whisker or cluster plots), multivariate statistics (distance matrices and dendrograms, principal components and scree plots), time series (ACF, PACF, ARIMA) and frequently used tests (mostly with exact p values not the normal approximations), such as:

- chi-square (O/E vectors, m by n contingency tables and wssq/ndof)
- McNemar test on n by n frequency tables
- Cochran Q test
- Fisher exact (2 by 2 contingency table) with all p values
- Fisher exact Poisson distribution test
- t (both equal and unequal variances, paired and unpaired)
- variance ratio
- F for model validation

- Bartlett and Levene tests for homogeneity of variance
- 1,2,3-way Anova (with automatic variance stabilizing transformations and nonparametric equivalents)
- Tukey post-ANOVA Q test
- Factorial ANOVA with marginal plots
- Repeated measures ANOVA with Helmert matrix of orthonormal contrasts, Mauchly sphericity test and Greenhouse-Geisser/Huyn-Feldt epsilon corrections
- MANOVA with Wilks lambda, Roy's largest root, Lawley-Hotelling trace, and Pillai trace for equality of mean vectors, Box's test for equality of covariance matrices, and profile analysis for repeated measurements.
- Canonical variates and correlations for group comparisons
- Mahalanobis distance estimation for allocation to groups using estimative or predictive Bayesian methods
- Cochran-Mantel-Haenszel 2x2xk contingency table Meta Analysis test
- Binomial test
- Sign test
- various Hotelling T-squared tests
- Goodness of fit and non parametric tests:
 - runs (all, conditional, up and down)
 - signs
 - Wilcoxon-Mann-Whitney U
 - Wilcoxon paired-samples signed-ranks
 - Kolmogorov-Smirnov 1 and 2 sample
 - Kruskal-Wallis
 - Friedman
 - Median, Mood and David tests and Kendall coefficient of concordance
 - Mallows Cp, Akaike AIC, Schwarz SC, Durbin-Watson
 - tables and plots of residuals, weighted residuals, deviance residuals, Anscombe residuals, leverages and studentized residuals as appropriate
 - half normal and normal residuals plots
- Multilinear regression by

- L_1 norm
- L_2 norm (weighted least squares)
- L_{∞} norm
- Robust regression (M-estimates)
- Also logistic, binary logistic, log-linear, orthogonal, reduced major axis, with interactive selection and transformation of variables in all cases.
- Partial Least squares (PLS)
- Survival analysis (Kaplan-Meier, ML-Weibull, Mantel-Haenszel) or, using generalized linear models with covariates (Exponential, Weibull, Extreme value, Cox)
- Correlation analysis (Pearson product moment, Kendal Tau, Spearman Rank) on all possible pairs of columns in a matrix, and canonical correlations when data columns fall naturally into two groups. Partial correlation coefficients can be calculated for data sets with more than two variables.
- Shapiro-Wilks normality test with the large sample correction
- Normal scores plots
- Binomial distribution, analysis of proportions, exact parameter confidence limits, likelihood ratio, odds, odds ratios and graphical tests such as log-odds-ratios plots with exact confidence limits for systematic variation in binomial p values
- Trinomial distribution (and confidence contour plots)
- Parameter estimates, confidence limits and goodness of fit for:
 - uniform
 - normal
 - binomial
 - Poisson
 - exponential
 - gamma
 - beta
 - lognormal
 - Weibull distributions
- All possible pairwise comparisons between columns of data by KS-2, MWU and unpaired t tests using the Bonferroni principle.
- Distance matrices for use in cluster analysis with extensive choice of pre-conditioning transformations and alternative link functions, e.g. Canberra dissimilarity and Bray-Curtis similarity.

- Nearest neighbors from a distance matrix
- · Classical-metric and non-metric scaling of distance matrices
- Principal components with eigenvalues, scree diagrams, loadings and scores from multivariate data sets
- Procrustes analysis to estimate the similarity between two matrices
- Varimax or Quartimax rotation of a loading matrix
- Canonical variates with eigenvalues, scree diagrams, loadings and scores from multivariate data sets. Group means can be plotted with confidence regions to assign comparison data to existing groups.
- K-means cluster analysis with plots
- Number needed to treat (NNT) and false discovery rate (FDR)

1.4 Calculations

Options are provided for the sort of calculations that are most often required inn data analysis, for instance singular value decomposition to find the rank of a matrix, or eigenvalue estimation.

