APPLICATIONS OF BIG DATA METHODS IN FINANCE: INDEX TRACKING
Kamil Simka\(^1\), Luca Margaritella\(^2\)
\(^1\) Faculty of Economic Sciences, University of Warsaw (kamilsimka@gmail.com)
\(^2\) School of Business and Economics (SBE), Maastricht University (luca.margaritella@gmail.com)

RESEARCH OBJECTIVES
Although the \textit{curse of dimensionality} does not relate to most financial settings, high-dimensional methods gained some relevance in the recent finance literature. \textbf{Index tracking} aims at finding an optimal sample of stocks able to mimic the behavior of an equity index. We solve the problem by imposing a \(\ell_1\) regularization in the optimization procedure and we consider the LARS of Efron et al. (2004) algorithm in order to solve the problem.

DATA
Most researchers focus their analyses on indexes of developed economies. But the biggest challenge for index tracking models is imposed by emerging markets. Such markets frequently lack in historical data and consist of numerous stocks. Here, market indexes of the Warsaw Stock Exchange are considered, namely the WIG, WIG20, and mWIG40 between January 2009 and December 2016 with weekly frequency. Data are collected from bossa.pl.

CONSTRAINT OPTIMIZATION PROBLEM
To track the indexes weekly returns \((R_i)\) a dynamic portfolio optimization algorithm is constructed by selecting a predefined number of the most significant stocks \((K)\). To this aim, we firstly imply the shrinking \(\ell_1\) penalty formulation of the Penalized least squares equation (LASSO, Tibshirani (1996)) to obtain variable selection. Secondly, the underlying parameters \(\omega\)'s are set to satisfy a collection of linear constraints, thus yielding the following \textbf{Constrained LASSO} (CLASSO) parametrization:

\[
\begin{align*}
\text{minimize} & \quad ||R_t - R_P\omega||_2^2 / T + \lambda ||\omega||_1 \\
\text{s.t.} & \quad \sum_{j=1}^{J^*} \omega_j = 1; \quad 0 \leq \omega \leq 1; \quad \#J^* = K
\end{align*}
\]

As LASSO imposes an unwanted bias to coefficients, we carry out a \textbf{post OLS} estimation with variables selected by the CLASSO procedure (P-CLASSO):

\[
\begin{align*}
\text{minimize} & \quad ||R_t - R_P\omega||_2^2 / T \\
\text{s.t.} & \quad \sum_{j=1}^{J^*} \omega_j = 1; \quad 0 \leq \omega \leq 1
\end{align*}
\]

EVALUATION OF TRACKING QUALITY
\[
\begin{align*}
\text{TE} &= \frac{1}{T}||R_{jt} - R_{Pj}||_2 \\
\text{RMSE} &= \frac{1}{T}||R_{jt} - R_{Pj}||_2 \quad \text{MAD} = \frac{1}{T}||R_{jt} - R_{Pj}||_1
\end{align*}
\]

FITTING PERFORMANCE

<table>
<thead>
<tr>
<th>Fund Name</th>
<th>TE</th>
<th>RMSE</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYXOR WIG20</td>
<td>0.422</td>
<td>0.059</td>
<td>0.299</td>
</tr>
</tbody>
</table>

Notes: the first table: results of post constrained lasso with the tuning parameter selected my means of cross-validation. the second table: results of Lyxor ETF WIG20. TE stands for the tracking error, RMSE the root mean square error and MAD the mean absolute deviation. Portfolio rebalancing every half-year, 1% transaction costs.

CONCLUDING REMARKS AND FUTURE WORK
We make a comparison of portfolio management approaches which are based on correlation selection criteria and the constrained post lasso. The second approach is more effective in terms of efficiency of tracking the market indexes. Additionally, The method of P-CLASSO has been compared with the LYXOR WIG20 ETF, which is the only ETF that tracks the WIG20 market index. In the lasso optimization procedure we have used the ten-fold cross-validation method in order to select the tuning parameter. The tracking quality measures are substantially lower for the P-CLASSO model. Further investigations will be needed in order consider the sub-period re-balancing of portfolios, as well as different transaction costs schemes.

R PACKAGES USED
data.table; dplyr; glmnet; hydroGOF; lars; lubridate; PerformanceAnalytics; quadprog; quantmod; xts.

REFERENCES