

PCRA Package and Data Overview

Doug Martin, Tom Philips, Jon Spinney and Kirk Li

July 7, 2026

1 The PCRA Package

This vignette is intended primarily as an overview of the structure and contents of the PCRA package; other vignettes and help material cover its functionality and use, and those of related packages, in greater depth. We assume that you have already installed R (version 4.4.3 or later) and a current version of RStudio, or will do so before continuing. To install R and RStudio, we recommend reading the relevant sections of Appendix A in *Hands-On Programming with R*, available at <https://rstudio-education.github.io/hopr/>.

The PCRA package was created to provide data-oriented computational support for the book *Robust Portfolio Construction and Risk Analysis* (RPCRA), co-authored by Doug Martin, Thomas Philips, Bernd Scherer, and Kirk Li. The book's publication by Springer is planned for 2026.

An important feature of the RPCRA book is its emphasis on reproducibility: most of its figures and tables can be reproduced with R code available in the demo folder of the PCRA package. Such reproducibility is made possible by the inclusion of many (but not all) of the data sets used by the R code that reproduces the figures and tables. The initial release of PCRA package contains thirty-one data sets. The most frequently used are CRSP® stocks data,¹ Standard and Poor's Global Market Intelligence (SPGMI) factors data, S&P Dow Jones indices including the S&P 500®, S&P Industrials®, S&P 400 Industrials®, and the S&P 425 Industrials®, as well as data sets from the Chicago Board Options Exchange (CBOE) LiveVol, Conference Board, Intercontinental Exchange (ICE) and Tradeweb Markets LLC.

These data sets may be used together with factor model fitting, portfolio optimization, and portfolio performance analysis methods in PCRA and other R packages to support realistic quantitative equity portfolio construction and performance analyses. These other packages include, but are not limited to, the time-series factor models package `facmodTS`, the cross-section factor models package `facmodCS`, the portfolio performance analysis package `PerformanceAnalytics`, and the portfolio optimization package `PortfolioAnalytics`, all of which are available at CRAN (<https://cran.r-project.org/>).

It is important to keep in mind that there currently exist two versions of the PCRA package:

¹CRSP® stands for the Center for Research in Security Prices, LLC, an affiliate of the University of Chicago Booth School of Business.

- PCRA-1.3.0 (the production version), available on CRAN at <https://cran.r-project.org/package=PCRA>
- PCRA-1.3.1 (the development version), available at <https://github.com/robustport/PCRA/>.

Before continuing, install the CRAN version of PCRA via RStudio's Tools > Install Packages menu: enter PCRA in the Packages field and click Install. After the installation is finished, load the package using R's `library` function, ignoring any warning messages that might appear.

```
library(PCRA)
```

The two PCRA versions have identical functionality except for (i) new features and data sets in the development version, and (ii) differences in their CRSP[®] stocks and SPGMI factors data. Specifically, due to data storage space constraints, the PCRA-1.3.0 package hosted on CRAN contains CRSP[®] stocks data only at a monthly frequency, and contains only a slightly rounded version of the SPGMI factors data, with the following names, respectively:

- `stocksCRSPmonthly`
- `factorsSPGMIr`.

For details on these two data sets, use the R `help` function:

```
help(stocksCRSPmonthly)
help(factorsSPGMIr)
```

In addition to the monthly CRSP[®] stocks data and rounded SPGMI factors data contained in PCRA-1.3.0, the PCRA-1.3.1 development version also contains daily and weekly CRSP[®] stocks data and unrounded SPGMI factors data. The additional PCRA-1.3.1 data sets are:

- `stocksCRSPweekly`
- `stocksCRSPdaily`
- `factorsSPGMI`.

Section 2.2 provides details concerning the rounding used to create the `factorsSPGMIr` data from the `factorsSPGMI` data, and shows that there are only minor differences on the order of 1% (and typically much less than 1%) in `factorsSPGMIr` data values relative to their `factorsSPGMI` counterparts. This suggests that one will see minimal differences in portfolio performance when using `factorsSPGMIr` in place of `factorsSPGMI`. Nonetheless, users of the PCRA-1.3.0 package who prefer to use the unrounded

factorsSPGMI data can do so in a manner described in the document “CRSP Stocks and SPGMI Factors in PCRA”, which is downloadable from CRAN at <https://cran.r-project.org/package=PCRA>, or can switch to the development version.

The PCRA package has an extensive collection of R functions that were used to create the book and to allow reproducibility. These can be viewed, along with the names of the included datasets, by typing:

```
ls("package:PCRA")
```

The R code chunks used to create this vignette are provided in the PCRA Overview Demo.R script in the demo folder of the PCRA package. Instructions for accessing and running such scripts are provided in the vignette “PCRA Reproducibility Code”, which we recommend reading in conjunction with this vignette.

