Improved Out-of-sample Equity Premium Forecasts with Many Mixed-root Predictors

SONGYEN CHEN

ID: D08723005

Dept. of Finance, NTU

December 6, 2024

Contents

Prologue

Elusive predictability of equity premium return ?

- **Predictability** Early Evidence emerged that stock returns are in-sample predictable even the weak predictability is revealed, e.g., Campbell and Shiller (1988), Fama and French (1988, 1989), Cochrane (2008), and Timmermann (2008).

- **Unpredictability** Goyal and Welch (2008) and Goyal et al. (2024) demonstrate that the historical mean benchmark is hard to beat in out-of-sample forecasting.

- Many works consider various predictors and methods to evidence the equity premium return forecasting. For examples, **(economic constraints)** Campbell and Thompson (2008), **(forecast combination)** Rapach, et al. (2010), **(option data)** Alexandridis et al. (2023), **(text data)** Lima and Godeiro (2022),

A fundamental fact about financial predictors:

- Many financial predictors are mixed-root, with nearly non-stationary $I(1)$, but not purely $I(1)$.
- But, the equity premium return process is weakly stationary $I(0)$.
- This results in an unbalance of integration order in predictive regression

$$r_{t+1} = \alpha + \beta X_t + u_{t+1},$$

forcing the true value of β to be zero, or very small if not zero (Phillips, 2015). The true predictability may be hidden.

- In large-dimensional scenarios with many predictors, this may lead to **inconsistent variable selection** with the conventional model selection techniques, e.g., BIC, LOOCV, LASSO since incorrect zero coefficients may be excluded.

Var.	Mean	S.D.	P25	P50	P75	Sk.	Kur.	ADF.pv	PP.pv	$AC_{(1)}$
$r(\%)$	0.20	4.37	-2.14	0.64	3.03	-0.67	5.29	0.01	0.01	0.02
DE	-0.77	0.31	-0.94	-0.81	-0.62	2.93	9.48	0.01	0.01	0.99
DY	-3.64	0.41	-3.96	-3.56	-3.35	0.01	2.10	0.38	0.41	0.99
DP	-3.64	0.44	-3.97	-3.57	-3.36	0.02	2.08	0.35	0.42	0.99
TBL(%)	4.46	3.26	1.69	4.64	6.13	0.66	3.63	0.05	0.20	0.99
EP	-2.87	0.43	-3.11	-2.90	-2.67	-0.51	5.43	0.01	0.02	0.99
BM	0.47	0.26	0.27	0.39	0.64	0.90	2.83	0.48	0.52	0.99
LTY(%)	6.22	2.89	4.23	5.99	8.02	0.51	2.98	0.44	0.43	0.99
LTR(%)	0.56	3.02	-1.22	0.38	2.31	0.37	5.24	0.01	0.01	0.06
INFL(%)	0.32	0.36	0.10	0.30	0.52	0.01	5.85	0.01	0.01	0.58
DFY(%)	1.02	0.44	0.73	0.91	1.20	1.74	7.45	0.01	0.01	0.96
DFR(%)	-5.62	3.90	-7.77	-5.54	-3.36	-0.28	4.86	0.01	0.01	0.53
NTIS	0.01	0.02	-0.01	0.01	0.02	-0.48	2.91	0.01	0.02	0.98
TMS	0.02	0.01	0.01	0.02	0.03	-0.24	2.70	0.04	0.01	0.96
RVOL	0.14	0.06	0.10	0.13	0.18	0.46	3.47	0.01	0.01	0.98
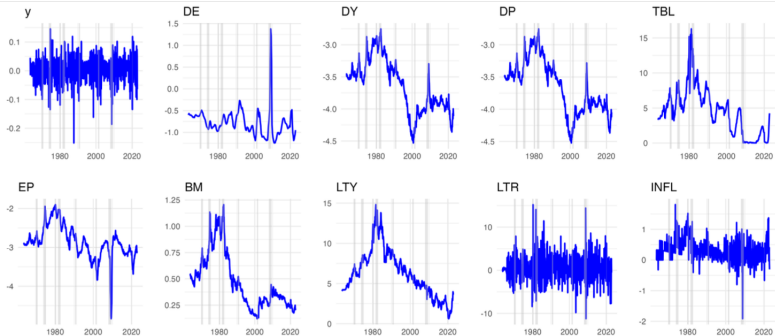
- Why do we not take the first differences of the predictors?

