Sparse random forests and dimension reduction

(Work in progress)

Workshop Économetrie Appliquée - Nantes - 24 septembre 2021

Amélie CHARLES
(Audencia Business School)

Olivier DARNÉ
(LEMNA, University of Nantes)

Outlines

- 1. Motivation
- 2. Methodology
- 3. Data
- 4. Empirical results
- 5. Limitations

▶ Considerable attention has focused on the forecasting of macroeconomic variables in a data-rich environment via the implementation of a variety of machine learning, variable selection and shrinkage methods
(e.g., Kim and Swanson, 2014, 2018; Li and Chen, 2014; Smeekes and Wijler, 2018)

▶ The method of **random forest (RF)** regression (Breiman, 2001) is particularly popular due to

- its wide applicability,
- allowance for nonlinearity in data,
- adaptability to high-dimensional data (e.g., Biau and Scornet, 2016; Medeiros et al., 2021),
- completely immune from overfitting (Goulet Coulombe, 2020),
- RF can dominate other machine learning and factor models in a high-dimensional context (e.g., Borup et al., 2020; Fortin-Gagnon et al., 2020; Medeiros et al., 2021)

▶ Nevertheless, some studies emphasize the need to select a reduced number of predictors prior to implementing RF.

- Borup et al. (2020) propose the **targeted RF (TRF)**, namely RF with an initial targeting step, where targeting is achieved via Lasso regularization
- Medeiros et al. (2021) consider another approach by combining RF with adaptive Lasso (Zhou, 2006), called **adaLasso/RF** or **hybrid method** (Masini et al, 2021)

▶ Targeting of predictors has been applied in various data-rich environment with penalized regressions

- dynamic factor models (Bai and Ng, 2008; Schumacher, 2010; Bessec, 2013)
- complete subset regressions (Kotchoni et al., 2019; Borup and Schütte, 2020)

▶ Other approaches for reducing the dimension are screening techniques such as the Sure Independence Screening (SIS) proposed by Fan and Lv (2008) and combined with penalized regressions (Zou and Zhang, 2009; Ferrara and Simoni, 2019).

▶ In this paper we compare some methods for variable selection and dimension reduction before to estimate RF with only the regressors selected, called **sparse random forest (sRF)**:

- soft thresholding:
  - Lasso regularized method
  - Elastic-Net regularized method
  - (Adaptive Lasso regularized method)
- hard thresholding:
  - the t-stat from an univariate predictive regression
  - the Granger non-causality test from a bivariate VAR
- dimensionality reduction by the SIS approach

Machine learning (ML)

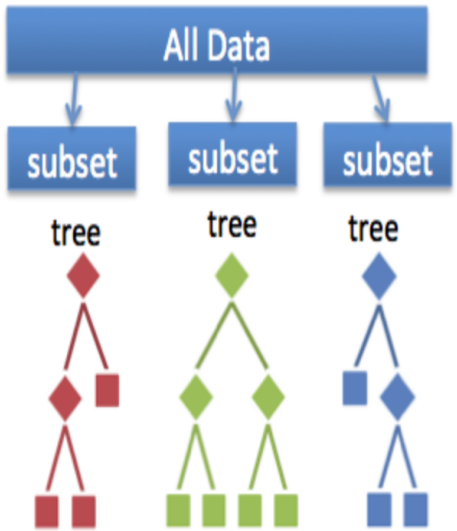
Unsupervised ML: only based on explicative variables X

Supervised ML: based on explicative variables X and their influence on the target Y

- Ensemble methods (Ensemble learning)
  - Sequential methods
    - Boosting
    - Adaptive Boosting
    - XGBoosting
  - Parallel methods
    - Bagging (*bootstrap aggregating*) (linear)
    - Random forest: bagging procedure that aggregates many uncorrelated decision trees (nonlinear)
- Parametric methods
  - Penalized regressions
  - Regressions on principal components
  - Sparse principal components

