Predicting Foreign Exchange Arbitrage

Stefan Huber
sjhuber

Amy Wang
amyywang

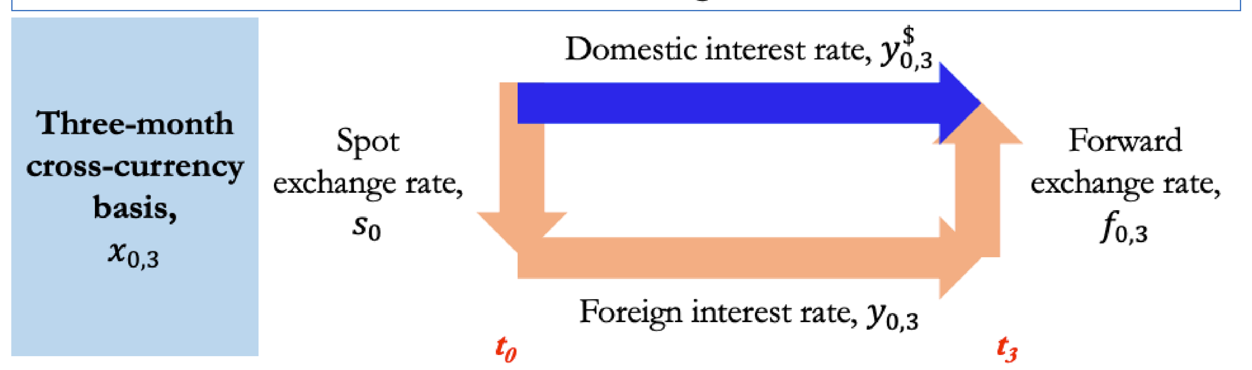
Introduction

The Covered Interest Parity condition ("CIP") should dictate prices on the trillion-dollar foreign exchange market. Yet the CIP fails. When CIP fails, cross-currency basis ("basis") appears. As illustrated below, a non-zero basis is an arbitrage opportunity, as all quantities in the basis are known today and a positive pay-off can be guaranteed without any cash outflow.

Our project seeks to predict the bases in the Post-Financial Crisis period of the Australian dollar ("AUD") and of the Japanese yen ("JPY"), both relative to the US dollar ("USD").

Employing Regression Trees, Random Forests, and three versions of Regularized Regressions, we predict the bases with reasonable success: our best test MSEs of ~70 for AUD and ~200-270 for JPY compare favorably to the standard deviation of the Post-Crisis bases, which are 11.2 and 23.8 for AUD and JPY, respectively.

$$x_{0,3} = y^{\$}_{0,3} - y_{0,3} - \frac{1}{3}(s_0 - f_{0,3})$$



Data

We collect 32 data series for each of the three relevant currencies: AUD, JPY, and USD. Our data capture activities in the financial markets, conditions of the economy, the state of international trade, and the stance of economic policies. With the exception of the Economic Policy Uncertainty ("EPU") Index, which is compiled by three leading economists and available at www.policyuncertainty.com, all data are obtained from Bloomberg:

Financial Markets	Economic Conditions	International Trade	Economic Policy
• *15 daily series*	• *7 monthly or quarterly series*	• *5 monthly or quarterly series*	• *5 monthly or quarterly series*
• Stock market indices	• GDP	• Current account balance	• Monetary base
(overall and by sector)	• CPI	• International trade balance	• Broad liquidity
• Stock trading volumes	• Unemployment rate	• Export	• Central bank asset
• 3-month interbank rate	• Labor force participation	• Import	• Government debt
• 10-year government bond yield	• Hourly wage	• International reserve balance	• Economic Policy Uncertainty Index
	• Mortgage outstanding		
	• Loans to bank		

Raw data are processed to form the final Train and Test sets following the procedures in the chart below. Note that before augmenting the data, we compute and use the percentage change instead of the original level for most of the data series. This is done to both normalize and extract more meaningful information.

Ultimately, each of our models is applied to 8 distinct Train/Test sets: for each of the AUD and JPY bases, we have either the Complete or the Post-Crisis sample, and within each sample, we split Train vs. Test using either a contiguous or a random approach.

Data Augmentation

- **Impute** daily observations based on the last available observation in low frequency series.
- **Expand** feature space in regularized regressions by squaring all features.

