Documentation for Open Source Cross-Sectional Asset Pricing Data

Andrew Chen and Tom Zimmermann
Data Release: October 2024
Created with Code Version: 1.4.1

This release is a small (but important) patch of the August 2024 release.

- The patch fixes a lookahead bias bug in AnnouncementReturn. All other predictors' data is the same as in the August 2024 release.
- For details see https://github.com/OpenSourceAP/CrossSection/issues/158.
- As such, these notes compare to the August 2023 release

New data relative to the August 2023 release

- Data through December 2023
- *Exception*: Predictors based on OptionMetrics are taken from our August 2023 release and end in December 2022
    - OptionMetrics revised their methodologies in March 2024. This revision includes new proprietary methods and the removal of the 300% implied vol cap.
    - These new methods seem to have caused large revisions to option-based predictors. See details in https://github.com/OpenSourceAP/CrossSection/issues/156
    - It's unclear what is the best way to handle these revisions. So for now using pre-2024 OptionMetrics data is the most transparent.
    - A version of our dataset with the new OptionMetrics data can be found in the folder `Dirty Version 2024.08/`

Major Updates relative to August 2023 release

- Two signals needed to be switched to new data sources after the original ones were discontinued:
    - betaVIX is originally based on VXO (volatility index based on S&P100) which was discontinued in September 2021. We switch to VIX afterwards.
    - Mom6mJunk is based on S&P ratings data but WRDS S&P credit ratings end in Feb 2017. We switch to Capital IQ S&P ratings data from 2016. In addition, we only assign a stock to "Junk" if it has a proper credit rating and the credit rating is low (previously, we interpreted missing credit ratings as "Junk" as well).
- Fixed typos in the signal documentation (signaldoc.csv) for some signals (DivYieldST, dCPVolSpread, AgeIPO).
- Fixed low number of observations in some months or years in two signals (FirmAgeMom, ForecastDispersion) that were due to filters that set some observations to missing.
- New code for the "zerotrade" signals that more closely follows Liu (2006). Also rationalized naming.
    - zerotrade1M, zerotrade6M and zerotrade12M are the 1-,6- and 12-month versions of the signal (as opposed to zerotradeAlt1, zerotrade, zerotradeAlt12 in earlier versions).

- FailureProbability requires book value of equity and its construction now follows Cohen, Polk, and Vuolteenaho (2003) (instead of just using "ceqq") as referenced by Campbell, Hilscher and Szilagyi (2008).
- We verified that there is no look-ahead bias in signals that use "cfacshr" or "cfacpr".
    - In the process, we lagged "ShareIss5Y" by an additional 5 months and we included alternative code (as a comment) for "ShareIss1Y" that closely follows Pontiff and Woodgate (2008) that gives very similar results to our implemented version.
    - We still need to check signals that are based on 13F data.
- For a complete list of closed issues see: https://github.com/OpenSourceAP/CrossSection/issues?q=is%3Aissue+closed%3A2023-08-16..2024-08-22+sort%3Aupdated-desc

Signals with Significant Changes

- zerotrade1M, zerotrade6M, zerotrade12M: construction now more closely matches Liu (2006)
    - https://github.com/OpenSourceAP/CrossSection/issues/132
- Mom6mJunk: Switches to Capital IQ S&P ratings data in 2006 and no longer interprets missing ratings as junk
    - https://github.com/OpenSourceAP/CrossSection/issues/135
- ForecastDispersion: Higher coverage in December, January and February
    - https://github.com/OpenSourceAP/CrossSection/issues/145
- FailureProbability, FailureProbabilityJune: Book equity construction now more closely follows description in Campbell, Hilscher and Szilagyi (2008)
    - https://github.com/OpenSourceAP/CrossSection/issues/141

Summary Stats

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

	date	n_distinct(signalname)
1	2023-12-29	194
2	2022-12-30	206
3	2021-12-31	206
4	2020-12-31	206
5	2019-12-31	206
6	2018-12-31	205
7	2017-12-29	206
8	2016-12-30	206
9	2015-12-31	206
10	2014-12-31	206
11	2013-12-31	207
12	2012-12-31	207
13	2011-12-30	207
14	2010-12-31	208
15	2009-12-31	208
16	2008-12-31	208
17	2007-12-31	208
18	2006-12-29	212

```
19 2005-12-30                    212
20 2004-12-31                    212
21 2003-12-31                    211
22 2002-12-31                    212
23 2001-12-31                    212
24 2000-12-29                    212
```

The drop at the end of 2023 is mainly due to our decision not to update OptionMetrics predictors due to inconsistencies between data vintages.