  "*In empirical finance, many financial predictors or factors are proxied and constructed through somewhat linear or nonlinear transformations based on specific financial theories and models, e.g., dividend-price ratio (Gordon growth theory), Tobin Q ratio (Tobin's Q theory), value premium (HML) (3-factor model), momentum (4-factor model),..., etc.*

  As such, we **avoid using differencing filters**, *as they can result in the loss of the original information contained in these variables*" (suggested by prof. Ethan Chiang (UNC)).

- Chen (2024) shows that such unbalance of integration orders has a **negative** effect on out-of-sample forecasts in terms of MSFEs.
    - r_t is weakly stationary $I(0)$, and X_t is nearly unit-root or purely $I(1)$,

$$r_{t+1} = \alpha + \beta X_t + u_{t+1}, \tag{1}$$

$$X_{t+1} = (1 - c/T)X_t + v_{t+1},$$

    with $\beta = bT^{-\eta}$ and $c \geq 0$,
    - $(u_t, v_t)^\top \sim$ white noise $\mathcal{N}(0, \Sigma)$,

$$\Sigma := \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{vu} & \sigma_v^2 \end{bmatrix},$$

    and $u_{t+1} = (\sigma_{uv}/\sigma_v^2)v_{t+1} + \epsilon_{t+1}$, with $\mathbb{E}[\epsilon_{t+1}|v_{t+1}] = 0$.
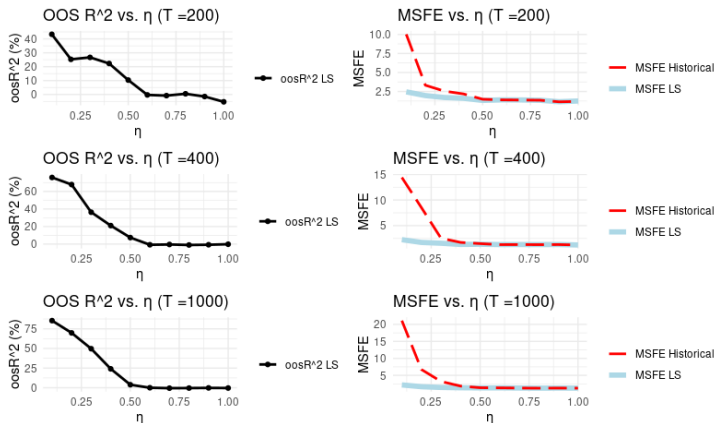
Illustration (Chen, 2024)

Figure: The out-of-sample R^2 of recursive LS estimates, and the corresponding MSFEs of recursive LS estimates and historical mean benchmark, respectively. The DGP is set up by: $r_{t+1} = 0.3 + \beta_T X_t + u_{t+1}$, $X_{t+1} = (1 - 1.1/T)X_t + v_{t+1}$, where $(u_t, v_t)^\top \sim$ white noise $\mathcal{N}(0, \Sigma)$, with $\Sigma := (1, -0.95; -0.95, 1)$. The true value $\beta_T = T^{-\eta}$. $T = \{200, 400, 1000\}$ and the number of replication is 1000. The recursive-window LS estimation with the first rolling window size is 180. The closer the value of β_T is to zero (fixed T), the more the behavior of r_t resembles the white noise u_t, which follows a weakly stationary $I(0)$.

- Proposition (Chen, 2024)

  - $\widehat{\beta}_{LS} = O_p(T^{-\eta}) = o_p(1)$.

  - $T\widehat{\beta}_{LS} \Rightarrow \frac{\sigma_{uv}^2}{\sigma_v^2} \frac{\int_0^1 J_v^c(s)dB_v(s)}{\int_0^1 J_v^c(s)^2} + \frac{\sigma_\epsilon}{\sigma_v} \frac{\int_0^1 J_v^c(s)dB_\epsilon(s)}{\int_0^1 J_v^c(s)^2}$, where $J_v^c(t)$ is a Ornstein–Uhlenbeck process with respect to $B_v(t)$, and $B_\xi(.)$ is a standard Brownian motion with respect to the stochastic process ξ.

