SSE user guide

2013-09-16 PsN 3.6.8

Introduction

SSE – Stochastic Simulation and Estimation – is a tool for model comparison and hypothesis testing. First, using a given model, henceforth called the input model, a number of simulated datasets are generated. Then the input model and a set of alternative models are fitted to the simulated data. Finally, a set of statistical measures are computed for the parameter estimates and objective function values of the various models.

Examples

```
sse moxonidine.mod -samples=1000 -alternative_models=alt1.mod,alt2.mod
```

```
sse pheno.mod -samples=500 -seed=12345 -threads=5
```

Input and options

<u>SSE-specific input</u>

A model file is required on the command-line.

-samples=N
The number of simulated datasets to generate is a required option. N must be greater than or equal to 2.

-alternative_models =alt1,alt2,alt3
A space-free comma-separated list of model files with alternative models is optional. If not given, only the input model will be fitted to the simulated data.

-no-estimate_simulation
By default, the simulated datasets are estimated with the input model. The OFV-values from these estimations are used as reference when computing statistics on results from estimations with the alternative models. By using the option -no-estimate_simulation (no option value), the input model will not be used to estimate (only to simulate). Instead, either a given reference OFV-value will be used (see option -ref_ofv) or the OFV-values from the first alternative model will be treated as the reference ones. Regardless of how option -no-estimate_simulation is set, the initial estimates in the simulation model will be used as reference when computing bias, rmse and rsebias for theta, omega and sigma. If simulating with uncertainty then the reference values will be different for each simulated dataset.

-ref_ofv=X
A reference OFV-value to use when computing statistics for the results from the alternative models. This option is only allowed together with -no-estimate_simulation.

-parallel_simulation=X
The number of parallel processes to start for the simulation step (not the estimation step) on a parallel computer. Default is the same setting as option -threads.

-add_models
This option tells PsN to add the alternative models listed with option -alternative_models to an old sse run. All models given via

option -alternative_models will be estimated from scratch, so the alternatives from the old run should not be listed in the input again. The -add_models option requires that the old sse run has been completed without errors. It is necessary to also set the general PsN option -directory to the sse directory of the old run, e.g. -directory=sse_dir50. The user must ensure that the -samples option in the new sse run is equal to or smaller than in the old run. If some simulated datasets are missing it will result in an error. The simulation model must still be given as input on the command-line. If the option -estimate_simulation is set (it is set by default) old estimation results will be reloaded if they exist, otherwise the simulation model will be estimated. The numbering of the extra models will start at the number of the last old alternative plus 1. Results for the added models are in file sse_results_add1.csv

-rawres_input=filename

A simple way to simulate with uncertainty. Instead of using identical parameter estimates for simulation of each new dataset, take parameter estimates from a raw_results.csv file, e.g. from a bootstrap run or the intial_estimates.csv file from a previous sse run with $PRIOR in the simulation model. The raw results file must be comma-separated and contain at least as many samples as the input -samples to sse, the labels for THETA/OMEGA/SIGMA in the file must match the labels in the simulation model given as input to sse, the theta columns must be directly followed by the omega columns which must be directly followed by the sigma columns, and the column header must be model either in the first column just as in a bootstrap raw_results file or in the second column as in a sse raw_results file. If a column header contains a comma, e.g. OMEGA(2,2), then that header must be enclosed in double quotes. This is done automatically in PsN raw results files. Note that is is possible to generate a file with initial parameter estimates outside of PsN, as long as the file follows the format rules.

-offset_rawres=N

Only relevant in combination with -rawres_input. Default 1. The number of result lines to skip in the input raw results file before starting to read final parameter estimates. In a regular bootstrap raw_results file, and also in an initial_estimates.csv file from an sse run, the first line of estimates refers to the input model with the full dataset, so therefore the default offset is 1.

-in_filter=<comma-separated list of conditions>

Only relevant in combination with -rawres_input. Default not used. The parameter estimates lines in the file can be filtered on values in the different columns. When specifying which column(s) the filtering should be based on, the exact column name must be used, e.g. minimization_successful. Filtering can only be based on columns with numeric values. The allowed relations are .gt. (greater than), .lt. (less than) and .eq. (equal to). If the value in the filter column is 'NA' then that parameter set will be skipped, regardless of the defined filter relation. Conditions are separated with commas. If the remaining number of lines after filtering is smaller than -samples, sse will stop with an error message. Then the user must either change the filtering rules or change -samples. If the user has created a file with parameter estimates outside of PsN, filtering can be done on any numeric column in that file. Do not set column headers containing .eq. or .lt. or .gt.in the user-generated file as this would interfere with the in_filter option syntax.

