Testing My LDA

Georgie Mansell

02/08/2021

Introduction

MSLDA is now an R package on github (https://github.com/g-l-mansell/MSLDA) which contains my different implementations of LDA:

```
library(MSLDA)
```

- LDA_original - as in the Blei 2003 paper - the only implementation that does not run in parallel - only some initialisations will work
- LDA_original_par - as above but the E-step is run in parallel - should give the same results with the same seed.
- LDA_noalpha - since the alpha update rule in LDA_original is flawed, this version treats alpha as a hyperparameter and so all initialisations should work.
- LDA_reshaped - this version is a further adaptation of LDA_original, which uses a count (document-term) matrix as an input rather than the document vectors - this should give the same results as LDA_noalpha with improved speed.
- LDA_smoothed - as in the Hoffman 2010 paper (batch LDA section) - this version assumes beta is a random variable.

This script will then contain the analysis, so everything is in one place to be easily rerun. Whether to rerun the analyses or just load the results of past runs will be controlled here:

```
rerun_original <- T
rerun_reshaped <- T
rerun_smoothed <- T
```

Edits to make

- Finish tidying the package functions
- Add all the analyses to this markdown
- Try to implement one of the methods in Rcpp - including using RcppParallel
- Try to follow Colins version with Poission prior

Check 1

Running all implementations on the simulated set of documents

```
load("data/MyCorpus.Rdata")
par(mfrow=c(1, 2))

res1 <- lda_original(docs, K=3, seed=83)
```

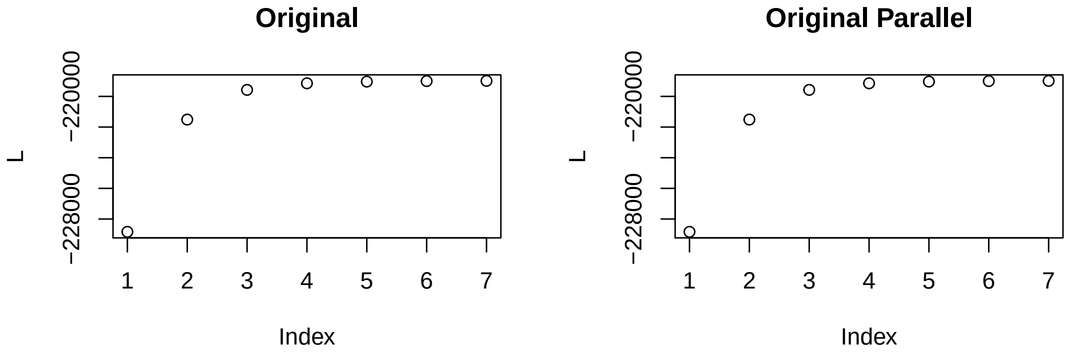
```
## [1] "Iteration 1"
## [1] "Iteration 2"
## [1] "Iteration 3"
## [1] "Iteration 4"
## [1] "Iteration 5"
## [1] "Iteration 6"
## [1] "Iteration 7"
```

```
plot(res1$Ls, ylab="L", main="Original")

res2 <- lda_original_par(docs, K=3, seed=83)
```

```
## [1] "Iteration 1"
## [1] "Iteration 2"
## [1] "Iteration 3"
## [1] "Iteration 4"
## [1] "Iteration 5"
## [1] "Iteration 6"
## [1] "Iteration 7"
```

```
plot(res2$Ls, ylab="L", main="Original Parallel")
```



Original

Original Parallel

```
res3 <- lda_noalpha(docs, K=3, seed=83)
```

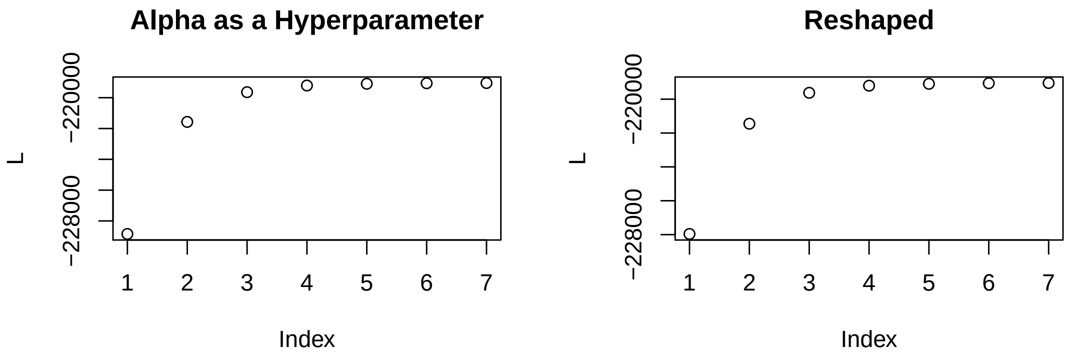
```
## [1] "Iteration 1"
## [1] "Iteration 2"
## [1] "Iteration 3"
## [1] "Iteration 4"
## [1] "Iteration 5"
## [1] "Iteration 6"
## [1] "Iteration 7"
```

```
plot(res3$Ls, ylab="L", main="Alpha as a Hyperparameter")

res4 <- lda_reshaped(counts, K=3, seed=83)
```

```
## [1] "Iteration 1"
## [1] "Iteration 2"
## [1] "Iteration 3"
## [1] "Iteration 4"
## [1] "Iteration 5"
## [1] "Iteration 6"
## [1] "Iteration 7"
```

```
plot(res4$Ls, ylab="L", main="Reshaped")
```



Alpha as a Hyperparameter

Reshaped

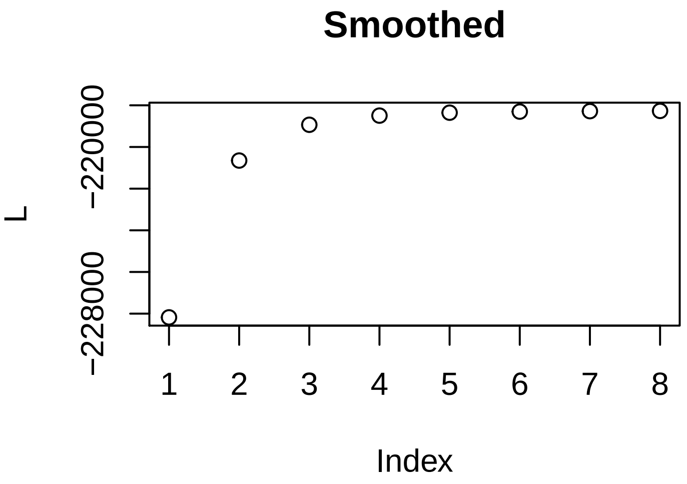
```
res5 <- lda_smoothed(counts, K=3, seed=83)
```

```
## [1] "Iteration 1"
## [1] "Iteration 2"
## [1] "Iteration 3"
## [1] "Iteration 4"
## [1] "Iteration 5"
## [1] "Iteration 6"
## [1] "Iteration 7"
## [1] "Iteration 8"
```

```
plot(res5$Ls, ylab="L", main="Smoothed")

#save(res1, res2, res3, res4, res5, file="analysis/InitalTest.Rdata")
sapply(list(res1, res2, res3, res4, res5), function(res) res$L)
```

```
## [1] -218988.7 -218988.7 -219047.5 -219047.3 -218269.4
```



Smoothed

Text data

Run all 3 versions on the simulated dataset of documents

Check that all three have monotonically increasing likelihood

Check that all 3 find reasonable thetas and beta when K=3