Documentation for Open Source Cross-Sectional Asset Pricing Data

Andrew Chen and Tom Zimmermann
Data Release: March 2022
Created with Code Version: 1.2.0

Major restructuring relative to April 2021

- SignalDocumentation.xlsx is replaced with SignalDoc.csv
  - This allows clean version tracking of the signal documentation using git.
  - SignalDocumentation.xlsx is now OldSignalDocumentation.xlsx (still on github if somebody needs it) and no longer updated
  - Comparisons to other meta studies are in Comparison_to_HLZ.csv and Comparison_to_MetaReplications.csv

New data relative to the April 2021 release

- Data through December 2021
- Two new predictors:
  - TrendFactor from Han, Zhou and Zhu 2016
  - Recomm_ShortInterest from Drake, Rees, and Swanson 2011
- Fama-French 1993 style 2x3 implementations for all signals

Fixes relative to April 2021 release

- Coskewness and CoskewACX: now uses Ken French's market return and risk-free rates
  - Old version used CRSP's NYSE/AMEX or NYSE only index and CRSP's risk free rates.
- Accruals: now more closely match Sloan 1996 by including depreciation.
  - https://github.com/OpenSourceAP/CrossSection/issues/51
- Delisting return adjustments now computed with compounding
  - https://github.com/OpenSourceAP/CrossSection/issues/49
- Quarterly Compustat lagging deals with subtle issues with rdq
  - https://github.com/OpenSourceAP/CrossSection/issues/50
- Various bug fixes
- For a complete list of closed issues see: https://github.com/OpenSourceAP/CrossSection/issues?q=is%3Aissue%20is%3Aclosed%20project%3Aopensourceap%2Fcrosssection%2F3

Signals with Significant Changes

- Accruals: now more closely match Sloan 1996 by including depreciation.
  - https://github.com/OpenSourceAP/CrossSection/issues/51
    - Thanks to a Ph.D. student who would like to remain anonymous.
    - It's worth noting that Xie 2001 defines accruals differently.
  - New long-short returns are quite different than the old ones.
    - post-pub, R-sq is only 24%.
    - however in-sample, R-sq is 73%
  - New accruals predictive power is still quite strong
    - in-sample t-stat increased from 5.07 to 6.70
    - post-pub t-stat decreased from 2.37 to 1.43
  - Overall, accruals is robust to the definition, though of course it is affected by post-publication decay.

- Coskewness and CoskewACX: now uses Ken French's market return and risk-free rates
  - Old version used CRSP's NYSE/AMEX or NYSE only index and CRSP's risk free rates.
    - Ang, Chen, and Xing (2006) Table 8 says they use coskewness "with respect to the market of all stocks listed on the NYSE," which the old code was more faithful too.
    - But the change simplifies the code a lot by avoiding the specialized download which can run into licensing issues.
  - New long short returns are very similar to old long short returns in-sample, but somewhat different post-pub for coskewness
    - R-sq is 85-98 percent, except for Coskewness in the post-pub sample, which produces an R-sq of 70%
  - New Coskewness and CoskewACX predictability is still strong in-sample, but a bit weaker post-pub
    - in-sample t-stats are almost the same as before
    - Post pub t-stats drop from around 1.6 to around 1.2

Summary Stats

[1] "Count of predictors with long-short returns by month"

	date	n_distinct(signalname)
1	2021-12-31	187
2	2020-12-31	197
3	2019-12-31	198
4	2018-12-31	197
5	2017-12-29	199
6	2016-12-30	201
7	2015-12-31	201
8	2014-12-31	201
9	2013-12-31	200
10	2012-12-31	201
11	2011-12-30	201
12	2010-12-31	202
13	2009-12-31	202
14	2008-12-31	202
15	2007-12-31	202
16	2006-12-29	206
17	2005-12-30	206
18	2004-12-31	205
19	2003-12-31	205
20	2002-12-31	207
21	2001-12-31	207
22	2000-12-29	207

[1] "Summary of portfolio full set mean monthly long-short returns"

	impname	before	insamp	between	postpub	last5years	2021	impid
1	PredictorPortsFull	0.56	0.68	0.45	0.31	0.19	0.82	1
2	HoldPer_1	0.53	0.72	0.49	0.32	0.19	0.80	2
3	HoldPer_12	0.27	0.49	0.29	0.20	0.12	0.68	2
R4	HoldPer_3	0.36	0.63	0.42	0.31	0.23	0.81	2
5	HoldPer_6	0.33	0.57	0.35	0.27	0.18	0.77	2
6	ME_gt_NYSE20pct	0.27	0.49	0.27	0.22	0.11	0.66	3
7	NYSEonly	0.39	0.45	0.13	0.22	0.18	0.80	3

```
8  Price_gt_5         0.35   0.56   0.34    0.22     0.10      0.75  3
9  VWforce            0.45   0.44   0.27    0.12    -0.01      0.25  3
10 Quintiles          0.44   0.64   0.41    0.26     0.13      0.49  4
11 QuintilesVW        0.31   0.41   0.28    0.11    -0.03      0.15  4
12 Deciles            0.41   0.80   0.50    0.36     0.25      0.65  5
13 DecilesVW          0.33   0.54   0.32    0.16     0.04      0.26  5
14 FF93style          0.12   0.20   0.11    0.03    -0.06      0.33  6
15 PlaceboPortsFull   0.12   0.25   0.36    0.10     0.08      0.39 NA


[1] "Rsq from regressing new long-short OP returns on old"
   samptype      p05      p10      p25      p50
1 full-samp 96.04197 98.35739 99.65859 99.91149
2   in-samp 98.21118 99.13929 99.63985 99.90619
3  post-pub 94.06720 97.87471 99.66754 99.92710
```

