Homework 8, due 11:59pm Sunday April 14

DATASCI/STATS 531

You are welcome to collaborate with your final project group for this homework. Nevertheless, you will submit individual solutions and you should take individual responsibility for the work that you submit under your own name. From the perspective of scholarship, it is appropriate to give credit to specific group members for helpful suggestions. If you can explain in your own words why you found this advice helpful, and why you think the advice is correct, then you can satisfy both the scholarship and individual contribution requirements for the grading scheme. If you poke around, you will find that solutions to these questions exist online. It may be simplest for you if you do not look for online solutions. If you are reluctant to give full and transparent credit to what you have borrowed from sources, that may be a sign you have borrowed too much.

In case it is helpful, here is a different way to summarize the course rules on scholarship, which is based on the rules of modern science:

If there is any information about how your report was produced which would diminish your own contribution then you have an obligation to share that information with the reader, so you do not claim credit for something you have not done yourself. Put another way, you have an obligation to help the reader quickly and fairly evaluate your individual contribution.

Your report should be written in Rmd, and you should submit both the Rmd source and a version compiled to HTML. For each question, you should give an answer (a letter or true/false claims) followed by some text explaining your answer. Most of the questions do not need any coding, but you should show your code when appropriate.

The goal is to think about some issues that might arise in final projects, and indeed some of the questions alert you to common issues arising in previous 531 final projects.

These questions are based on material developed for a course on Simulation-based Inference for Epidemiological Dynamics. Note that this vague ackowledgement of a large but unspecified intellectual debt is not intended as a role model for homework reports; it is okay in this context but insufficient to explain your own contribution in a course assignment.

Question 1. From a scientific perspective, conclusions should not depend on the units we choose. However, we must get the details straight to correctly describe a POMP model and its `pomp` representation. Suppose our data are two years of weekly aggregated case reports of a disease and we have a continuous time model solved numerically using an Euler timestep of size dt. Which one of the following is a correct explanation of our options for properly implementing this in a `pomp` object called `po`?

(A) The measurement times, `time(po)`, should be in units of weeks, such as $1, 2, \ldots, 104$. The latent process can be modeled using arbitrary time units, say days or weeks or years. The units of dt should match the time units of the **latent** process.

(B) The measurement times, `time(po)`, should be in units of weeks, such as $1, 2, \ldots, 104$. The latent process can be modeled using arbitrary time units, say days or weeks or years. The units of dt should be in weeks (in practice, usually a fraction of a week) to match the units of the **measurement** times.

(C) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`$=1/52$, $2/52, \ldots, 2$. The latent process and dt should use the same units of time as the measurement times.

(D) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`$=1/52$, $2/52, \ldots, 2$. The latent process can also use arbitrary units of time, which do not necessarily match the units of the measurement times. The units of dt should match the units used for the **latent** process.

(E) The measurement times do not have to be in units of weeks. For example, we could use `time(po)`=1/52, 2/52, . . . , 2. The latent process can also use arbitrary units of time, which do not necessarily match the units of the measurement times. The units of dt should match the units used for the **measurement** times.

Question 2. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.so’:
status = 1
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'  -I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c:39:5:
error: called object type 'int' is not a function or function pointer
    W = 0;
    ^
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/5
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
running command 'PKG_CPPFLAGS="-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'" '/Library/Frameworks/R.framework/Resources/bin/R'
CMD SHLIB -c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_be9007eb030e47cb34264e3e779b6da9.c
2>&1' had status 1
```

Which one of the following is the most plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 3. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.so’: status = 1
```

```
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include' -I'/Users/ionides/sbied/questions'
-I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c:33:16:
error: use of undeclared identifier 'pop'; did you mean 'pow'?
    double m = pop/(S_0+I_0+R_0);
               ^~~
               pow
/Applications/
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
  running command 'PKG_CPPFLAGS="-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'" '/Library/Frameworks/R.framework/Resources/bin/R' CMD SHLIB
-c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_b675d99e691eda865610f570058ea3be.c 2>&1' had status 1
```

