Parallel Processing for Faster ARIMA Model Selection

Drew Schmidt

April 10, 2012

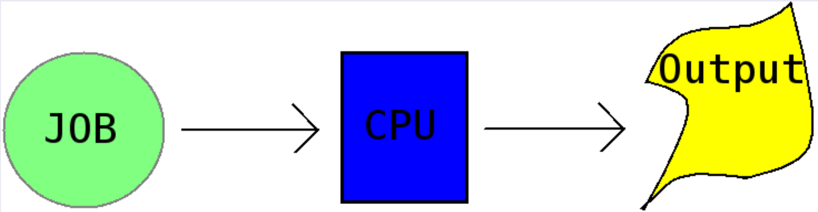
Contents

What is Parallelization?

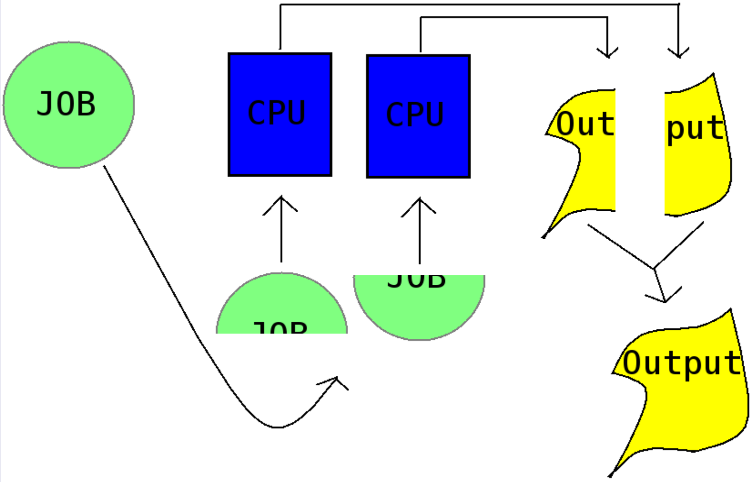
Parallel Processing

Parallel processing (as opposed to *serial* processing) is the use of multiple processors and/or multiple processing cores (and/or hyperthreading) in computation.

Serial Processing



Parallel Processing



Serial Processing Example

$n =$ big number
for $i = 1, 2, 3, 4, \ldots, n$
    do something
done

Parallel Processing Example - 2 Cores

$n =$ big even number

`for` $i = 1, 3, 5, 7, \ldots, n-1$ do something `done`	`for` $i = 2, 4, 6, 8, \ldots, n$ do something `done`

`combine`

Why Should We Care About It?

Why Should We Care About Parallelization?

- Processors haven't really gotten "faster" in the last 10 years
- Parallel is faster than serial
- Other people care about parallelization (i.e., looks good on a résumé)
- Usually easier than other ways of improving performance

Speedup

Parallel processing is generally faster. Ideally, we would want

$$(\text{multi core run time}) = (\text{number cores}) * (\text{single core run time})$$

In practice, this is rarely the case because of overhead inherent to (most) implementations.

A Speedup Analogy

- one core $\rightarrow$ many cores
- one checkout lane $\rightarrow$ many checkout lanes

Parallel ARIMA Model Selection

ARIMA Model Parameter Combinations

Suppose you want to fit a seasonal ARIMA model to monthly data

$$ARIMA(p, d, q)_1(P, D, Q)_{12}$$

Choices for parameters:

- Generally any of p, q, P, Q will range from 0 to 10 (rarely a sensible need to go above 5).
- Both of d and D should generally be 0 or 1; large degrees of differencing can inject structure which isn't actually there.

ARIMA Model Parameter Combinations

This gives us a general guideline of checking a number of models on the order of $5^4 2^2 = 5184$ (or debatably $10^4 2^2 = 58564$).

This can not be done by hand.

And this is with just *one* form of seasonality! Consider data with monthly, weekly, and daily seasonality:

$$ARIMA(p, d, q)_1(P_1, D_1, Q_1)_{12}(P_2, D_2, Q_2)_{52}(P_3, D_3, Q_3)_{365}$$

Our loose guidelines put the order of models to check at 26,873,856 *for just one time series*!

Checking Several Models At Once

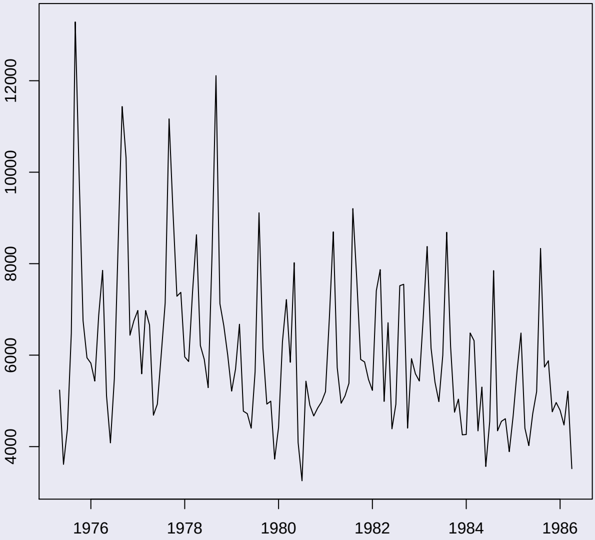
With so many models to evaluate and since any two such evaluations are independent, we can spread the workload across several cores.

PRIVATE SELL OF SPECIAL TRANSPORT CARS IN RFA[1]

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1975					5237	3612	4385	6479	13286	9928	6755	5943
1976	5824	5433	6868	7853	5112	4082	5483	8502	11436	10313	6437	6751
1977	6976	5592	6975	6659	4688	4929	6019	7140	11166	9140	7288	7373
1978	5962	5861	7378	8632	6212	5906	5290	8238	12111	7127	6643	5998
1979	5214	5698	6675	4775	4721	4404	5635	9111	6150	4930	4994	3727
1980	4421	6296	7215	5844	8021	4090	3253	5431	4905	4671	4843	4980
1981	5201	6826	8696	5730	4949	5109	5386	9204	7656	5906	5851	5469
1982	5230	7409	7869	4989	6708	4389	4926	7518	7549	4405	5923	5593
1983	5435	6821	8376	6154	5412	4984	5999	8686	6184	4753	5038	4258
1984	4265	6484	6316	4344	5302	3566	4549	7848	4346	4553	4610	3890
1985	4675	5655	6484	4405	4022	4719	5203	8333	5740	5877	4762	4964
1986	4794	4473	5212	3518								

[1] Series 394 from the Mcomp package for R

PRIVATE SELL OF SPECIAL TRANSPORT CARS IN RFA



Model Selection

We fit all seasonal ARIMA models with $d = D = 0$ and each of p, q, P, Q ranging from 0 to 6 for a total of 2401 models. The winning model is the one with smallest Bayesian Information Criterion (BIC) score, computed as

$$-2 loglik + k \ln(n)$$

On a core i5 with 4 cores (no hyperthreading)

Cores	Processes	Run Time	Avg # Models/Second
1	1	44.181 seconds	217.3785
4	4	17.126 seconds	560.7848
4	8	14.566 seconds	659.3437

18 Month Forecast from $ARIMA(1,1,1)_1(1,0,1)_{12}$ — $R^2_{\text{Test}} \approx .800$