Random forest

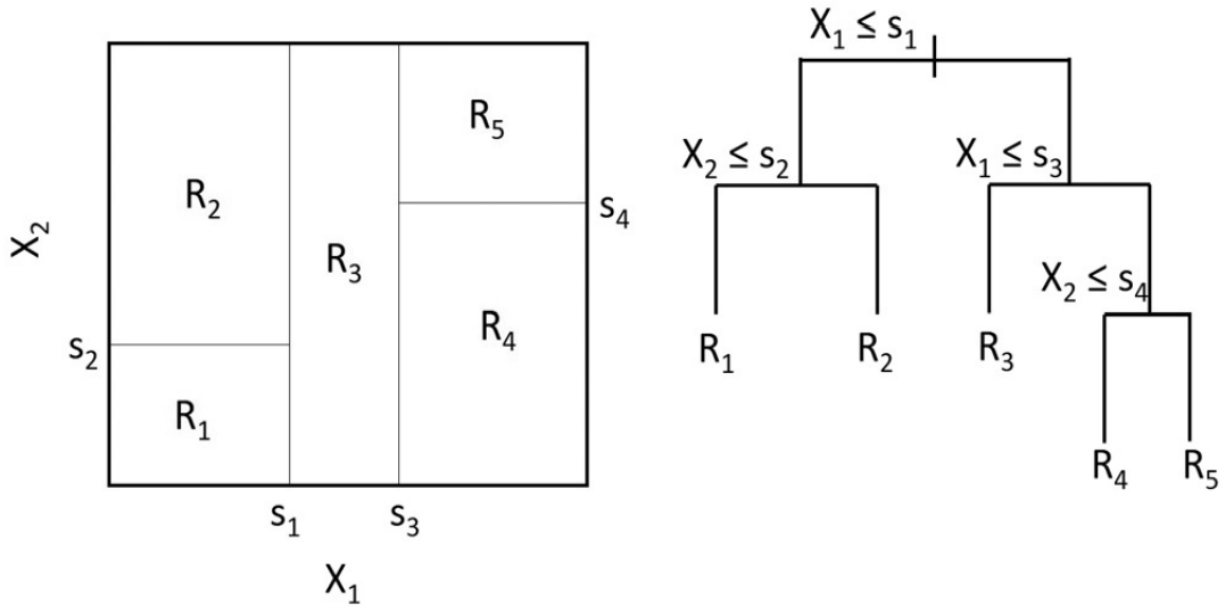
Random Forest (RF) is an ensemble learning method based on classification and regression trees (Breiman, 2001).
RF allows to reduce the variance of regression trees and is based on bootstrap aggregation (bagging) of randomly constructed regression trees.
This approach combines thus several randomized decision trees and aggregates their predictions by averaging.



A random forest takes a random subset of features from the data, and creates n random trees from each subset. Trees are aggregated together at end.

Consider a regression with only two predictors $X = (X_1, X_2)$, where each predictor X_i takes values in some given interval, and Y is the dependent variable.

Figure: Example of a regression tree (cf. Hastie et al., 2001)



The final result is a partitioning into five regions R_k, $k = 1, \ldots, 5$, suggested to estimate the conditional expected value (prediction) of Y.
Each region corresponds to a terminal node of the tree.

Given a dependent variable Y_{t+h}, a set of predictors X_t, and a number of terminal nodes K, the splits are determined to minimize the sum of squared errors of the following regression model

$$Y_{t+h} = \sum_{k=1}^{K} c_k I_k(X_t; \theta_k)$$

- $I_k(X_t; \theta_k)$: an indicator function such that

$$I_k(X_t; \theta_k) = \begin{cases} 1 & \text{if } X_t \in R_k(\theta_k) \\ 0 & \text{otherwise} \end{cases}$$

- θ_k: the set of parameters defining the kth region
- c_k: node means, i.e. $c_k = \sum_{j \in R_k} Y_j / N_k$, with N_k the number of variables in the kth region

For each bootstrap sample b, with $b = 1, \ldots, B$, a tree with K_b regions is estimated for a randomly selected subset of the original regressors.

K_b is determined to leave a minimum number of observations in each region.

The final forecast is the average of the forecasts of each tree applied to the original data

$$\widehat{Y}_{t+h} = \frac{1}{B} \sum_{b=1}^{B} \left(\sum_{k=1}^{K_b} \widehat{c}_{k,b} I_{k,b}(X_t; \widehat{\theta}_{k,b}) \right)$$

Practice

- At each cell of the tree, a split is performed by maximization of the CART-criterion (Classification And Regression Trees) by selecting *mtry* variables randomly among the p original ones, $mtry \in \{1, \ldots, p\}$;
- The minimum number of observations in every terminal node is set to 5;
- The number of bootstrap samples $B = 500$;
- We use the RF algorithm implemented in the `randomForest` package in R and the parameter *mtry* is estimated during the validation phase by implementing the `caret` package in R.

Soft thresholding

Lasso regularization. Lasso (*least absolute shrinkage and selection operator*) (Tibshirani, 1996) regression is characterized by an L_1 penalty function, allowing for sparsity.