Which one of the following is the most plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 4. Suppose you obtain the following error message when you build your pomp model using Csnippets.

```
Error: error in building shared-object library from C snippets: in ‘Cbuilder’: compilation error:
cannot compile shared-object library ‘/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.so’: status = 1
compiler messages:
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG
-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'  -I/usr/local/include   -fPIC  -Wall -g -O2
-c /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c
-o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.o
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c:39:36:
error: too many arguments to function call, expected 2, have 3
      rep = nearbyint(rnorm(1,mean,sd));
                      ~~~~~        ~~
```

```
/Librar
In addition: Warning message:
In system2(command = R.home("bin/R"), args = c("CMD", "SHLIB", "-c",  :
running command 'PKG_CPPFLAGS="-I'/Users/ionides/Library/R/x86_64/4.1/library/pomp/include'
-I'/Users/ionides/sbied/questions'" '/Library/Frameworks/R.framework/Resources/bin/R'
CMD SHLIB -c -o /var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.so
/var/folders/fv/pt62sh2d6_gf9fp3t7b466vr0000gr/T//RtmpD16GmG/
5324/pomp_77886fb66d95b4b9904440d86a4425b3.c 2>&1' had status 1
```

Which one of the following is the most plausible cause for this error?

(A) Using R syntax within a C function that has the same name as an R function.

(B) A parameter is missing from the `paramnames` argument to `pomp`.

(C) Indexing past the end of an array because C labels indices starting at 0.

(D) Using `beta` as a parameter name when it is a declared C function.

(E) A missing semicolon at the end of a line.

Question 5. Let V_n be a Markov process and let $W_n = h(V_n)$ for some function h. Which of the following statements are true?

i) W_n is a Markov process for all choices of h.

ii) W_n is a Markov process for some choices of h.

iii) W_n is not a Markov process for any choice of h.

iv) If $V_n = (X_n, Y_n)$ where X_n and Y_n are a POMP model, and $h(X_n, Y_n) = X_n$ then W_n is a Markov process.

v) If $V_n = (X_n, Y_n)$ where X_n and Y_n are a POMP model, and $h(X_n, Y_n) = Y_n$ then W_n is a Markov process.

(A) i,iv,v

(B) ii,iv

(C) ii,v

(D) iii

(E) None of the above

Question 6. Suppose that 10 replications of a particle filter, each using 10^3 particles, runs in 15 minutes with no parallelization. To look for a more precise likelihood evaulation, you consider running 20 replicates, each with 10^4 particles. How many minutes will this take, if you distribute the calculation across 4 cores?

(A) 50

(B) 60

(C) 75

(D) 120

(E) 300

Question 7. A particle filter is repeated 5 times to evaluate the likelihood at a proposed maximum likelihood estimate, each time with 10^4 particles. Suppose the log likelihood estimates are -2446.0, -2444.0, -2443.0, -2442.0, -2440.0. Which of the following is an appropriate estimate for the log likelihood at this parameter value and its standard error.

(A) Estimate $= -2443.0$, with standard error 1.0

(B) Estimate $= -2443.0$, with standard error 2.2

(C) Estimate $= -2443.0$, with standard error 5.0

(D) Estimate $= -2441.4$, with standard error 2.2

(E) Estimate $= -2441.4$, with standard error 1.4

Question 8. What is the log likelihood (to the nearest unit) of the Dacca cholera data for the POMP model constructed in pomp via

```
d <- dacca(deltaI=0.08)
```

with cholera mortality rate 8% and other parameters fixed at the default values.

(A) -3764

(B) -3765

(C) -3766

(D) -3767

(E) -3768

Question 9. Effective sample size (ESS) is one of the main tools for diagnosing the success of a particle filter. If you plot an object of class `pfilterd_pomp` (created by applying `pfilter` to a pomp object), the ESS is displayed. Suppose one or more time points have low ESS (say, less than 10) even when using a fairly large number of particles (say, 10^4). What is the proper interpretation?

(A) There is a problem with data, perhaps an error recording an observation.