  - $t(\beta) \Rightarrow \sqrt{(1-\gamma_{uv}^2)}Z + \gamma_{uv}\frac{\int_0^1 J_v^c(s)dB_v(s)}{\sqrt{\int_0^1 J_v^c(s)^2ds}}$, where $\gamma_{uv} := \sigma_{uv}/\sigma_u\sigma_v$ and Z is a standard Gaussian.

  - $\mathrm{MSFE}(\bar{r}_t) = \sigma_u^2 + \frac{1}{T}\sigma_u^2 + O(T^{1-2\eta})$,
    $\mathrm{MSFE}(\widehat{r}_{t+1|t}^{LS}) = \sigma_u^2 + \frac{K(c)}{T}\sigma_u^2 + \frac{\sigma_{uv}^2}{\sigma_v^2 T}\frac{4(1-e^{-2c})}{2c} + O(T^{-2\eta})$, where $K(c) \geq 1$, $\forall c > 0$, $\bar{Y}_t = t^{-1}\sum_{s=1}^t Y_s$ be the historical sample average, and $\widehat{Y}_{t+1|t}^{LS} = \widehat{\beta}_{LS}X_t$.

  -

$$\mathrm{MSFE}(\bar{r}_T) \leq \mathrm{MSFE}(\widehat{r}_{T+1|T}^{LS}), \quad \eta \geq 1/2$$
$$\mathrm{MSFE}(\bar{r}_T) > \mathrm{MSFE}(\widehat{r}_{T+1|T}^{LS}), \quad \eta < 1/2.$$

What this talk tries to do

- Chen, Chen, and Vincent (2024) argue that this may be induced from a problem of omitted variables, and propose a new robust model framework that incorporates highly persistent predictors with approximately co-integrations regardless of various levels of high persistence.

- A disadvantage of Chen, Chen, and Vincent (2024) is that the dimension of predictors k cannot be too large, particularly $k < T/2$.

- In this project, I extend the work of Chen et al. (2024) that a **c**ontinuously-updating **a**daptive **L**ASSO **m**echanism (CALM) is proposed to select the effective predictors for the model of Chen et al. (2024) when the number of mixed-root predictors is sufficiently large. From empirical evidence, the out-of-sample equity premium forecasts can be significantly improved.

- The idea is based on the FWL (Frisch–Waugh–Lovell) theorem and the adaptive LASSO (Zou, 2006).

Model Framework

- Let r_{t+1} be the equity premium return at $t+1$, $t. = 1, 2, \cdots, T$, and denote the potential predictors $\boldsymbol{X}_t := (\boldsymbol{X}_{1t}, \boldsymbol{X}_{2t}, \cdots, \boldsymbol{X}_{kt})^\top \in \mathcal{X} \subseteq \mathbb{R}^k$, $k < T/2$, $\mathcal{X} := \mathcal{X}^U \uplus \mathcal{X}^N$, $\mathcal{X}^U \cap \mathcal{X}^N = \phi$, where $\mathcal{X}^U$ is the set of m $I(1)$ predictors, and $\mathcal{X}^N$ is the set of non-$I(1)$ predictors, including those nearly $I(1)$ predictors and weakly stationary $I(0)$ predictors.

- DGP:

$$\begin{pmatrix} r_{t+1} \\ \boldsymbol{X}_{t+1} \end{pmatrix} = \begin{pmatrix} \mu_o \\ \boldsymbol{\rho}_o \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \\ \boldsymbol{\rho} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_t \\ \boldsymbol{X}_{t-1} \end{pmatrix} + \begin{pmatrix} u_{t+1} \\ \boldsymbol{v}_{t+1} \end{pmatrix},$$

where $(u_{t+1}, \boldsymbol{v}_{t+1}^\top)^\top \sim$ martingale difference sequences with possibly nonzero contemporaneous correlation.