- O Zeros of polynomials
- O Zeros of a user-defined function
- O Zeros of n nonlinear functions in n variables
- O Integrals of n user-defined functions in m variables
- O Convolution integrals
- O Bound-constrained quasi-Newton optimization
- O Eigenvalues
- O Determinants
- O Inverses
- O Singular value decomposition with right and left singular vectors
- O LU factorisation as A = PLU with matrix 1 and infinity norms and corresponding condition numbers
- O QR factorisation as in A = QR
- O Cholesky factorisation as in $Q = RR^T$

- O Matrix multiplication $C = AB, A^TB, AB^T$ or A^TB^T
- O Evaluation of quadratic forms $x^T A x$ or $x^T A^{-1} x$
- O Solve full-rank matrix equations Ax = b
- O Solve over-determined linear systems Ax = b in the L_1 , L_2 or L_{∞} norms
- O Solve the symmetric eigenvalue problem $(A \lambda B)x = 0$
- O Areas, derivatives and arc lengths of user supplied functions
- Analysis of cooperative ligand binding (zeros of binding polynomial, Hessian, minmax Hill slope, transformed binding constants, cooperativity indices, plotting species fractions)
- Power and sample size calculations for statistical tests used in clinical trials, including plotting power as a function of sample size (1 or 2 binomial proportions; 1, 2 or k normal ANOVA samples; 1 or 2 correlation coefficients; 1 or 2 variances; chi-square test)
- O Probabilities and cdf plots for the non-central t, non-central chi-square, non-central beta or non-central F distributions
- Estimation of exact parameter confidence limits for the binomial, normal, Poisson, etc. distributions and plotting confidence contours for the trinomial distribution.
- O Robust calculation of location parameter with confidence limits for one sample (median, trimmed and winsorized means, Hodges-Lehmann estimate, etc.).
- Time series smoothing by moving averages, running medians, Hanning or the 4253Htwice smoother
- Time series, sample autocorrelation functions and partial autocorrelation functions and plots for chosen numbers of lags and associated test statistics
- O Auto- and cross-correlation matrices for two time series
- O Shannon, Brilloin, Pielou, and Simpson diversity indices.
- O Kernel density estimation

1.5 Simulation

Models can be evaluated for plotting, and exact data can be simulated and random error added to mimic experimental situations, as now listed.

• Generating exact data from a library of models or from user-defined models

• User defined models can be multiple equations in several variables or sets of nonlinear differential equations, and can have conditional logical branching (equivalent to if...elseif...else) for models that swap at critical values of subsidiary functions or independent variables

• Logical operators like IF, IFNOT, AND, OR, NOT, XOR, etc. can be used to control model execution