(B) There is a problem with the model which means that it cannot explain something in the data.

(C) The model and data have no major problems, but the model happens to be problematic for the particle filter algorithm.

(D) At least one of **??**, **??** and **??**.

(E) Either **??** or **??** or both, but not **??**. If the model fits the data well, the particle filter is guaranteed to work well.

Question 10. When carrying out inference by iterated particle filtering, the likelihood increases for the first 10 iterations or so, and then steadily decreases. Testing the inference procedure on simulated data, this does not happen and the likelihood increases steadily toward convergence. Which one of the following is the best explanation for this?

(A) One or more random walk standard deviation is too large.

(B) One or more random walk standard deviations is too small.

(C) The model is misspecified, so it does not fit the data adequately.

(D) A combination of the parameters is weakly identified, leading to a ridge in the likelihood surface.

(E) Too few particles are being used.

Question 11. People sometimes confuse likelihood profiles with likelihood slices. Suppose you read a figure which claims to construct a profile confidence interval for a parameter ρ in a POMP model with four unknown parameters. Suppose that the code producing the plot is available to you as an Rmarkdown file. Which one of the following confirms that the plot is, or is not, a properly constructed profile confidence interval.

(A) The CI is constructed by obtaining the interval of rho values whose log likelihood is within 1.92 of the maximum on a smoothed curve of likelihood values plotted against ρ.

(B) The code involves evaluation of the likelihood but not maximization.

(C) The points along the ρ axis are not equally spaced.

(D) The smoothed line shown in the plot is close to quadratic.

(E) Both **??** and **??** together.

Question 12. For each of the following, say whether the statement is true or false.

(A) A profile likelihood must lie *above* every slice.

(B) Confidence intervals can be read from likelihood slices.

(C) A poor man's profile must lie above the true profile.

(D) A poor man's profile must lie below the true profile.

Question 13. The iterated filtering convergence diagnostics plot in Fig. **??** comes from a student project. What is the best interpretation?

(A) Everything seems to be working fine. The likelihood is climbing. The replicated searches are giving consistent runs. The spread of convergence points for σ_ν and H_0 indicates weak identifability, which is a statistical fact worth noticing but not a weakness of the model.

(B) The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. Additional searching is needed, experimenting with **larger** values of the random walk perturbation standard deviation for these parameters to make sure the parameter space is properly searched.

(C) The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. Additional searching is needed, experimenting with **smaller** values of the random walk perturbation standard deviation for these parameters to make sure the parameter space is properly searched.