Sample Construction

		Complete	Post-Crisis
Observations		Longest possible horizon: 2004 - 2017	Post financial crisis only: 2010 - 2017
m	*AUD*	3,312	1,885
	JPY	3,309	1,884
Features (before squaring)		Features with complete obs only.	All available features.
n	*AUD*	63	64
	JPY	58	64

Train vs. Test Split

- Two methods to subset Test in both samples.
- **Contiguous** split: 200 obs at the end + 200 obs in the middle of Post-Crisis.
- **Random** split: 400 obs chosen at random in Post-Crisis.

Model

Regression Tree

- Impurity in node j with mean μ_j: $D_j = \sum_{i \in j} (y^{(i)} - \mu_j)^2$

Random Forest

- 2000 trees with a minimum of 5 leaves
- Size of variable subset considered chosen by OOB prediction error

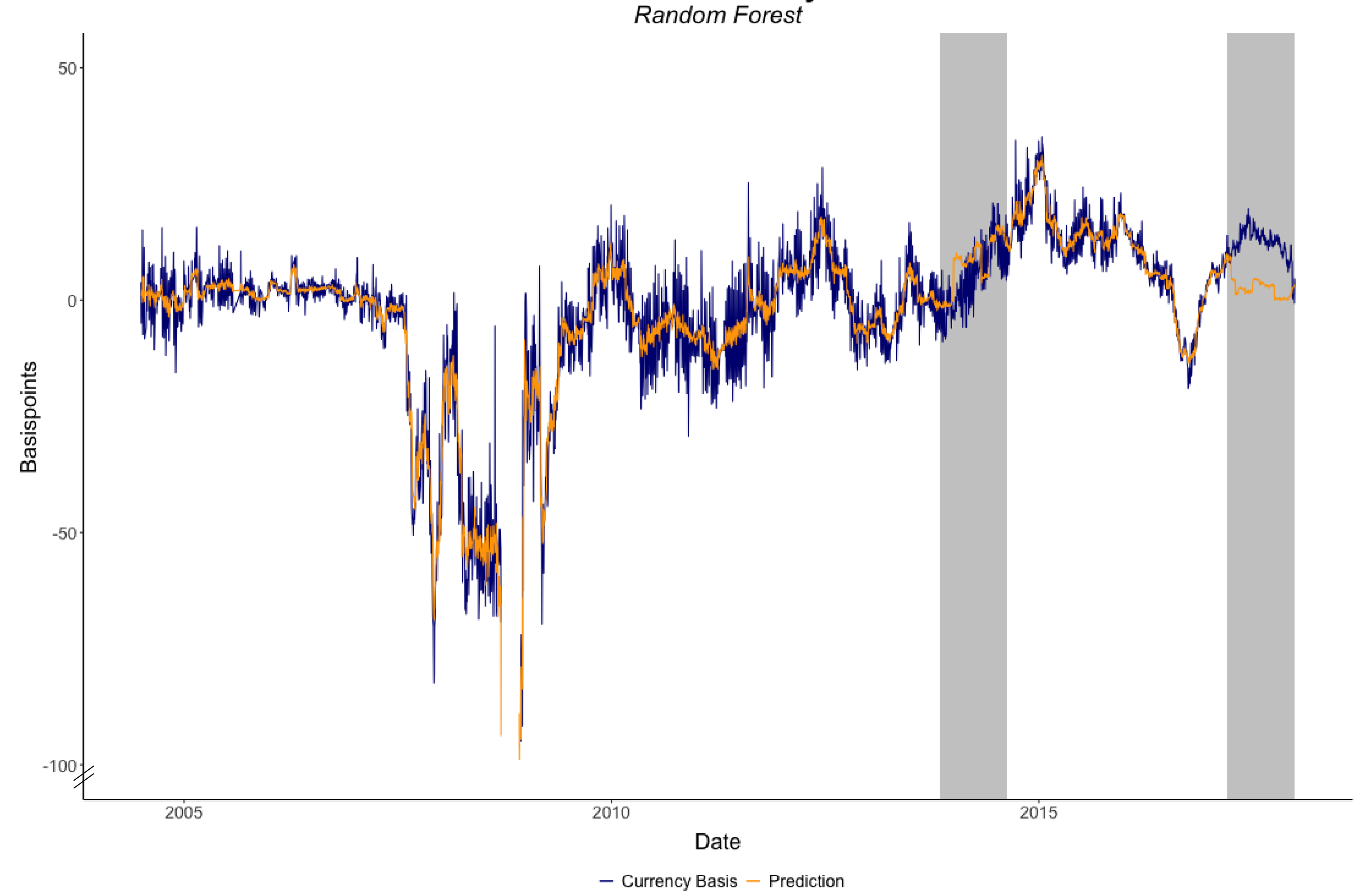
Regularized Regression

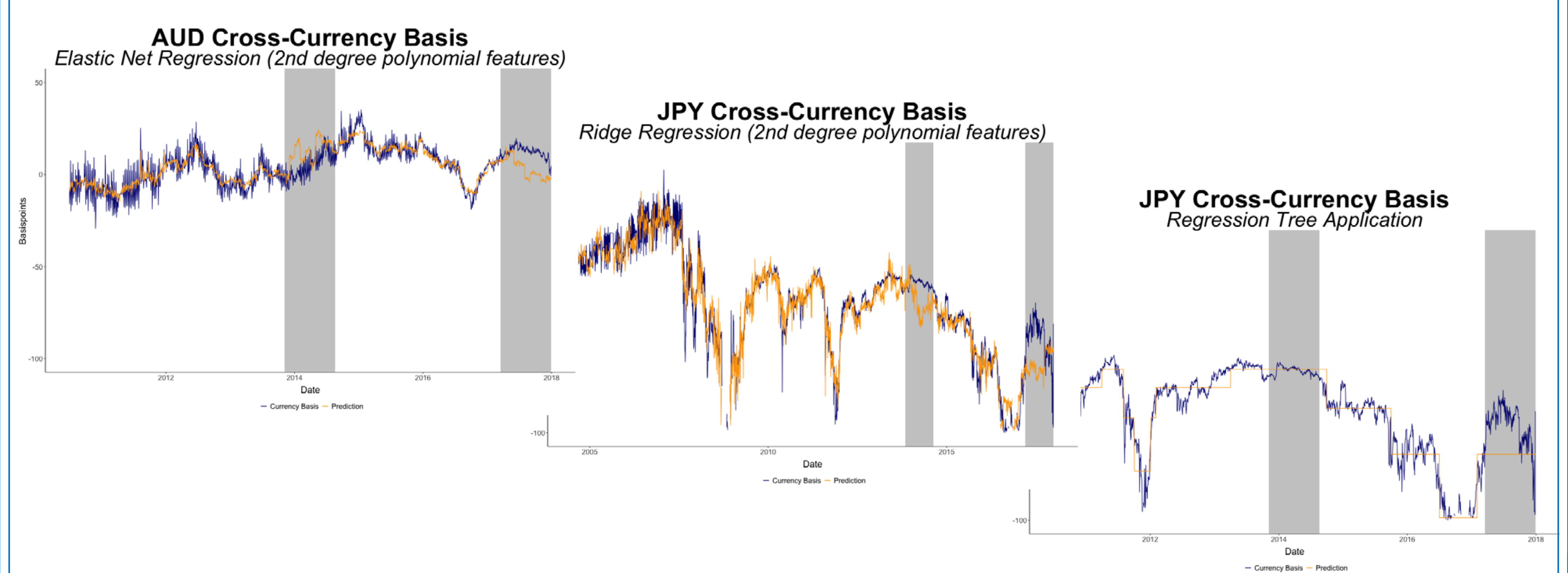
$$\min_{\theta} \sum_{\{i=1\}}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \lambda(\alpha\|\theta\|_2^2 + (1-\alpha)\|\alpha\|_1)$$

- Lasso: $\alpha = 0$, Ridge: $\alpha = 1$, Elastic Net: $\alpha = 0.5$
- Choice of λ via 10-fold cross-validation

