

BAN432 fall 2022  
First group assignment

Group 11

2022-10-22

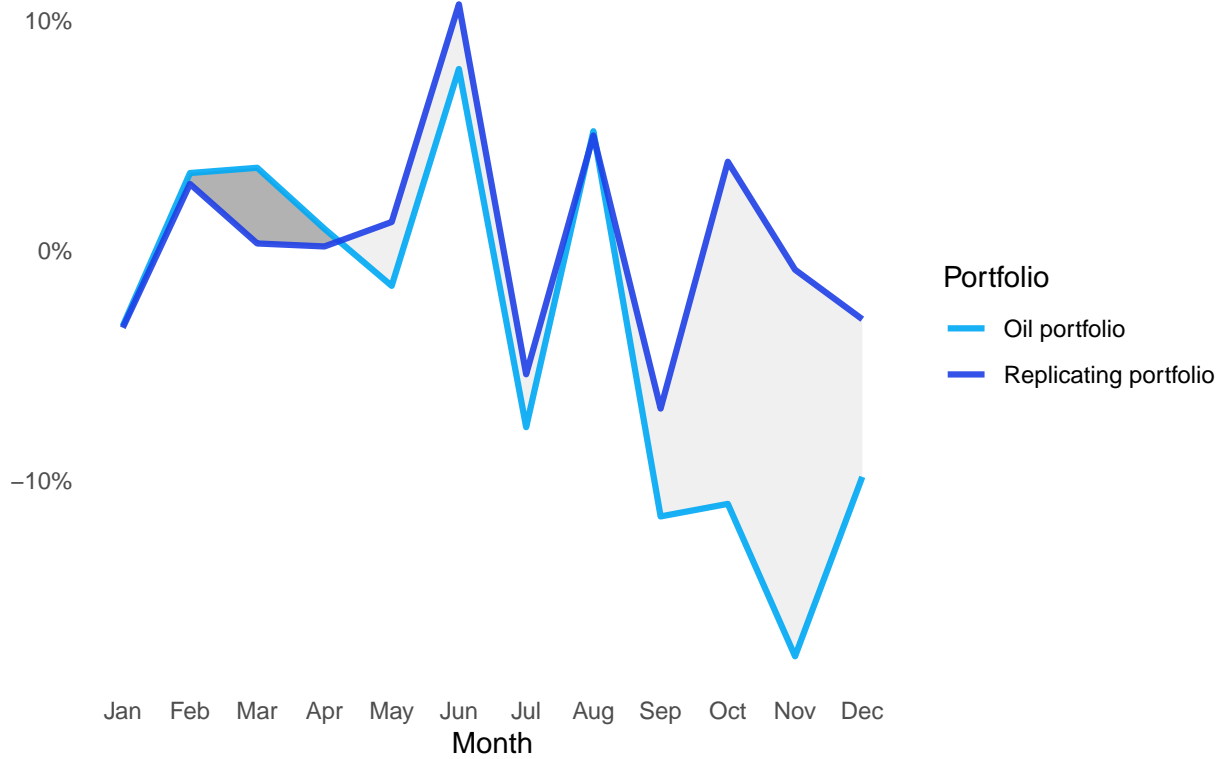
**Bi-grams**

## Selecting by ll

Table 1: Top 10 Log-Likelihood bigrams

	ll
proved_reserves	1004.0300
working_interest	991.5506
hydraulic_fracturing	566.1345
proved_undeveloped	516.7899
gross_acres	462.5419
estimated_proved	444.2142
shale_play	387.4694
undeveloped_reserves	366.5558
natural_production	319.3342
reserves_december	285.7961

Figure 1: Bi-grams replicating portfolio performance



```
## [1] "RMSE of bi-grams replicating portfolio is 0.00503"
```

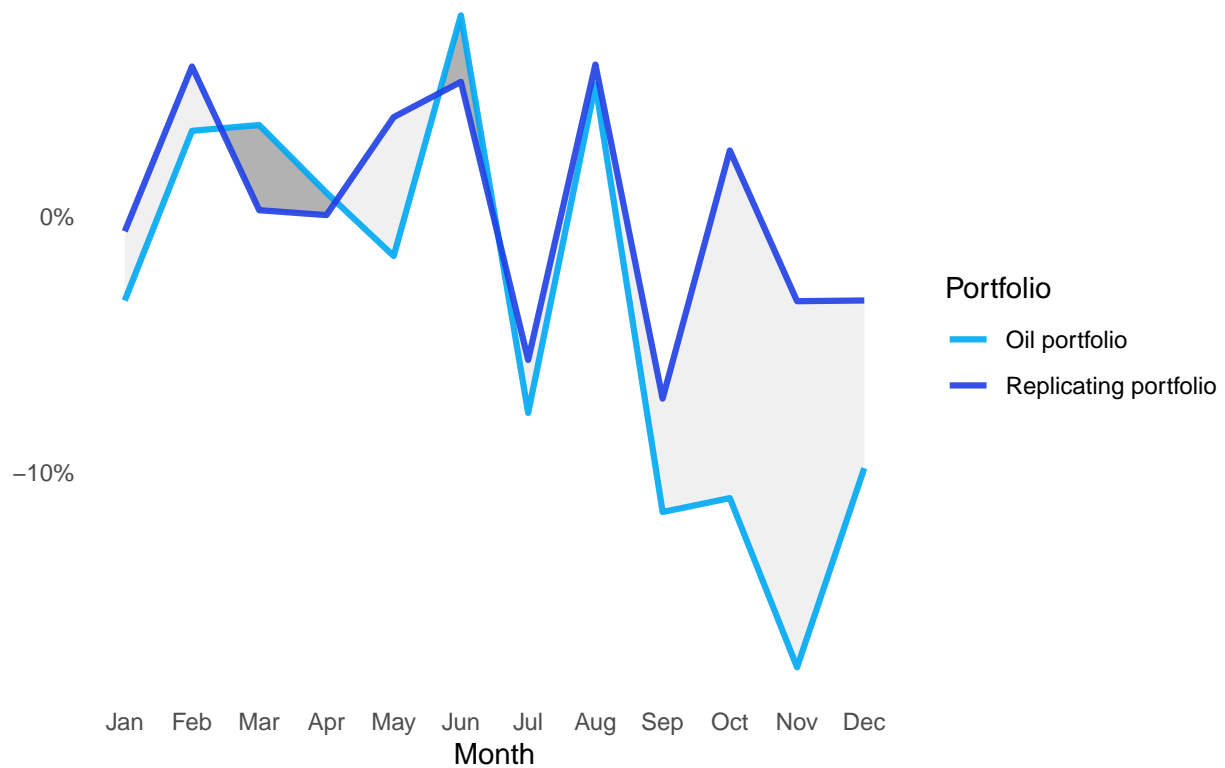
## Uni-grams

```
## Selecting by ll
```

Table 2: Top 10 Log-Likelihood unigrams

	ll
natural	2906.320
wells	2311.355
production	2059.049
proved	1989.841
drilling	1861.884
reserves	1633.491
block	1458.118
field	1440.347
exploration	1386.708
acres	1007.866

Figure 2: Uni-grams replicating portfolio performance



```
## [1] "RMSE of uni-grams replicating portfolio is 0.00433"
```

Oil firm tracking portfolio constructed based on the uni-grams document term matrix performs significantly better than the one constructed with the use of bi-grams, as demonstrated by the Root Mean Squared Error measure. However, the use of uni-grams or bi-grams does not affect the overall relative shape of returns as shown by the charts above.

## Optional

Figure 3: Random portfolios RMSE distribution