Example:
-in_filter=minimization_successful.eq.1,significant_digits.gt.3.5

-out_filter=<comma-separated list of conditions

Default not used. The user may choose to only compute results based on estimations which fulfill certain conditions. The default is to only skip runs where the ofv cannot be read from the lst-file or is equal to 0. Filtering of output can be done on any numeric column in a standard sse raw_results file, for example minimization_successful, significant_digits and covariance_step_successful. The allowed relations are .gt. (greater than), .lt. (less than) and .eq. (equal to). If the value in the filter column is 'NA' then that parameter set will be skipped, regardless of the defined filter relation. Conditions are separated with commas. If the remaining number of estimation results after filtering is less than 2, sse will stop with an error message.

Example:
-out_filter=minimization_successful.eq.1,significant_digits.gt.3.5

-recompute=<raw results filename, including directory name>

Default not set. Setting this option makes PsN recompute output statistics based on the specified raw_results file. Note that the filename must be given including the directory name. The user may change the -out_filter settings for the recomputation. Apart from -out_filter, the input model must be set, and -samples. Alternative models are not needed, information about them will be read from the raw results file. Option -directory will be ignored, instead the directory specified as part of the file path will be used.

Example: -recompute=sse_dir12/raw_results_run1.csv

Some general PsN-options which are useful in combination with sse

For a complete list of common options see common_options_defaults_versions.pdf, or psn_options -h on the commandline. The option -abort_on_fail (see common_options_..) will only affect the estimations of the simulated datasets. PsN always aborts if a simulation in sse fails.

-clean=X

To be able to reuse old simulation datasets and runs do not set clean to more than 2.

-directory=sse_dirX
The directory in which the script will run NONMEM can be named. The default name is “sse_dirX” where X is increased by 1 each time you run the script. If the run is aborted or crashes, setting the directory to the one from which the script was running earlier can be done. PsN will then not run the model files that had finished, saving time. Note that same set of options must be given as when the run was started the first time.

-seed=X
A seed for the random number generator can be specified. This makes the run reproducible. It acts as a starting point for the random number generator to produce the random number that will be present in the simulation control file.

-threads=X
The number of parallel processes to start for the estimation step (not the simulation) on a parallel computer.

-help
With -help sse will print a longer help message.

Special cases

To simulate odd-type data, relevant code (with a simulation block) is needed in the input model file, as well as a $SIMULATION row. It should contain one ordinary (from a normal distribution) seed number and one extra seed number (from a uniform distribution) “$SIMULATION (11111) (11111 UNIFORM)”. The random number generator handled by PsN will replace the seed numbers in a controlled manner (derived from “–seed=” if provided).

Please note that NONPARAMETRIC is not yet completely supported, since it requires additional results handling features.

Simulation with uncertainty

Simulation with uncertainty is supported using three methods. One method is to use options -rawres_input and -offset_rawres, see above. The second method is to define $PRIOR NWPRI in the input model, including the PLEV option required when simulating from a prior. Please note that the update_inits script has some functionality for automatically adding $PRIOR NWPRI to a model, based on output from an estimation. However the update_inits feature is still experimental and the generated model needs to be manually checked. The third method is to define $PRIOR TNPRI in the input model, including the PLEV option required when simulating from a prior. If no $SIM record is present in the input model containing $PRIOR, PsN will create one and add option TRUE=PRIOR. If PLEV option is not set in $PRIOR, NONMEM will halt saying that the value of PLEV is inappropriate. When $PRIOR is present in the input model, PsN will store the initial estimates sampled from the prior in the file initial_estimates.csv. This file is formatted so that it can be used with option -rawres_input in a later sse run. The first row, model 0, in initial_estimates.csv is the initial estimates of the input model, and these will be skipped if -offset_rawres is equal to the default value.

If $PRIOR TNPRI is set, then NONMEM requires $MSFI in the first $PROB. PsN will automatically copy the msf-file to the run directory, meaning that it is not necessary to set PsN option -extra_files for the msf-file.

Output

The output from each SSE run is collected in a new directory with a name of the form sse_dirX, where X is an integer. The most recent SSE run is found in the directory with the highest number. The only exception is when the user restarts a run from a specified existing directory using the -directory option, then the results will be saved in that directory.

When $PRIOR is present in the input model, PsN will store the initial estimates sampled from the prior in the file initial_estimates.csv, see Simulation with uncertainty above.