Results

AUD Cross-Currency Basis
Random Forest





Summary of MSEs from ML Algorithms on Contiguous Split

	Tree	Forest	Lasso		Ridge		Elastic Net	
			Linear	Polynomial	Linear	Polynomial	Linear	Polynomial
Test MSE								
AUD Complete	222	***69***	655	282	1231	189	726	208
AUD Post-Crisis	333	***70***	247	115	140	154	238	101
JPY Complete	1234	***274***	394	555	314	295	376	527
JPY Post-Crisis	***187***	337	908	1162	820	678	899	1049
Train MSE								
AUD Complete	130	75	242	127	255	120	240	126
AUD Post-Crisis	27	21	33	21	37	38	33	21
JPY Complete	217	84	367	119	405	110	364	121
JPY Post-Crisis	51	14	55	24	67	70	56	24

Note: for algorithms trained on the Contiguous Split, the reported MSE is the average between the MSEs on the two contiguous test blocks.

Legend ***Lowest*** Graphed

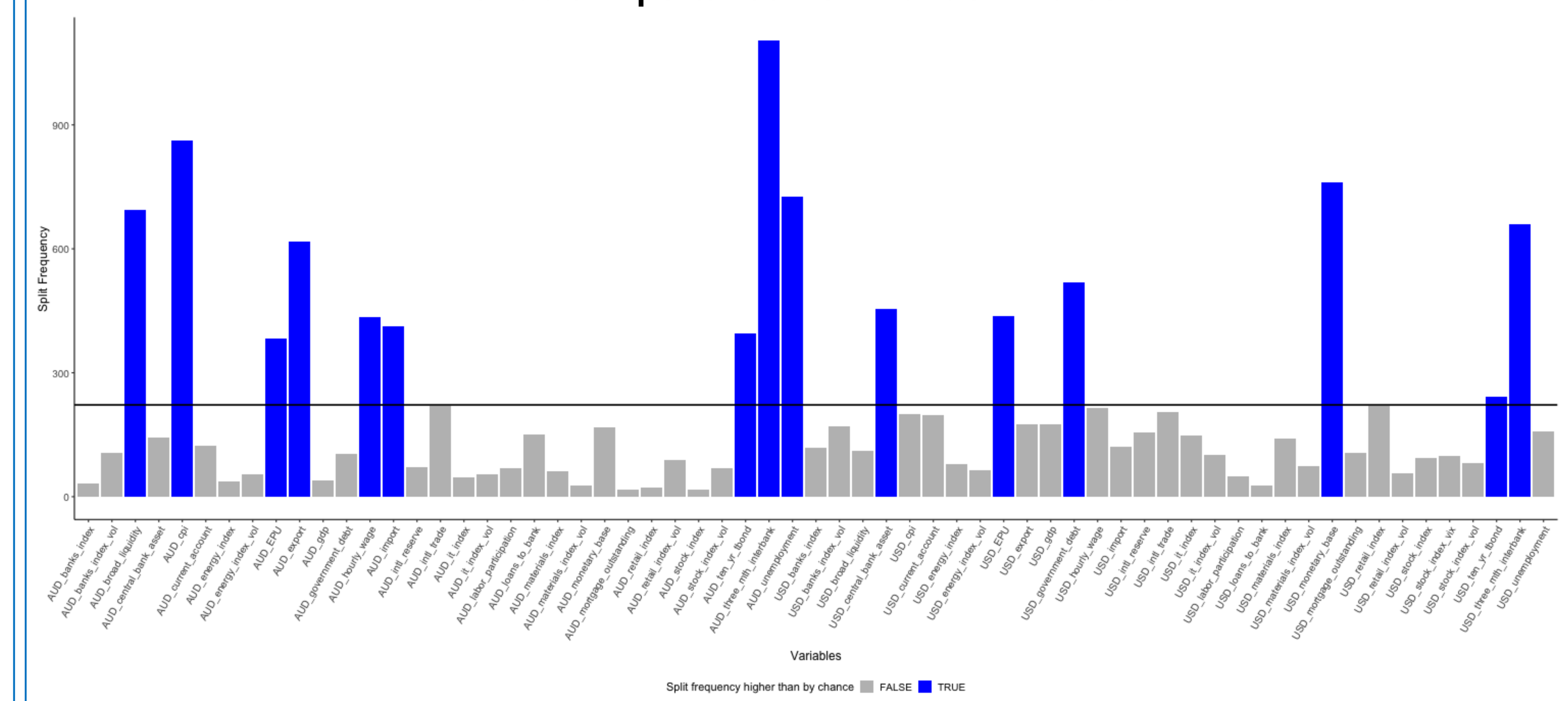
Discussions

For the AUD basis, the regression tree and the regularized regressions perform similarly. Comparing the MSEs from the two specifications of regularized regressions (linear vs. polynomial feature space), we note that the prediction error in the AUD basis is likely caused by a **bias** problem, as the MSE decreases with the inclusion of the higher-dimensional features in both the training and the test set. In predicting the JPY basis, we find indications of a **variance** problem, i.e. overfitting, since the polynomial improves the training error but increases the test error.