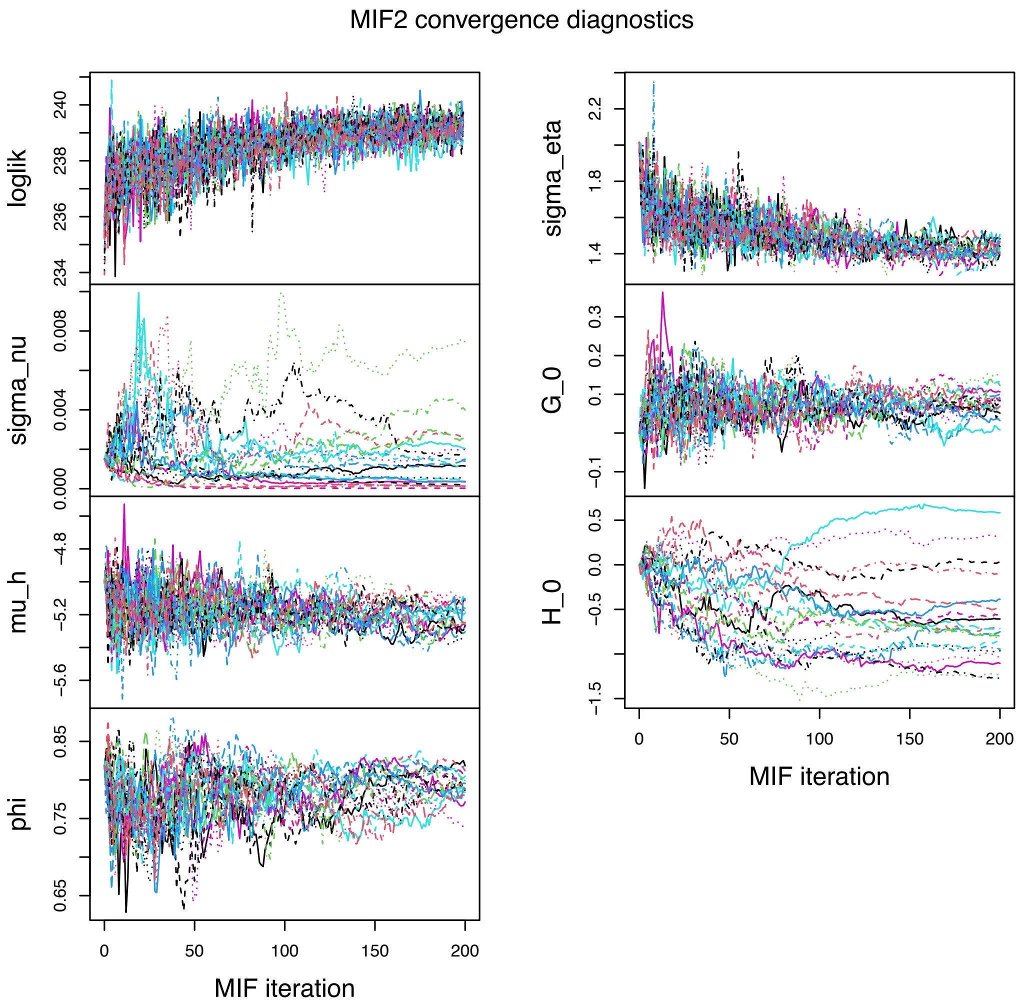

Figure 1: Iterated filtering convergence diagnotic plot for ??

