

Documentation for Open Source Cross-Sectional Asset Pricing Data

Andrew Chen and Tom Zimmermann

Data Release: October 2025

Created with Code Version: 2.0.0

New data relative to the October 2024 release

- Data through December 2024
- *Exceptions:*
 - Predictors based on OptionMetrics implied vol (CPVolSpread, dCPVolSpread, dVolCall, dVolPut, RIVolSpread, skew1, SmileSlope) are taken from [our August 2023 release](#) and end in December 2022
 - OptionMetrics [revised their methodologies](#) in March 2024. This revision includes new proprietary borrow fee calculations.
 - 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.
 - Predictors based on the Gompers, Ishii, Metrick governance index (Activism1, Activism2, and Governance) end in 2007 as that data is no longer updated
 - ProbInformedTrading relies on the Duarte et al. data which ends in 2012. See <https://github.com/OpenSourceAP/CrossSection/issues/171>.
- Not a data update but a data download update: There are now packages to download the data directly in Python and R:
 - Python: <https://github.com/mk0417/open-asset-pricing-download>
 - R: <https://github.com/tomz23/OpenSourceAP.DownloadR>
- Also, there is now an html browser that lets you easily access summary stats for all reproduced signals:
<https://openassetpricing.com/SignalDoc-Browser.html>

Signals with Non-Trivial Revisions to Long-Short Monthly Returns

- Non-trivial revisions means $R^2 < 95\%$ from regressing this vintage's monthly returns on the previous vintage (see table below)
- Traceable to issue fixes
 - ChNAnalyst: fix bug in size filter
 - See <https://github.com/OpenSourceAP/CrossSection/issues/182>
 - PriceDelayTstat: fix typo in formula
 - See <https://github.com/OpenSourceAP/CrossSection/issues/177>
 - Recomm_ShortInterest: fix asrol fill forward bug
 - See <https://github.com/OpenSourceAP/CrossSection/issues/178>
 - OptionVolumel: improve replication by matching expiration filter
 - See <https://github.com/OpenSourceAP/CrossSection/issues/166>
 - MS: Python translation + using market equity at the permco level
 - See <https://github.com/OpenSourceAP/CrossSection/issues/167>
- Traceable to Python translation
 - MomOffSeason*YrPlus: Python translation significantly simplifies the Stata code, no longer uses asrol.

- RDAbility, CustomerMomentum, EarnSupBig, AbnormalAccruals: Python translation by necessity does not perfectly replicate Stata due to the complicated calculation (rolling regression with many missing values).
 - See the Stata comparison test [override explanations](#)
- Source of revision unclear
 - CompositeDebtIssuance, ExchSwitch, InvGrowth, Investment, ExclExp, RDS, GrAdEXp: For these signals, the Python translation almost perfectly replicates the Stata code.
 - See the Stata comparison test [override explanation](#)
 - CRSP revisions could be the source
 - Nevertheless, the mean returns are quite close to the previous vintage

Summary Stats

```
[1] "Count of predictors with long-short returns by month"
      date n_distinct(signalname)
1  2024-12-31           195
2  2023-12-29           198
3  2022-12-30           208
4  2021-12-31           207
5  2020-12-31           207
6  2019-12-31           207
7  2018-12-31           207
8  2017-12-29           208
9  2016-12-30           208
10 2015-12-31           208
11 2014-12-31           207
12 2013-12-31           208
13 2012-12-31           209
14 2011-12-30           209
15 2010-12-31           208
16 2009-12-31           208
17 2008-12-31           208
18 2007-12-31           209
19 2006-12-29           212
20 2005-12-30           212
21 2004-12-31           212
22 2003-12-31           212
23 2002-12-31           212
24 2001-12-31           212
25 2000-12-29           212
```

The drop after the end of 2022 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 2024  impid
1 PredictorPortsFull 0.57    0.69    0.41    0.30    0.41    0.23    1
2 HoldPer_1         0.56    0.73    0.45    0.31    0.41    0.24    2
```

3	HoldPer_12	0.25	0.47	0.24	0.18	0.25	0.21	2
4	HoldPer_3	0.38	0.62	0.39	0.29	0.42	0.32	2
5	HoldPer_6	0.28	0.57	0.32	0.24	0.36	0.37	2
6	ME_gt_NYSE20pct	0.30	0.51	0.26	0.20	0.24	0.07	3
7	NYSEonly	0.50	0.47	0.25	0.24	0.34	0.13	3
8	Price_gt_5	0.38	0.56	0.32	0.20	0.29	0.17	3
9	VWforce	0.46	0.45	0.23	0.12	0.11	0.05	3
10	Quintiles	0.44	0.64	0.39	0.28	0.36	0.15	4
11	QuintilesEW	0.43	0.63	0.38	0.29	0.38	0.18	4
12	QuintilesVW	0.35	0.41	0.27	0.12	0.07	0.02	4
13	Deciles	0.40	0.80	0.49	0.37	0.48	0.30	5
14	DecilesEW	0.38	0.76	0.47	0.36	0.47	0.27	5
15	DecilesVW	0.35	0.53	0.32	0.19	0.25	0.18	5
16	FF93style	0.13	0.18	0.10	0.02	-0.02	-0.05	6
17	PlaceboPortsFull	0.10	0.24	0.34	0.13	0.21	0.19	NA

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

	samptype	p05	p10	p25	p50
1	full-samp	88.46425	95.31884	98.89308	99.83597
2	in-samp	90.66791	95.53936	98.97383	99.83753
3	post-pub	87.55878	94.34306	99.21712	99.85083

	signalname	samptype	rsqr	new_rbar	old_rbar
	<char>	<char>	<num>	<num>	<num>
1:	ChNAnalyst	full-samp	7.099251e-04	0.4585037	0.2794887
2:	PriceDelayTstat	full-samp	1.246887e+01	0.1361000	0.1020680
3:	Recomm_ShortInterest	full-samp	3.925202e+01	0.6109566	0.7890873
4:	MomOffSeason16YrPlus	full-samp	6.490678e+01	0.3210458	0.3192998
5:	CompositeDebtIssuance	full-samp	8.081694e+01	0.2972230	0.2720846
6:	ExchSwitch	full-samp	8.148425e+01	0.6292463	0.6375997
7:	OptionVolume2	full-samp	8.372378e+01	0.2639945	0.2571851
8:	MS	full-samp	8.397739e+01	0.7895390	0.7973460
9:	MomOffSeason11YrPlus	full-samp	8.495197e+01	0.1389744	0.2058205
10:	MomOffSeason06YrPlus	full-samp	8.625975e+01	0.6121516	0.6607700
11:	RDAbility	full-samp	8.809204e+01	0.2959931	0.2029253
12:	OptionVolume1	full-samp	8.876878e+01	0.5244311	0.5136342
13:	CustomerMomentum	full-samp	9.157069e+01	0.7646749	0.7559220
14:	InvGrowth	full-samp	9.276639e+01	0.7553523	0.6876475
15:	Investment	full-samp	9.281368e+01	0.1414084	0.1733874
16:	AbnormalAccruals	full-samp	9.327675e+01	0.1723683	0.1845056
17:	ExclExp	full-samp	9.361456e+01	0.1724273	0.1573885
18:	EarnSupBig	full-samp	9.373152e+01	0.3922751	0.4366822
19:	RDS	full-samp	9.431404e+01	0.2664878	0.2689263
20:	GrAdExp	full-samp	9.496743e+01	0.3265617	0.3564051

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.py, Price.py, and STreversal.py 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, yyyyymm, [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.
- o 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.
- o 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