Overall, the **random forest** algorithm delivers the best performance, with the lowest train and test MSE. Note that in the second contiguous test period (2017), no algorithm seems to fit very well. One potential reason is that the cross-currency bases were on an elevated level with low variance, a state that has not been observed in the train data.

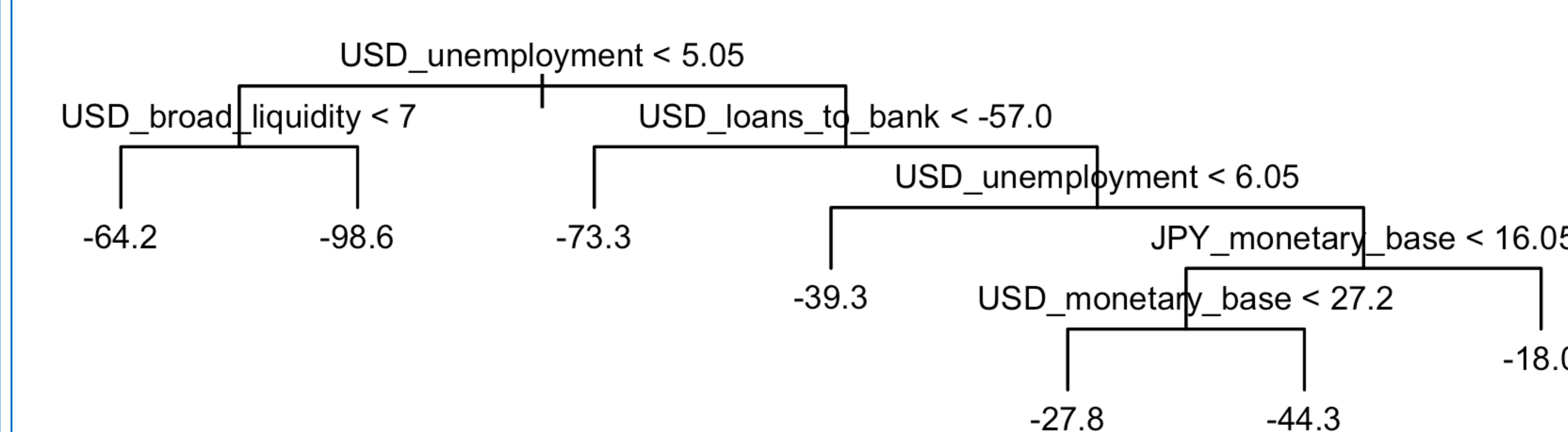
We also explore the set of variables with the most predictive power. The figure below shows how often the trees in the random forest (as applied to the Complete AUD sample) split on each of the considered features within the first seven splits. The horizontal line indicates how often a feature would appear if no feature had predictive power.

Variable Importance AUD Random Forest



In the chart below, we exemplarily show a fitted tree for the JPY basis. Interestingly, the JPY basis is predicted by mostly US features; this contrasts with the prediction of AUD basis, which relies on both Australian and US features. This raises the question whether we have included the appropriate set of features for JPY basis.

Tree for JPY Cross-Currency Basis



Future Work

Given the observed bias issue with the AUD data, we would collect more economic features and use higher order polynomial features in the regularized linear regressions in the future.

To improve the performance on the JPY data, we will expand the set of algorithms employed. Specifically, we will apply the boosting technique, and we will consider training a neural network.

Finally, we want to extend the analysis to a larger set of currency bases.

References

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131(4), 1593-1636.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent, https://web.stanford.edu/~hastie/Papers/glmnet.pdf
Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010

Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 77:1-17. http://dx.doi.org/10.18637/jss.v077.i01.