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

	impname	before	insamp	between	postpub	last5years	2023	impid
1	PredictorPortsFull	0.55	0.69	0.43	0.32	0.36	0.08	1
2	HoldPer_1	0.52	0.73	0.47	0.34	0.36	0.07	2
3	HoldPer_12	0.24	0.48	0.26	0.22	0.28	0.02	2
4	HoldPer_3	0.36	0.63	0.41	0.32	0.38	0.03	2
5	HoldPer_6	0.31	0.57	0.33	0.27	0.32	0.02	2
6	ME_gt_NYSE20pct	0.30	0.51	0.26	0.23	0.21	-0.08	3
7	NYSEonly	0.49	0.47	0.22	0.27	0.33	0.11	3
8	Price_gt_5	0.36	0.57	0.32	0.21	0.23	-0.21	3
9	VWforce	0.47	0.45	0.25	0.15	0.09	-0.20	3
10	Quintiles	0.43	0.64	0.39	0.28	0.30	-0.03	4
11	QuintilesVW	0.35	0.41	0.26	0.12	0.05	-0.38	4
12	Deciles	0.38	0.80	0.49	0.37	0.41	0.21	5
13	DecilesVW	0.34	0.53	0.32	0.19	0.19	-0.06	5
14	FF93style	0.11	0.18	0.10	0.03	-0.03	-0.23	6
15	PlaceboPortsFull	0.11	0.26	0.36	0.13	0.16	0.30	NA

[1] "Rsq from regressing new long-short OP returns on old"

	samptype	p05	p10	p25	p50
1	full-samp	95.98326	99.34860	99.87172	99.97999
2	in-samp	98.01056	99.74157	99.96915	99.99957
3	post-pub	94.95243	98.83016	99.75853	99.96377

R2s are somewhat low for the signals with changes to the code in this year's update, with the additional five month lag in ShareIss5Y making the biggest difference. AnnouncementReturn has a relatively "low" R2, given the patch (see above). Still, R2s are extremely high for the vast majority of signals. Below are the lowest full sample R2 signals:

	signalname	samptype	rsq	new_rbar	old_rbar
	<char>	<char>	<num>	<num>	<num>
1:	ShareIss5Y	full-samp	4.80923	0.5892337	0.4557178
2:	AgeIPO	full-samp	67.00795	0.6276788	0.6102104
3:	Mom6mJunk	full-samp	77.07229	1.4740566	1.0514921
4:	ShareIss1Y	full-samp	80.94094	0.7570156	0.6833197
5:	FirmAgeMom	full-samp	83.00328	1.5209500	1.7825054
6:	iomom_supp	full-samp	91.99627	0.4055613	0.4734619
7:	IndIPO	full-samp	94.11073	0.4496679	0.4466776
8:	CredRatDG	full-samp	94.32627	0.5590158	0.6781290
9:	ChNAnalyst	full-samp	95.35690	0.3285226	0.1614635
10:	TrendFactor	full-samp	95.90370	1.5890551	1.6152437
11:	ForecastDispersion	full-samp	95.95527	0.5812503	0.4771931

```
12:           iomom_cust full-samp 96.02525 0.3131280 0.2774074
13:     CustomerMomentum full-samp 96.84792 0.8153887 0.8659754
14:                RDIPO full-samp 97.39252 0.6831812 0.6374495
15:   AnnouncementReturn full-samp 97.63745 0.9938312 1.1231940
16:         ReturnSkew3F full-samp 97.73729 0.2365294 0.2533373
17:     AbnormalAccruals full-samp 98.07943 0.1881846 0.1779257
18: Recomm_ShortInterest full-samp 98.46960 0.8097401 0.8268106
19:               OrgCap full-samp 99.05608 0.2783493 0.2943467
20:      AnalystRevision full-samp 99.11953 0.6352937 0.6281084
```

Directory

- SignalDoc.csv
  - Describes each signal and contains hand-collected statistics
- 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