The estimated coefficients of shrinkage estimator are given by

$$\widehat{\beta} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- where λ is the penalty parameter (or tuning parameter)
- λ is selected via cross-validation
- selection bias between highly correlated variables
- non oracle property

Soft thresholding

Elastic-Net regularization. Zou and Hastie (2005) develop the Elastic Net (EN) regularization method, which is a generalization including Lasso and Ridge as special cases, to avoid collinearity.

$$\widehat{\beta} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right)$$

- where $\alpha \in [0;1]$
- If $\alpha = 1 \Rightarrow$ EN = Ridge
- If $\alpha = 0 \Rightarrow$ EN = Lasso
- property of grouped selection or the 'grouping effec'
- non oracle property
- λ is selected via cross-validation
- α is chosen sequentially with $0 < \alpha < 1$

Hard thresholding

Univariate predictive regression. As in Kotchoni et al. (2019) and Fortin-Gagnon et al. (2020), we estimate an univariate predictive regression for each predictor $X_{i,t}$:

$$y_{t+1} = \alpha + \sum_{j=0}^{3} \rho_j y_{t-j} + \beta_i X_{i,t} + \varepsilon_t$$

The subset of relevant predictors X_t^* is obtained by gathering those series whose coefficients β_i have their t-stat lower than the critical value

$$t_c = 1.65 : \; X_t^* = \{X_i \in X_t | t_{Xi} > t_c\}$$

Hard thresholding

Bivariate VAR model. As in Martínez-Martín and Rusticelli (2020), we estimate bivariate VAR model for each variable $X_{i,t}$:

$$\begin{cases} y_t = \alpha_0^1 + \sum_{j=1}^{p} \rho_j^1 y_{t-j} + \sum_{j=1}^{p} \beta_j^1 X_{i,t-j} + \varepsilon_{1,t} \\ X_{i,t} = \alpha_0^2 + \sum_{j=1}^{p} \rho_j^2 X_{i,t-j} + \sum_{j=1}^{p} \beta_j^2 y_{t-j} + \varepsilon_{2,t} \end{cases}$$

where the lag order p is fixed according to AIC criteria.

We obtain the subset of relevant predictors X_t^* by gathering those series whose the p-values for Wald tests of Granger non-causality based on heteroskedasticity-robust variance estimators are lower than the 10% level.

SIS approach

Fan and Lv (2008) propose an approach referred to as Sure Independence Screening (SIS hereafter). Sure screening refers to the property that "*all important variables survive after applying a variable screening procedure with probability tending to 1*"

Let $\omega := (\omega_1, \ldots, \omega_p)'$ the vector of marginal correlations of predictors with the response variable Y_t, such as

$$\omega = X'Y$$

- where the $(N \times p)$ matrix X is first standardized-columnwise.
- ω is thus a vector of marginal correlations of predictors with the response variable, rescaled by the standard deviation of the response

For any given $\lambda \in]0;1[$ we sort the p componentwise magnitudes of the vector ω in a decreasing order and define a submodel $\mathcal{M}(\lambda)$ such as

$$\mathcal{M}(\lambda) = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } \lfloor \lambda N \rfloor \text{ largest of all}\}$$

- where $\lfloor \lambda N \rfloor$ denotes the integer part of λN
- SIS $\Rightarrow$ shrink the full model $\{1, \ldots, p\}$ down to a submodel $\mathcal{M}(\lambda)$ with size $d = \lfloor \lambda N \rfloor < N$

Data

- Large monthly macroeconomic datasets which consist of US variables from the FRED-MD database following McCracken and Ng (2016)
  - output and income (Group 1);
  - labor market (Group 2);
  - housing (Group 3);
  - consumption, orders and inventories (Group 4);
  - money and credit (Group 5);
  - interest rates and exchange rates (Group 6);
  - prices (Group 7) and
  - stock market (Group 8)

- We restrict the sample period to 1970-2019, as most series become available from 1970, and remove those with missing values during the period $\Rightarrow p = 125$ variables

- We also consider four autoregressive terms but not lags of all variables or principal component factors as in Medeiros et al. (2021).

- All variables are transformed to achieve stationarity as proposed by McCracken and Ng (2016)

Target variable: inflation

We only focus on the consumer price index (CPI) prediction since RF appears to be relevant for inflation forecasting (Medeiros et al., 2021).

The inflation is computed as $Y_t = \log(P_t) - \log(P_{t-1})$, where P_t is the CPI in period t.