Directory

- SignalDoc.csv
  - Describes each signal and contains hand-collected statistics
  - Information was previously in BasicInfo and AddInfo tabs of SignalDocumentation.xlsx
  - Other tabs of SignalDocumentation.xlsx can be found on Github site
- Firm Level Characteristics/
  - Note: downloadable predictors do not include size, price, or short-term reversal. To make them, use Size.do, Price.do, and STreversal.do in our signals code, or 12_CreateCRSPPredictors.R in our portfolios code. We also do not include bid-ask spreads from TAQ, but that was not shown to predict returns anyway.
  - Full Sets/
    - PredictorsIndiv.zip
      - A zip file containing all of the csvs in Predictors/
    - PlacebosIndiv.zip
      - Similar to PredictorsCsvs.zip, but for not predictive or indirect signals
    - signed_predictors_dl_wide.zip
      - A single csv with all (downloadable) predictors, signed so higher signal implies higher mean returns based on OPs.
  - Individual/
    - Predictors/
      - These files are for convenient retrieval of a particular predictor.
      - Each csv in this folder has columns: permno, yyyymm, [signalname], where [signalname] is the acronym for a signal used in the paper. The [signalname] column has values of the characteristic (a.k.a. signal).
    - Placebos/
      - Same as Predictors/, but for characteristics that were not predictive or indirect signals based on the original results. See SignalDoc.csv for details.
- Portfolios/
  - Full Sets OP/
    - Full sets of portfolios following the original papers (OP)
    - PredictorSummary.xlsx
      - Summary stats for all portfolios

- PredictorLSretWide.csv
  - Columns: date, AbnormalAccruals, Accruals, …, zerotradeAlt12.
  - Description:
    - long-short return during the month indicated by date (if sorted, return between previous date and date), implemented based on original papers. See SignalDoc.csv for details.
    - These are performance of trading strategies based on cross-sectional predictors, or, if you like, realized factor premiums.
- PredictorPortsFull.csv
  - Columns: signalname, port, date, ret, signallag, Nlong, Nshort
  - Description:
    - Sets of portfolios formed on each predictor (e.g. 5 portfolios formed by sorting on momentum). Also includes the long-short portfolios.
    - Includes what some people call "test asset returns." Can also be used to study monotonicity.
    - Implementations based on the original papers. See SignalDoc.csv.
- Placebo*.csv
  - Portfolios and summaries based on not-predictors and indirect signals.

- Full Sets Alt/
  - Full sets of portfolios with alternative implementations. Each zip file consists of just one csv file with the same filename excluding the suffix.
  - PredictorAltPorts_Deciles
    - Like PredictorPortsFull, but only continuous predictors, and implemented as deciles. Stock weights follow OP.
  - PredictorAltPorts_DecilesVW
    - Like PredictorAlt_PortsDeciles, but stocks weights are all VW
  - PredictorAltPorts_FF93style
    - Like PredictorPortsFull, but uses Fama-French 1993 style 2x3 sorts
  - PredictorAltPorts_HoldPer_*.zip
    - Like PredictorPortsFull.csv, but uses "rebalancing periods" of 1, 3, 6, and 12-months instead of the original rebalancing periods.
    - These rebalancing periods should really be called signal updating periods, since value-weighting or equal-weighting is always enforced monthly, see footnote in the paper.
  - PredictorAltPorts_Quintiles and PredictorAltPorts_QuintilesVW
    - Same as PredictorAltPorts_Deciles and PredictorAltPortsDecilesVW but with quintile sorts
  - PredictorAltPorts_LiqScreen_*.csv
    - Like PredictorPortsFull.csv, but with various liquidity adjustments. Please see paper
  - Placebo*.zip

            - Portfolios based on not-predictors and indirect signals.
    - Individual/
        - For convenience, these are csvs for portfolio sorts on a specific characteristic. For example, you can pull the simple B/M sorted portfolios directly from a BM.csv file in here.
            - All csvs are in wide format with columns (date, port1, port2, …, port[N], portLS)
        - Original Cuts/
            - Each csv here has returns from assigning stocks to portfolios based on a given predictor, and implementing following the original papers
        - Original CutsVW/
            - Like Original Cuts/ but value-weighted
        - Cts Deciles/
            - Like Original Cuts/, but using only continuous predictors and sorted into deciles
        - Cts DecilesVW/
            - Like Cts Deciles, but value-weighted
        - Cts Quintiles/
            - Like Cts Deciles/, but using quintiles
        - Cts QuintilesVW/
            - Like Cts Quintiles/, but value weighted
            -
        -
- DailyPortfolios/
    - Daily portfolio returns. Aggregates up to the monthly strategies (almost). Only contains returns (%, daily), please see monthly data for supporting statistics like # of stocks.
    - DailyPortSummary.xlsx
        - Some summary stats for the daily portfolios, since we don't provide results in the paper.
        - Sumstats sheet should be self explanatory. Note number of signals in each predictor port varies because the original paper vary in the number of portfolios they form.
        - Timingcheck: regressions of monthly returns on daily returns aggregated to monthly, done by groups of (signalname, portfolio).
        - Further details, see daily portfolio construction R script.
    - Predictor.zip
        - Following original papers.
        - Columns: date, port1-portN, portLS
        - Other than dates, values are returns, as indicated by the filename.
    - Other implementations (below) match previous descriptions (see Portfolios / Full sets alt)
        - PredictorVW.zip
        - CtsPredictorDecile.zip
        - CtsPredictorDecileVW.zip
        - CtsPredictorQuintile.zip
        - CtsPredictorQuintileVW.zip