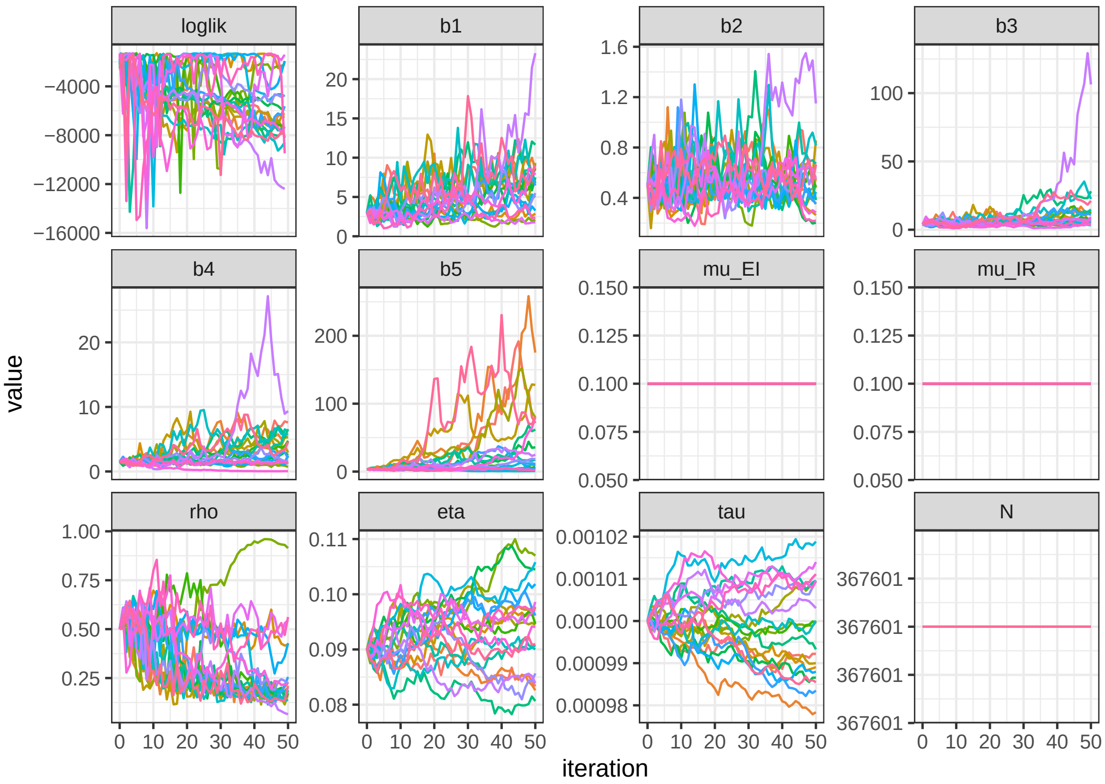

Figure 2: Iterated filtering convergence diagnotic plot for **??**

(D) The consistently climbing likelihood is promising, but the failure of σ_ν and H_0 to converge needs attention. This indicates weak identifiability which cannot be solved by improving the searching algorithm. Instead, we should change the model, or fix one or more parameters at scientifically plausible values, to resolve the identifiability issue before proceeding.

(E) Although the log likelihood seems to be climbing during the search, until the convergence problems with σ_ν and H_0 have been addressed we should not be confident about the successful optimization of the likelihood function or the other parameter estimates.

Question 14. The iterated filtering convergence diagnostics plot in Fig. **??** comes from a student project, calculated using 10^3 particles. Which one of the following is the best interpretation of this diagnostic plot?

(A) Everything seems to be working fine. There is a clear consensus from the different searches concerning the highest likelihood that can be found. Therefore, the search is doing a good job of maximization. Occasional searches get lost, such as the purple line with a low likelihood, but that is not a problem.

(B) The seaches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should use more particles and/or more iterations to achieve this.

(C) The seaches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should compare the best likelihoods obtained with simple statistical models, such as an auto-regressive moving average model, to look for evidence of model misspecification.

(D) The seaches obtain likelihood values spread over thousands of log units. We would like to see consistent convergence within a few log units. We should look at the effective sample size plot for the best fit we have found yet, to see whether there are problems with the particle filtering.

(E) All of **??**, **??** and **??**.

Question 15. In the following call to `mif2`, which of the statements below are true? You may assume that `obj` is a pomp object with parameters `alpha`, `Beta`, `gamma`, and `delta`.

```
obj %>%
  mif2(
    Nmif=100,
    partrans=parameter_trans(log=c("Beta","alpha","delta")),
    paramnames=c("Beta","alpha","delta"),
    rw.sd=rw.sd(Beta=0.05,alpha=ivp(0.02),gamma=0.05),
    cooling.fraction.50=0.1
  ) -> obj
```

(A) 50 IF2 iterations will be performed.

(B) `Beta` and `alpha` are estimated on the log scale.

(C) `gamma` is not estimated.

(D) `delta` is not estimated.

(E) The magnitude of the perturbation on `Beta` at the end of the run will be $0.05{\times}0.1^{100} = 5{\times}10^{-102}$.

(F) The magnitude of the perturbation on `gamma` at the end of the run will be $0.05{\times}0.1^{100/50} = 5{\times}10^{-4}$.

(G) `alpha` is an initial-value parameter; it will be perturbed only at the beginning of the time series.

(H) After the call, `obj` is an object of class 'mif2d_pomp'.

Question 16. Assume that `obj` is the result of the call in **??**, we consider the result of the following call.

```
obj %>%
  mif2(
    rw.sd=rw.sd(Beta=0.05,alpha=ivp(0.02)),
    cooling.fraction.50=0.2
  )
```

Explain whether each of the following are true or false.

(A) 100 more IF2 iterations will be performed.

(B) The settings of the previous calculation are re-used, with the exception of `rw.sd` and `cooling.fraction.50`.

(C) The starting point of the new calculation is the end point of the old one.

(D) `Beta` and `alpha` are estimated on the log scale.

(E) `gamma` is not estimated.

(F) `delta` is not estimated.

(G) The cooling occurs more quickly than in the previous call.