SSE creates two files summarizing the results that can be opened in e.g. Excel. The raw_results.csv file is a standard PsN file containing raw result data for each estimation run (termination status, parameter estimates, uncertainty estimates, etc…). The sse_results.csv file is a file specific to the SSE routine containing summary statistics and comparisons common to all estimation runs.

The header line for the reference/true values uses complete indexes for OMEGA and SIGMA, e.g OMEGA(2,1), SIGMA(3,3). The header lines for the input model and the alternative models use naïve numbering, for example if OMEGA(1,1), OMEGA(2,2), OMEGA(3,2) and OMEGA(3,3) are defined in the modelfile, they will in the sse output be numbered OM_1, OM2, OM_3 and OM_4. See also section Principles of statistical analyses.

Explanation for data items in sse_results.csv:

The 'true' values in the formulas below are always the initial parameter estimates of the simulation model. When simulating without uncertainty the true value is constant over i, but when simulating with uncertainty the true value varies with index i. If for any i the true value is equal to 0 in a computation of a relative value, then that sample is skipped and N is adjusted.

sd = standard deviation of estimated parameter x_i.

Standard error CI_{bias} *= parametric confidence intervals = mean bias ±* Z**rse where* $Z = [2.58, 1.96, 1.645]$ *for CI = [99%, 95%, 90%]*

Items under the OFV statistics heading by default refer to likelihood ratio test comparisons between the input model and the alternative models. The exceptions are described under section Principles of statistical analyses below.

Known problems

The simulated datasets are $TABLE output from NONMEM, and NONMEM formats and rounds off values when printing tables. This leads to two known problems.

1) In NONMEM6 1013201 is rounded to 1013200 (five significant digits), and if this makes a significant change to the model estimation, for example if the value is a covariate, then the sse results will be wrong. In NONMEM7 it is possible to set the FORMAT or RFORMAT option in $TABLE to make sure no important information is lost. With NONMEM6 the user must make sure the rounding to five significant digits does not harm the results. PsN cannot detect this problem.

2) A column of integers in the input dataset, e.g. OCC for occasion, will be formatted as a floating point number in table output. When the table output is used as input in the estimation step of sse, statements like IF (OCC .EQ. 1) will not work, because to NONMEM 1 is not equal to 1.000000e00. As a workaround the user must write code that will work both for integers and floating point numbers, for example (IF OCC .GT. 0 AND OCC .LT. 2). Alternatively, the user can set the RFORMAT option in the estimation models, see NONMEM help files.

Recovering a crashed/stopped sse

If the sse run is halted before the simulations are finished, it is possible to reuse the already finished runs. Run the same command as before, set option -directory if it was not set in the original call. Please note that PsN will print a message “Starting NNN NONMEM executions” that indicates that all simulations are rerun, but the check for reusable results is done after this message is printed. Some copying of results is done, and some existing result files will be overwritten with identical copies.

If the sse run is halted after the simulations are finished, PsN can reuse the results from both simulations and finished estimations. Run the same command again with -directory set. Do not set the rerun option. PsN will print that all NONMEM executions are started even if results are reused, see above.

Technical overview of algorithm

The program creates N (N= number of samples) simulation model files that are modified copies of the input model file given on the command line. Then the program creates N*M (M=number of alternative model files + 1 (for the input model file)) estimation model files. Each model file has one estimation copy per simulated dataset. N simulated datasets are generated by running NONMEM once for each simulation model file, and then each simulated dataset is used for parameter estimation M times, once with each model (simulation + alternatives). The estimated parameters and OFV values from the M*N runs are collected in raw_results.csv together with other NONMEM outputs, and statistical measures are written to sse_results.csv.

Important features of the N simulation model files generated from the input model:

- In $DATA record IGNORE and ACCEPT statements are kept intact.
- If the input model contains a $SIMULATION record, the record is kept intact, except that the

random seeds are set to different values in each copy, and it is made sure that TRUE=PRIOR is set if the input model has $PRIOR. If there is no existing $SIMULATION record, the program creates a new one according to $SIMULATION (<random seed>) ONLYSIMULATION and adds TRUE=PRIOR if the input model has $PRIOR.

- If any $ESTIMATION record contains LIKELIHOOD, -2LOGLIKELIHOOD, -2LLIKELIHOOD or LAPLACIAN, then ONLYSIMULATION NOPREDICTION is set in $SIMULATION.
- All $ESTIMATION records are removed.
- Any $COVARIANCE record is removed.
- Old $TABLE records are kept, and the filename in option FILE=<filename> will have trailing numberes removed and then be numbered -sim-i, where i is the order number of the simulation model file.
- In the newly generated $TABLE record for the simulated dataset the option FILE=<simdata> is set, where <simdata> is a unique name for each simulation model file.
- If option -rawres_input is set, then replace the THETA/OMEGA/SIGMA parameter values in the ith simulation model with the final estimates from the ith results line in the input raw_results file.