• Special functions that can be used within models include

- Airy functions Ai(x), Bi(x) and derivatives
- Hyperbolic and inverse hyperbolic functions $\cosh(x)$, $\sinh(x)$, $\tanh(x)$, $\operatorname{arc-} \cosh(x)$, $\operatorname{arcsinh}(x)$, $\operatorname{arctanh}(x)$
- Bessel functions J0, J1, Y0, Y1, I0, I1, K0, K1
- Normal integral phi(x), error function erf(x) and complements phic(x) and erfc(x), Dawson's integral
- Exponential, sine and cosine integrals, E1(x), Ei(x), Si(x), Ci(x)
- Fresnel, Spence, Debye, Fermi-Dirac, Abramovitz, Clausen integrals
- Kelvin *bei*(*x*), *ber*(*x*), *kei*(*x*) and *ker*(*x*) functions
- Elliptic integrals RC, RF, RD, RJ and Jacobi functions *sn*(*u*, *m*), *cn*(*u*, *m*) and *dn*(*u*, *m*)
- Binomial coefficients, gamma function, incomplete gamma function, log(gamma(x)), digamma (*psi*(*x*)), and trigamma functions
- Struve confluent hypergeometric functions
- Legendre polynomials of degree n and order m, spherical harmonics
- Cumulative distribution functions for the normal, t, F, chi-square, beta and gamma distributions
- Inverse functions for the normal, t, F, chi-square, beta and gamma distributions
- Impulse functions: Heaviside, Kronecker delta, Dirac delta, triangular spike, Gaussian
- Periodic wave impulse functions: square, rectified triangle, Morse dot, sawtooth, rectified sine, rectified half sine, unit impulse
- One-line commands can be used for vector arithmetic (initialisation, norms, dot products), for evaluating polynomials or Chebyshev expansions, or for calling for mathematical constants (like Euler's gamma)
- Submodels can be defined, and these can be called from a main model for function evaluation, root finding, or adaptive quadrature or they can be called dynamically with arbitrary arguments
- A one line command is all that is necessary to estimate a convolution integral for any two sub-models over any range

• Adding random error to exact data sets to simulate experimental error

• Systems of differential equations After simulating, selected orbits can be stored and, with autonomous systems, vector field phase portraits can be plotted to identify singularities.

• Generating random numbers and 1,2,3-D random walks from the:

- uniform
- normal
- chi-square
- F
- logistic
- Weibull
- Cauchy
- Poisson
- binomial distributions.
- Generating n by m normally distributed random matrices.
- Generating random permutations of lists and Latin squares.

1.6 Documentation

Help documents are available in several formats as follows:

- ASCII text files can be read by any text editor (e.g. Notepad), or using a $SIMF_{I}T$ viewer, which is safer as it does not allow editing.
- HTML documents can be viewed by a built in HTML browser.
- PostScript documents (.ps) can be read using GSview or transformed into Portable Document Format.
- Portable Document Format (.pdf) files can be read and printed using Adobe Acrobat reader.
- Binary files that can only be read using a built in interpreter.
- A short HTML program can be viewed from the main SIMFIT program manager, w_simfit.exe.
- Each program has a dedicated self contained tutorial.
- Many of the specialized controls have individual tutorials.
- There is an extensive set of readme files (w_readme.0 gives details) which describe advanced features and technical details.

- This document (simfit_summary.pdf) summarizes the package and contains collages.
- A document (MS_office.pdf) describes the interface to MS Office and, in particular, explains how to use the macros to extract data from MS Excel spreadsheets.
- A document (PS_fonts.pdf) lists the PostScript font encodings, for those who want to create special effects.
- A detailed reference manual (w_manual.pdf) is provided in .ps and .pdf formats. The pdf version incorporates hyperlinks between the contents, index and page references and provides book marks.
- There is a full set of test files containing appropriate data to test every $SIMF_{I}T$ procedure.

1.7 Tutorials and worked examples

The SIMFIT website provides a set of some 220 short documents explaining how to use the individual programs. These are also collected together into an 800 page document called w_examples.pdf which can also be read from the website and is available from the SIMFIT main menu.

1.8 Graphical techniques

The primary hardcopy is encapsulated PostScript (.eps) but these can be used retrospectively to generate .pdf, .svg, .png, .jpg, .tif, or .bmp. The following procedures are available.

- Grouping of data into histograms (with error bars if appropriate) and cdfs
- Calculation of means and error bars with arbitrary confidence limits from replicates can be done interactively or from data files
- Error bars can be non-symmetrical or sloping if required and multiple non-orthogonal error bars can be plotted
- Error bars can be added to 2D and 3D bar charts and 3D cylinder plots
- Extrapolation of best-fit linear and nonlinear curves to arbitrary end points
- Automatic transformation of error bars into various axes (Hill, Lineweaver-Burk, Scatchard, log-odds, etc.)
- Immunoassay type dilution plots using logs to base 2, 3, 4, 5, 6, 7, 8, 9 as well as e and 10, and with labels as logs, powers of the base or fractions
- Multiple axes plots
- Pie charts with arbitrary displacements, fill-styles, colours