- Fortin-Gagnon et al. (2020) and Medeiros et al. (2021) consider the consumer price index (CPI) as an I(1) series
- Kotchoni et al. (2019) and Borup et al. (2020) as an I(2) series, i.e. CPI acceleration

Variable selection

All models are estimated on seven subsamples with the same number of observations ($T = 180$ observations) to compare the number of selected variables across the various approaches of dimension reduction along the time:

- 1970-1984,
- 1975-1989,
- 1980-1994,
- 1985-1999,
- 1990-2004,
- 1995-2009,
- 2000-2014.

Table: The percentage of selected variables.

Period	Lasso	EN	SIS	Granger	t-stat
1970-1984	17%	28%	50%	29%	33%
1975-1989	19%	21%	50%	28%	36%
1980-1994	**5%**	**8%**	51%	22%	35%
1985-1999	16%	36%	51%	13%	18%
1990-2004	21%	24%	51%	17%	18%
1995-2009	13%	19%	51%	36%	36%
2000-2014	19%	29%	51%	28%	33%
Average	**16%**	23%	51%	23%	30%
S.D.	5.5%	9.1%	0.4%	6.3%	8.3%

- Number of selected variables is relatively stable across horizons
- High volatil for EN vs. low volatil for SIS
- The soft-thresholding selection is sparse
- EN exhibits the property of grouped selection

Comparison of variable selection.

Period	Total	Lasso/EN	Lasso/SIS	Lasso/Granger	EN/SIS	EN/Granger
1970-1984	5%	95%	73%	41%	69%	33%
1975-1989	5%	88%	68%	36%	81%	41%
1980-1994	3%	100%	100%	67%	100%	60%
1985-1999	2%	100%	81%	29%	72%	21%
1990-2004	3%	100%	81%	33%	84%	29%
1995-2009	3%	100%	71%	35%	75%	50%
2000-2014	5%	96%	71%	46%	62%	30%
Average	**4%**	97%	78%	41%	78%	38%
S.D.	1.0%	4.5%	11%	13%	12%	13%

Period	Granger/SIS	Lasso/t-stat	EN/t-stat	t-stat/SIS	Granger/t-stat
1970-1984	59%	59%	39%	63%	51%
1975-1989	64%	56%	63%	72%	51%
1980-1994	71%	100%	100%	76%	40%
1985-1999	82%	52%	32%	83%	30%
1990-2004	86%	37%	32%	70%	35%
1995-2009	53%	76%	71%	83%	37%
2000-2014	61%	79%	59%	77%	35%
Average	68%	66%	57%	75%	40%
S.D.	12%	21%	25%	7.1%	8.2%

The most frequently selected variables are the order-one lagged inflation (5 subperiods), Real M2 Money Stock (4 subperiods) and Moody's Baa Corporate Bond Minus FEDFUNDS (3 subperiods).

Forecasting exercise

Forecasting design

- Out-of-sample rolling forecasts
- The out-of-sample window is
  - from January 1985 to December 1989 for the first subperiod
  - from January 1990 to December 1994 for the second subperiod
  - and so on ...
- producing $H = 60$ forecasts for each subsamples
- The forecasting horizon considered is one month ahead ($h = 1$)
- RF and sRF models are re-estimated at each step
- estimation sample size remains fixed and the forecasts do not overlap

Forecasting evaluation

- the mean squared error (MSE)
- the out-of-sample (OOS) R^2: $R^2_{OOS} = 1 - (MSE_{sRF}/MSE_{RF})$
- the equal predictive ability (EPA) tests of Diebold and Mariano (1995),
- the superior predictive ability (SPA) tests of Hansen (2005) and
- the model confidence sets (MCS) of Hansen, Lunde, and Nason (2011)

The results in terms of MSE show that the sRF models based on soft-thresholding and SIS approach produce MSE better than those of the full RF in most subperiods, except the 2010-2014 and 2015-2019 subperiods.

This result is confirmed by the R^2_{OOS} because these sRF models display a positive value.

The smallest MSE are exhibited by the EN regularization whatever the subperiod, except for the 2010-2014 subperiod.

The sRF models based on hard-thresholding give the highest MSE and negative R^2_{OOS}, suggesting that this pre-selection approach does not appear relevant for the sRF models.

Work in progress

Applying the EPA, SPA and MCS tests

Applying other dimensionality reduction methods, such as:

- generalized least squares screening (GLSS) of Yousuf (2018)
- distance correlation screening (DC-SIS) of Li et al. (JASA, 2012)
- partial distance correlation Screening (PDC-SIS) approach of Yousuf and Feng (JBES, 2021)