- VAR (1) coefficient matrix (e.g., Phillips, 2023):

$$\boldsymbol{\rho} = \boldsymbol{I}_k - \boldsymbol{C}/T^s, \ \boldsymbol{C} := \text{diag}(c_1, c_2, \cdots, c_k),$$

for all $c_i > 0$, and $s \geq 0$. This allows for various levels of persistence.

- VEC representation for $CI(1,1)$ among $I(1)$ predictors: $\boldsymbol{X}_t^U \in \mathcal{X}^U$,

$$\Delta \boldsymbol{X}_{t+1}^U = \boldsymbol{\Gamma} \boldsymbol{X}_t^U + \boldsymbol{\nu}_{t+1}^U, \ \boldsymbol{\Gamma} = \boldsymbol{B} \boldsymbol{A}^\top,$$

with $rank(\boldsymbol{\Gamma}) = r < m$, $\boldsymbol{B}$ is the $m \times r$ adjusting factor, representing that how predictors adjusts deviations back toward the long-run equilibrium, and $\boldsymbol{A} = (\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_r)$ is $m \times r$ co-integrating matrix with r co-integrating vectors which captures the long-run equilibrium.

Model specification

Similar to Chen et al. (2024), a constrained framework can be constructed by:

$$r_{t+1} = \boldsymbol{\mu}_o + \boldsymbol{X}_t^\top \boldsymbol{\beta}_1 + \boldsymbol{X}_{t-1}^\top \boldsymbol{\beta}_2 + e_{t+1}, \tag{2}$$

subject to two constraints:

1. (contemporaneous co-integrations)

$$\boldsymbol{R}\boldsymbol{\beta}_{1m} = \boldsymbol{0}_{(m-1)\times 1}, \ \boldsymbol{R} = (\widehat{\boldsymbol{\theta}}_{(m-1)\times 1}, \boldsymbol{I}_{m-1})_{(m-1)\times m}, \tag{3}$$

   where $\widehat{\boldsymbol{\Theta}} =: (1, -\widehat{\boldsymbol{\theta}}^\top)^\top$ is the $m \times 1$ normalizing estimated co-integrating vector with largest eigenvalue, among m $I(1)$ predictors, which are identified by the ADF and PP tests and then selected by Johansen's standard procedure;

2. (approximately intertemporal co-integrations)

$$\boldsymbol{\beta}_2 + \widehat{\boldsymbol{\rho}}^\top \boldsymbol{\beta}_1 = \boldsymbol{0}_{k\times 1}, \tag{4}$$

   where $\widehat{\boldsymbol{\rho}}$ is the VAR (1) estimator for $\boldsymbol{\rho}$, componentwise.

CALM algorithm

Given $\widehat{\alpha}_j^{(0)}$ is the initial LS or ridge estimator. At the ith iteration: for $i \geq 2$,

-
$$(\widehat{\boldsymbol{\mu}}^{(i)^\top}, \widehat{\boldsymbol{\alpha}}^{(i)^\top})^\top = \underset{\mu^{(i)} \in \mathbb{R}, \boldsymbol{\alpha}^{(i)} \in \mathbb{R}^{k_{(i)}}}{\operatorname{argmin}} \left\{ ||r_{t+1} - \ddot{\boldsymbol{v}}_t^\top \boldsymbol{\alpha}^{(i)}||_2^2 + \lambda^{(i)} \sum_{j=1}^{k_{(i)}} \widehat{\tau}_j^{(i-1)} |\alpha_j^{(i)}| \right\},$$

  where $\ddot{\boldsymbol{v}}_t$ is the standardized residual from regressing $\boldsymbol{X}_t$ on $\boldsymbol{X}_{t-1}$, and $\widehat{\tau}_j^{(i-1)} = 1/|\widehat{\alpha}_j^{(i-1)}|$ is the adaptive penalty weight.

