Exercises on model parameter identifiability

Markus Schartau and Andeas Oschlies

Autumn School: Data Assimilation in Biogeochemical Cycles, 21.-27.09.14, at Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy

General information– The following exercises deal with the specification of confidence- or credibility regions for the identification of uncertainties in model parameter estimates, as described in the lecture. Examples are based on a marine plankton ecosystem model with eleven state variables that describe mass flux of carbon (C) and nitrogen (N) (variable stoichiometry of organic C:N elemental ratio). The model is applied for simulations of a mesocosm experiment (duration 23 days), which included three enclosed water volumes (called mesocosms or bags) that were sampled in time. Data-model comparison will be done with daily averaged observables of the three mesocosms. We will keep initial conditions fixed for all model simulations. Environmental variables (temperature, salinity) data were hourly interpolated between measurements at noon. Photosynthetic available radiation (PAR) data resolve diurnal (night-day) changes, derived from irradiance measurements at noon and astronomical calculations of daylength.

What to expect – The major aim is to give insight to conditions that are realistic, to provide some practical guidance, and to foster (maybe) new ideas. The analysis of the identifiability of model parameter values must be seen as part of an (usually interative) approach to deriving representative posterior error distributions of model parameter estimates. Examples picked for the exercises are typical for marine- and aquatic plankton ecosystem models. We will only look at two-dimensional parameter spaces for reasons of better visualization. These 2D-visualizations will, however, be helpful when applying methods of optimization in a high-dimensional parameter space (e.g. methods that require an approximation and inversion of a covariance matrix of model parameters). For example, an analysis on parameter identifiability can help to justify the outsourcing of specific parameters from optimization; this way problems of inversion (nearly singular matrices) are reduced or eliminated. Furthermore, the conditions addressed in the exercises should be helpful to setup a successful application of MCMC methods that eventually provide posterior error distributions of a high-dimensional parameter space.

Literature – We find many descriptions in textbooks about the theory on defining confidence regions that are derived from χ^2-distributions with a prescribed degree of freedom (df). These are mathematical sound but are hardly applicable under realistic (highly non-linear) conditions. Students and researchers, myself included, often face

situations where it becomes unclear in whether the statistical preconditions are sufficiently met to apply idealized methods of Frequentist statistics. It becomes particularly delicate if parameter estimation is approached with Bayesian statistical considerations whereas concepts of Frequentist statistics are applied to come up with a practical (and "objective") solution of parameter identification. Very good background information and useful practical methods are described in Press et al. (2001) (online version here) in chapter 15 Modelling of Data. An example of how data from mesocosm experiments are used to estimate model parameters and constrain processes responsible for variations in C:N ratio of particulate organic matter (POC:PON) can be found in Schartau et al. (2007). Parameter identifiability is implicitly covered in a study on model reduction (Ward et al., 2013). Recently, some studies in the research field *System Biology* have addressed parameter identifiability explicitly, see e.g. Kreutz et al. (2012) and Raue et al. (2012). A publication by Raue et al. (2010) includes, according to my opinion, one of the best descriptions of the general problem of parameter identifiability and explains feasible methods, e.g. a parameter profile screening. I adopted their concept of separating structural- from practical parameter identifiability (see lecture slides). The following exercises were partially inspired by examples given in their paper.

Auxiliary material needed – a) Computer with MATLAB or OCTAVE, b) Equation Look Up Sheet (ELUS), and c) lecture slides (notes)

0 Outline of exercises (total duration: 60 minutes)

1 Comparison between optimized model solution with data from mesocosm experiment (5 to 10 minutes)

2 Identifying/specifying confidence limits of optimal parameter estimates via resampling strategy (10 minutes)

3 Looking at regions of confidence/credibility of cost (objective) function values of two-dimensional parameter variations (10 minutes)

4 Derive error estimates of optimized parameter values with approximation of Hessian matrix (10 minutes)

5 *Homework* – Comparison of confidence/credibility regions if only chlorophyll *a* data were available (5 to 10 minutes)

1 Comparison between optimized model solution with data from mesocosm experiment (5 to 10 minutes)

At first, we will performe a model run and compare the model solution with observations, enter:

```
>> model_solve
```

You will find the abbreviations of state variable names in the ELUS. Identify those state variables whose results do not match observations; specify critical periods (e.g. growth phase, bloom phase or post-bloom period).

2 Identifying/specifying confidence limits of optimal parameter estimates via resampling strategy (10 minutes)

We now apply a resampling strategy. A series of resample sets (default=2000) of the reference model solution and of the data is generated with:

```
>> model_solve_resample
```

Note that resample sets will differ between calls (no fixed seed is used for the random number generator). You will first see figures similar to (1), but this time with resample sets included. The trajectory of one (randomly selected) resample set is shown as dashed red line. Questions:
a) *Is the 'single' resample set characteristical for the data*?
b) *How would a single (randomly selected) resample set would look like if we had assumed independence of data*?

Based in the resample sets (noisy "twin" model results and resampled observations) the program also calculates cost function values (or χ^2 values) at the point of optimal parameter estimates (reference solution). To see these results enter:

```
>> make_fig_conf_lim
```

We find large variations of cost function values (between resample sets) at this specific point, well knowing that these solutions comply with our (identical twin) reference solution. These solutions are indifferent with respect to our error assumptions. Thus, the distribution of these cost function values can be used as representative probability density estimates of "acceptable" or "tolerable" model solutions (in contrast to a predefined χ^2-distribution with prescribed degree of freedom). Compare the different confidence limits that are obtained for a cost function that has a) a fixed covariance matrix, b) time variable variance information (no correlation) at dates of observation. Figures are stored in the directory

```
./FIGURES
```

Questions:
a) *If we compare the distribution of J what can we say about the degree of freedom*? Note that we have 102 observations.
b) *Do smaller cost function values automatically mean better fits to data (YES/NO)*?