- Bar charts with arbitrary positions, sizes, fill-styles, colours and error-bars
- Presentation box and whisker plots, and pie or bar charts with 3D perspective effects
- Orbits and vector field diagrams for systems of differential equations
- Dendrograms and 3D-cluster plots for use in cluster analysis
- Scree diagrams and score or loading scatter plots for principal components analysis (score plots can have Hotelling T^2 elliptical confidence regions)
- ◆ 3D-surfaces and 2D-projections of contours
- Curves in space and projections onto planes
- vast array of plotting characters and maths symbols
- Standard PostScript fonts, Symbol, ZapfDingbats and Isolatin1 encoding
- Professional quality PostScript files that can easily be edited to change titles, legends, symbols, line-types and thicknesses, etc.
- Interactive PostScript facility for arbitrarily stretching and clipping overcrowded plots such as dendrograms without changing aspect ratios of fonts or plotting symbols but by just changing white-space between graphical objects
- A PostScript editor is supplied for scaling, rotating, shearing, translating, editing, making collages and inlays from .eps files
- Transformation of .eps files into scalable vector graphics format (svg) or compressed bitmap graphics formats (e.g. bmp, pcx, tif, jpg, png, pdf)
- Plotting user defined parameteric equations such as $r(\theta)$, x(t), y(t) in 2-space and x(t), y(t), z(t) in 3-space
- Facility to import PostScript specials automatically into the PostScript file creation stream in order to redefine fonts, colours, plotting symbols, add logos, etc.
- Graphical deconvolution of summation models after fitting
- ◆ 2D and 3D Biplots for multivariate data sets
- Metafiles can be archived then replayed to continue editing

2 The Simdem package

SIMD_EM is a set of three dynamic link libraries, available with 32-bit (cdecl or stdcall compatible) or 64-bit binaries, that can be called from any executables that support one or other of these conventions. It provides programmers with access to the SIMF_IT libraries on order to wrote Windows programs without having to know anything about the Windows API.

SIMD_EM was originally designed for Fortran programmers, but the DLLs can be linked to Windows executables compiled in any programming language. The package has 70 individual programs, each designed to demonstrate a particular GUI or graphical feature, and the source code is provided so that the calling sequences can be understood. The driver (simdem.exe or x64_simdem.exe) allows users to execute compiled examples then view the corresponding source code to understand the calling sequences. It can also run the SIMFIT PostScript editor. Executables linked to SIMD_EM have access to all the SIMFIT input-output, menu, tutorial, and graphical procedures.

SIMD $_E$ M is ideally suited for use with the NAG Fortran Builder and Silverfrost Plato IDEs and it is bundled with these compilers.

32-bit programs compiled using the S1MD_EM library must be distributed with the following run–time system

w_clearwin.dll
w_menus.dll
w_graphics.dll
salflibc.dll

while 64-bit versions require this run-time system

x64_clearwin.dll x64_menus.dll x64_graphics.dll salflibc64.dll clearwin64.dll.

3 Gallery

Here are examples of plotting styles, followed by collages constructed from SIMFIT .eps files. Zoom in to the graphs to appreciate that the resolution is only limited by the resolution of the display or plotting device.







Time t















Simfit Cylinder Plot with Error Bars



Phase Portrait for the Lotka-Voterra Equations



SIMFIT

Pie Chart Fill Styles











2.50 2.00 1.50 0.00 0.50 -1.25 0.00 y(2)

Orbits for a System of Differential Equations





15



SIMFIT Three Dimensional Scatter Diagram





Using CSAFIT to analyse Flow Cytometry Histograms



Three Dimensional Multivariate Biplot















Best Fit Epidemic Differential Equations



Twister Curve with Projections onto Planes











SIMFIT 3D plot for z = f(x,y)



Bivariate Normal Distribution $\mu_x = \mu_y = 0, \ \sigma_x = \sigma_y = 1, \ \rho = 0$



4 Collages









Simfit Cylinder Plot with Error Bars





0%

Diffusion From a Plane Source



Slanting and Multiple Error Bars



20