- Denote $\widehat{\mathcal{A}}_{(i)}^{\boldsymbol{X}} := \{X_{j,t} : \widehat{\alpha}_j^{(i)} \neq 0\}$ as the selected set of predictors at the ith iteration. Let $k_{(i)} := |\widehat{\mathcal{A}}_{(i)}^{\boldsymbol{X}}|$ be the cardinality of $\widehat{\mathcal{A}}_{(i)}^{\boldsymbol{X}}$ at the ith iteration. The iteration stops at $i = 2$ if $k_{(0)} < [T/2]$, and succeeds if $k_{(0)} \geq T/2$. When $k^* = k_{(i)} = k_{(i+1)} < [T/2]$, the iteration stops. Then, redo the estimation procedure of (2) to (4) with predictors of dimension k^*.

- $\lambda^{(i)}$ is the non-negative tuning parameter selected by the stationary time-series cross validation (TSCV) procedure proposed by Racine (2000).

Data and Predictors

- The equity premium return r_{t+1}: log excess return.
- The set of financial predictors: 14 predictors, conducted from Goyal and Welch (2008).
- The set of technical indices: 14 trading indices based on three types of trading strategies, following Neely et al. (2014).
- The set of anomaly portfolios: 516 predictors, conducted from Chen and Zimmermann (2022).
- Overall, total number of predictors X_t is 544.
- The sample period covers from 1963/7-2022/12, monthly frequency. The full sample length is: $T = 713$.
- Recursive-window estimation with the initial window size is 180 (15 years).

Table: Definitions of Financial Predictors

Predictor	Definition
DP	log dividends to price ratio of the S&P 500 index
DY	log dividend yield of the CRSP index
EP	log earnings to price ratio of the S&P 500 index
DE	log dividends to earnings ratio of the S&P 500 index
BM	book-to-market ratio of the Dow Jones Industrial Average Index
RVOL	square root of the sum of squared daily returns on the S&P 500 indexs
NTIS	aggregate net equity issues by the NYSE listed stocks
TBL	three-month Treasury-bill rate
LTY	long-term government bond yield
LTR	long-term government bond returns
TMS	long-term government bond yield minus the Treasury bill rate
DFY	difference between BAA and AAA-rated corporate bond yields
DFR	difference between corporate bond and long-term government bond returns
INFL	monthly rate of change of CPI

Empirical Evidence

Evaluation

1. Out-of-sample R^2 (Campbell and Thomason, 2008):

$$oosR^2(\%) := 1 - \frac{\sum_{t=R}^{T-1}(r_{t+1} - \widehat{r}_{t+1|t})^2}{\sum_{t=R}^{T-1}(r_{t+1} - \bar{r}_{t+1|t})^2} \in (-\infty, \infty),$$

where $\widehat{r}_{t+1|t}$ and $\bar{r}_{t+1|t} = \sum_{s=1}^{t} r_s/t$ are the conditional forecasts based on monthly point forecasts of r_{t+1} at time t under the the alternative and the historical mean benchmark models, respectively.

2. Cumulative Differences of Sum of Forecast Errors (CDSFE):

$$CDSFE_t := \sum_{h=1}^{t}(r_{h+1} - \bar{r}_{h+1|h})^2 - \sum_{h=1}^{t}(r_{h+1} - \widehat{r}_{h+1|h})^2.$$

- The positive value of $CDSFE_t$ implies the competing model beats the historical mean benchmark in terms of lower MSFEs.
- A positive (negative) slope of $CDSFE_t$ indicates a dynamic trend of improving (worsening) forecasts.

Main result

- Denote the proposed estimator as: CCALM (constrained CALM)
- Baseline model: conventional kitchen-sink (KS) model:
  $r_{t+1} = \mu_o + \boldsymbol{X}_t^{\top}\boldsymbol{\beta} + u_{t+1}, \ \forall \boldsymbol{X}_t \in \mathcal{X}$.
- Other competing models:
  - simple forecast combination (SFC, Rapach, et al., 2010),
  - diffusion index factor model (DI, Bai and Ng, 2008) with pre-screening by the highest in-sample adjusted R^2 of PCs,
  - random forests (RF, Breiman, 2001) with moving block bootstraps for 500 random trees,
  - L_2-boosting (L2B, Bühlmann and Yu, 2003),
  - LASSO (Tibshirani, 1996), and
  - elastic-net (EN, Zou et al., 2005).
- LASSO, EN, and CCALM utilize the time-series cross validation to select tuning parameters.

