Parallel programming in Stata

Brian Quistorff

Department of Economics
University of Maryland

Cluster Mini-talks, 2014

Easy Wins For Parallelism

It is easy to parallelize a set of operations that don't depend on each other and where there is a clear way to delineate what is the same/different between tasks. Common situations include:

1. Bootstrapping: Data is the same, random seed for resampling changes.
2. Individual-level structural model: Model parameters the same, person-data changes.

Easy Wins For Parallelism

It is easy to parallelize a set of operations that don't depend on each other and where there is a clear way to delineate what is the same/different between tasks. Common situations include:

1. Bootstrapping: Data is the same, random seed for resampling changes.
2. Individual-level structural model: Model parameters the same, person-data changes.

Easy Wins For Parallelism

It is easy to parallelize a set of operations that don't depend on each other and where there is a clear way to delineate what is the same/different between tasks. Common situations include:

1. Bootstrapping: Data is the same, random seed for resampling changes.
2. Individual-level structural model: Model parameters the same, person-data changes.

Can't I just use Stata-MP?

- Stata-MP (which is on the cluster) has coded some some commands to take advantage of multiple processors but doesn't allow explicit parallelization by the programer.
- See which commands can be sped up: `http://www.stata.com/statamp/statamp.pdf`

Can't I just use Stata-MP?

- Stata-MP (which is on the cluster) has coded some some commands to take advantage of multiple processors but doesn't allow explicit parallelization by the programer.
- See which commands can be sped up: `http://www.stata.com/statamp/statamp.pdf`

Stata's -parallel- module

- Documentation:
  `http://fmwww.bc.edu/repec/bocode/p/parallel.pdf`
- Can run a .do file or command in parallel.
- I'll show an example of parallelizing a bootstrap procedure.

Stata's -parallel- module

- Documentation:
  `http://fmwww.bc.edu/repec/bocode/p/parallel.pdf`
- Can run a .do file or command in parallel.
- I'll show an example of parallelizing a bootstrap procedure.

Stata's -parallel- module

- Documentation:
  `http://fmwww.bc.edu/repec/bocode/p/parallel.pdf`
- Can run a .do file or command in parallel.
- I'll show an example of parallelizing a bootstrap procedure.

Basic coding guides

- Keep intermediate files around until sure computation worked (because debugging is harder than normal).
- Make switching between parallel and non-parallel both easy stylistically and result in the same output. For the latter, pre-compute seeds for all runs. Otherwise, have to worry about problems of joint randomization in parallel RNGs of Stata (seeOzier - Perils of simulation).
- (see code)

Basic coding guides

- Keep intermediate files around until sure computation worked (because debugging is harder than normal).
- Make switching between parallel and non-parallel both easy stylistically and result in the same output. For the latter, pre-compute seeds for all runs. Otherwise, have to worry about problems of joint randomization in parallel RNGs of Stata (seeOzier - Perils of simulation).

- (see code)

Basic coding guides

- Keep intermediate files around until sure computation worked (because debugging is harder than normal).
- Make switching between parallel and non-parallel both easy stylistically and result in the same output. For the latter, pre-compute seeds for all runs. Otherwise, have to worry about problems of joint randomization in parallel RNGs of Stata (seeOzier - Perils of simulation).

- (see code)

Debugging

- Check the intermediate __pll${pid}_do${pll_instance}.log along with your own log
- Use -set trace on- (and tracedepth)
- Always test with smaller reps/dataset before running big.

Debugging

- Check the intermediate __pll${pid}_do${pll_instance}.log along with your own log
- Use -set trace on- (and tracedepth)
- Always test with smaller reps/dataset before running big.

Debugging

- Check the intermediate __pll${pid}_do${pll_instance}.log along with your own log
- Use -set trace on- (and tracedepth)
- Always test with smaller reps/dataset before running big.

Other notes

- -parallel- doesn't copy settings such as matsize (so re-include).
- -parallel- doesn't copy scalars or matrices (so use globals or mata objects).
- -parallel- doesn't work in batch mode for windows w/o extra work (ask Brian)
- Copied mata objects will move after -parallel- so earlier pointers will be incorrect (so regenerate).

Other notes

- -parallel- doesn't copy settings such as matsize (so re-include).
- -parallel- doesn't copy scalars or matrices (so use globals or mata objects).
- -parallel- doesn't work in batch mode for windows w/o extra work (ask Brian)
- Copied mata objects will move after -parallel- so earlier pointers will be incorrect (so regenerate).

Other notes

- -parallel- doesn't copy settings such as matsize (so re-include).
- -parallel- doesn't copy scalars or matrices (so use globals or mata objects).
- -parallel- doesn't work in batch mode for windows w/o extra work (ask Brian)
- Copied mata objects will move after -parallel- so earlier pointers will be incorrect (so regenerate).

Other notes

- -parallel- doesn't copy settings such as matsize (so re-include).
- -parallel- doesn't copy scalars or matrices (so use globals or mata objects).
- -parallel- doesn't work in batch mode for windows w/o extra work (ask Brian)
- Copied mata objects will move after -parallel- so earlier pointers will be incorrect (so regenerate).