Important features of the N estimation model files generated from the input model and of the estimation model files generated from alternative model files (N files per alternative model):

- The N model files generated from the input model are numbered with the number of the simulated dataset used as input in the modelfile. The files generated from the alternative models are numbered with X-Y, where X is the order number of the alternative model and Y is the order number of the simulated dataset.
- In $DATA record, IGNORE is set to @ (i.e. ignore text lines), replacing any old IGNORE=<single character>. All old IGNORE=(list) and ACCEPT statements are kept. The user must be cautious regarding statistics if IGNORE or ACCEPT statements differ between models.
- Any $SIMULATION record is removed from the first $PROBLEM.
- All $ESTIMATION records are kept. If MSFO=<filename> is specified, <filename> will first have any trailing numbers removed, and then <filename> will have the same numbering appended as the name of the modelfile itself (see above).
- In $TABLE records of the first $PROBLEM, the filename in FILE=<filename> will have trailing numberes removed and then be numbered as the modelfile, see above.
- Input is set to <simdata> from one of the simulation model files.
- If an alternative model has a second $PROBLEM (this is not allowed for the input model unless the first $PROB has $PRIOR TNPRI), then it is assumed that the first $PROB has MSFO=<filename> in $ESTIMATION and that the second $PROBLEM has an $MSFI record. PsN will set the same numbered filename in MSFO and MSFI. It is also assumed that the second $PROBLEM has a $TABLE record. In the first $TABLE of the second $PROBLEM PsN will set FILE=simtabX-Y.dat, where X-Y is numbers as above. The filename in $DATA of the second $PROB will be set to the same <simdata> as in the first $PROBLEM, but IGNORE statements will not be changed. This means that the user must ensure that there is an IGNORE=@ in $DATA of the second $PROBLEM. PsN will not change the random seeds in $SIMULATION, but since the

simulated datasets used as input are different, the table output will also be different.

- If the input model is to be estimated and it has $PRIOR set, then option PLEV is removed from $PRIOR.

Principles of statistical analyses

All estimated parameters values; thetas, omegas and sigmas, are compared with the initial parameter values in the input model file, i.e. the values used for generating the simulated datasets. The matching is based on numbering, so the value of theta 3 used for simulation will be compared with all estimated theta 3. It is up to the user to ensure that the matching is correct, i.e. that the simulation value of CL is always compared with the estimated values of CL. If for example OMEGA(1,1), OMEGA(2,2), OMEGA(3,2) and OMEGA(3,3) are defined in the modelfile, they will in the sse output be numbered OM_1, OM2, OM_3 and OM_4, i.e. values are numbered naively without considering if they are on the diagonal or not. Same for SIGMA. If OMEGA/SIGMA is diagonal in the original model but of block form in an alternative model, the matching will be incorrect. To avoid that error, use block form also in the original model, setting off-diagonal elements to 0 (NONMEM will keep these fixed). If there are more parameters in an alternative model than in the simulation or vice versa, no matching can be done and the spaces for the comparison statistics will be empty. If parameters are different from a model to another one, numbering should be done considering that, or if not comparisons should not be interpreted.

SSE does not check whether minimization was successful or not. Statistical computations include also parameter estimates from NONMEM runs terminated with e.g. rounding errors, unless the option -out_filter is used, see above. SSE will skip all runs where the ofv cannot be read from the lst-file or is equal to 0. This means that e.g. the mean may be based on different number of runs (samples) for different alternative models. If the ofv is available but a parameter value is missing, then SSE will print a warning and compute the statistics without that value.

If a run neither minimizes nor terminates, but gets into an infinite loop, it won't be killed by PsN; so, in order for the sse results to be computed, a manual kill can be contemplated.

In sse_results.csv no individual parameter estimates are reported. Mean, median, sd and other measures are the means etc. over the simulated datasets.

The OFV values are treated differently. By default, the OFV value obtained when estimating parameters using the input model is used as the reference for each simulated dataset, and the OFV values obtained when estimating the alternative models are compared with this reference. There are two alternative methods. The first is to set the option -no-estimate_simulation without setting -ref_ofv. In this case the OFV-values from the first alternative model are used as references. The second alternative is to set -no-estimate_simulation together with -ref_ofv=X, and then X is used as the reference OFV value for all alternative models.