Table: Out-of-sample Performance

Criteria	KS	SFC	DI	RF	L2B	LASSO	EN	adaLASSO	CCALM
oosR^2(%)	-18.94	-0.13	-1.61	-1.91	0.26	-5.95	-6.12	-6.22	0.67
in$\bar{R}^2$(%)	0.09	0.11	0.10	-0.09	0.13	0.09	0.12	0.11	2.65

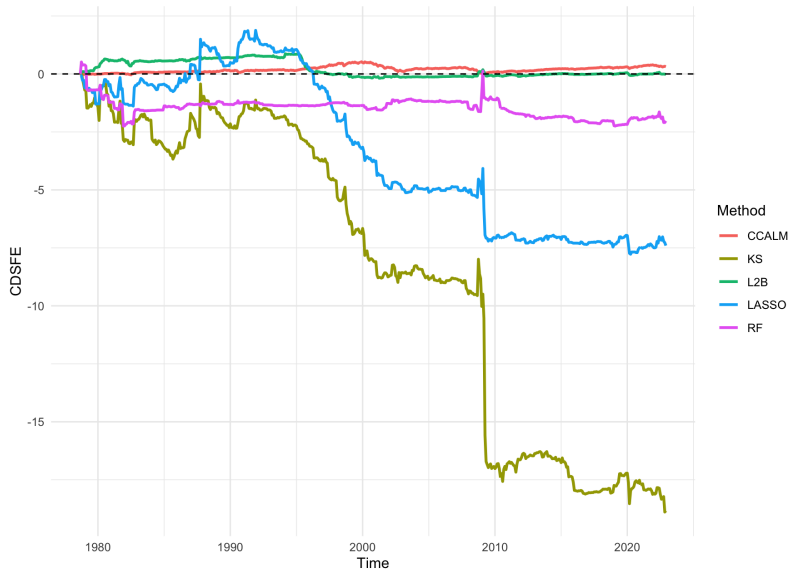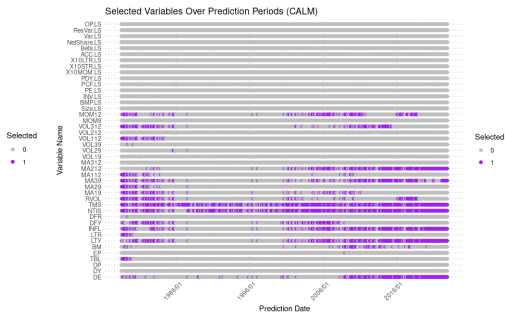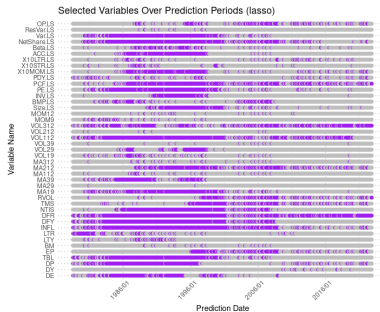Note: oosR^2(%) denotes the out-of-sample R^2, and in$\bar{R}^2$(%) measures the in-sample adjusted R^2.

CDSFEs

Illustration: variable selection for the first 40 predictors

Selected Variables Over Prediction Periods (lasso)

Selected Variables Over Prediction Periods (CALM)

Concluding Remarks

Summary

- **To accurately forecast a weakly stationary $I(0)$ target variable, it is recommended to use a set of weakly stationary $I(0)$ predictors. However, if the use of (nearly) $I(1)$ non-stationary predictors becomes unavoidable, the users must proceed carefully.**

- This project proposes a new variable selector, **CALM**, to select effective predictors to examine the weakly stationary $I(0)$ equity premium forecasts when dealing with many mixed-root predictors.

- From the preliminary empirical findings, the new approach outperforms conventional methods (e.g., Kitchen-sink, LASSO) when the predictor set consists of many possibly stationary and non-stationary predictors for out-of-sample equity premium forecasts.

- The theoretical properties of the proposed estimator are yet to be established in my future work.

THANK YOU FOR YOUR LISTENING