3 Looking at regions of confidence/credibility of cost (objective) function values of two-dimensional parameter variations (10 minutes)

We will now consider the confidence limits that we have identfied in (2) and have a look on cost function values for variations of two parameters while all other remain at their optimal values. Call

```
>> make_plot_vari2D
```

Select

```
twin_cost_cov_p1Pc_p2aggreg_loss.mat
```

This shows us an almost ideal situation, with the maximum potential photosynthesis rate parameter (P_c) and the loss parameter due to particle aggregation (Φ_{agg}), see ELUS; so to say one *source* and one *sink* parameter. Figure 1 shows the result if a cost function with a covariance is used, Figure 2 was generated with a cost function assuming no correlations (as in ELUS). Compare the region of confidence (credibility regions). Question:
Which of the two Figures appears to show a correlation greater than the other?

Again, use

```
>> make_plot_vari2D
```

and consider other selections and compare between Figures (all with P_c as the first parameter): files that start with "obs" used cost functions with resample data, Figures are stored in the directory. The relevant parameters and the associated equations are listed in ELUS.

```
./2D_vari_results/FIGURES
```

Question:
Which parameter concerns us with respect to non-identifiability?

4 Derive error estimates of optimized parameter values with approximation of Hessian matrix (10-15minutes)

In this exercise we will derive second derivatives of the cost function with respect to the parameters, namely a Hessian matrix (in our context also called Fisher information). With

```
>> approx_hessian_Aggreg_Pc_cov
```

we will be asked on the incremental step size for approximating second derivatives of a Hessian matrix (see ELUS). By default the values are set to 10% (of each optimal parameter value). You may check two things:
a) *When increasing the incremental step sizes (given in percentage), how does the error ellipse change*?
b) *When gradually decreasing the incremental step sizes, at which "minimum" incremental step size does the calculation of the error ellipse (inversion of the approximated Hessian) fail?*
You may look (for each increment size) at the estimated errors as well:

```
>> [u_err_p1, u_err_p2]
```

Now have a look on the 2x2 elements of the Hessian itself:

```
>> pvar_hess
```

Question: *What does it mean when an off-diagonal element of the Hessian is negative?*
You may now select **one** pair of parameters (either good or problematic case) and compare between cost function with covariances

```
>> approx_hessian_*_cov
```

and cost function without correlations (only variances)

```
>> approx_hessian_*_uncorr
```

Questions:
a) *Are the approximated error ellipses representative?*
b) *Do the error ellipses reflect information that we had already seen in the Figures of J- and* χ^2*-distributions?*
c) *What is the reason for situations where an error ellipse cannot be derived?*
d) *What are the upper and lower limits of each parameter if you (here visually in 2D) apply the screening profile method?* (For this see lecture notes and only compare between "obs_cov_" and "obs_uncorr_")

5 *Homework* – Comparison of confidence/credibility regions if only chlorophyll *a* data were available (10-15 minutes)

The homework exercise goes through all points addressed before. In the directory:

```
>> cd ./CHLa_only
```

you will find all scripts as in the directory above, and you may use them in the same sequence described in exercises 1) through 5). However, this time only data from chlorophyll *a* concentrations (and their model counterparts; twins). This, for example, represents the situtation where only remote sensing data of chlorophyll *a* concentrations can be assimilated into a marine ecosystem- or biogeochemical model locally. Since we have learned that the available data contains less information than data that are truly independent. This due to the correlation between the observed variables, i.e. negative correlation between DIN and PON. Given this, you can analyse how conditions of parameter identifiability have changed with the consideration of chlorophyll *a* only.

ALL THE BEST!

References

Kreutz, C. , A. Raue, and J. Timmer (2012). Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Systems Biology*, **6**:120, 1-9.
`http://www.biomedcentral.com/1752-0509/6/120`

Press, W. , S. A. Teukolsky, W. T. Vetterling, B. P. Flannery (2007). *Numerical Recipes*, ISBN 0521431085.
accessible online version: `http://apps.nrbook.com/fortran/index.html`

Raue, A. , C. Kreutz, T. Maiwald, U. Klingmuller, and J. Timmer (2010). Addressing parameter identifiability by model-based experimentation. *IET Systems Biology*, **5**(2), 120-130.

Raue, A. , C. Kreutz, F. J. Theis, and J. Timmer (2013). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society*, **371**(1984-20110544), 1471-2962.

Schartau, M. , A. Engel, J. Schröter, S. Thoms, C. Völker, and D. Wolf-Gladrow (2007). Modelling carbon overconsumption and the formation of extracellular particulate organic carbon. *Biogeosciences*, **4**, 433-453
open access: `http://biogeosciences.net/4/433/2007/bg-4-433-2007.html`

Ward, B. , M. Schartau, A. Oschlies, A. Martin, M. Follows, and T. R. Anderson (2013). When is a biogeochemical model too complex? Objective model reduction and selection for North Atlantic time-series sites. *Progress in Oceanography*, **116**, 49-65.
from repository: `http://eprints.soton.ac.uk/356914/`

I myself like reading in