1.1 Other Data in the PCRA Package

In addition to the relatively large CRSP[®] and SPGMI data sets, the PCRA package contains many other data sets used throughout the RPCRA book. These may also be used for a wide variety of computing exercises, but strictly for academic (i.e., non-commercial) purposes. The following call to R’s data function produces the display shown in Table 1, listing each data set name with its brief description:

```
data(package = "PCRA")
```

Detailed descriptions of each of the PCRA data sets can be obtained using R’s help function, e.g., to see the details of the SP500 index data, type:

```
help(SP500)
```

Data Set Name	Manual Page Title
CRSPLiquidMktCapGrpsCnts	CRSP Liquid Stocks Cap Groups Counts
CboeOptionStrategies	CboeOptionStrategies
ConferenceBoardETI	ConferenceBoardETI
FRBInterestRates	Federal Reserve Board Interest Rates
HFstrategies	Hedge Fund Strategies Returns
SP400Industrials	SP400Industrials
SP425Industrials	SP425Industrials
SP500	SP500
SP500from1967to2007	SP500from1967to2007
SPIndustrials	SPIndustrials
ShortDurationCredit	ShortDurationCredit
USTreasuryTradeweb	USTreasuryTradeweb
crsp.returns8	CRSP Returns for 8 stocks in 4 cap groups
datFF3W	Fama-French 3-Factor Model Weekly Time Series
datFF4W	Fama-French 4-Factor Model Weekly Time Series
factorsSPGMI	SPGMI Data 14 Factors for 294 Stocks
factorsSPGMIr	Rounded Version of SPGMI Data 14 Factors
gfunds5	Five German Investment Funds
invensysEPS	Invensys Earnings per Share
retDD	CRSP Returns of Stock with Ticker DD
retEDS	CRSP Returns of Stock with Ticker EDS
retFNB	CRSP Returns of Stock with Ticker FNB
retKBH	CRSP Returns of Stock with Ticker KBH
retMER	CRSP Returns of Stock with Ticker MER
retOFG	CRSP Returns of Stock with Ticker OFG
retPSC	CRSP Returns of Stock with Ticker PSC
retVHI	CRSP Returns of Stock with Ticker VHI
retWTS	CRSP Returns of Stock with Ticker WTS
stocksCRSPdaily	CRSP daily stocks data for 294 stocks
stocksCRSPmonthly	CRSP monthly stocks data for 294 stocks
stocksCRSPweekly	CRSP weekly stocks data for 294 stocks

Table 1: PCRA package data sets and their corresponding descriptions

1.1.1 The S&P Data Sets

Standard and Poor's has graciously consented to the inclusion of five data sets in the PCRA package, one of which (SP500) is updated annually as new year-end data on the S&P 500® Index becomes available. The

other four data sets are historical and are copied from paper copies of the S&P Analysts' Handbook. All five data sets are of type `data.frame`.

S&P provides investors with a great deal of information on the S&P 500® at <https://www.spglobal.com/spdji/en/indices/equity/sp-500/>, and some (but not all) of this information can be found in `SP500`, which has (among other items), the per-share earnings, revenues, and price level of the S&P 500® Index from 1925 to the most recent year-end. It is updated once a year, typically in April, after S&P's analysts finalize all the year-end data for the index. In addition, it provides year-end CPI and Nominal GDP levels sourced from the Federal Reserve's FRED database at <https://fred.stlouisfed.org/>.

It is instructive to compare `SP500` with `SP500from1967to2007`, `SPIndustrials`, `SP400Industrials` and `SP425Industrials`. The first is updated annually; a superset of the data added each year can be viewed at <https://www.spglobal.com/spdji/en/documents/additional-material/sp-500-eps-est.xlsx/>. The latter four data sets are, in contrast, derived from scans of old paper copies of the S&P Analysts' Handbook. Some columns that appeared in past editions (e.g., `Cash_Flow`, `Depreciation`, and `TotalReturnIndex`) are no longer included, while new columns (e.g., `SP500CapEx`) have been added.

1.1.2 The CBOE Dataset

The `CboeOptionStrategies` data set is a `data.frame` containing total return indices for ten options-related strategies, price and total return indices for the S&P 500® Index, the yield on 3-month T-bills, and two equity volatility indices (VIX and VXO), covering June 1986 to December 2021. Information on these ten strategies, as well as many others, can be found at <https://www.cboe.com/us/indices/indicessearch/>.

1.1.3 The Conference Board Dataset

The PCRA data set `ConferenceBoardETI` is a `data.frame` with the values of the Conference Board's Employment Trends Index™ from November 1973 to September 2023. The Employment Trends Index is a proprietary composite of eight employment indicators, and information on it can be found at <https://www.conference-board.org/topics/employment-trends-index/>. The Conference Board publishes a number of indicators of the health of the U.S. and global economies, including the Leading Economic Index® (<https://www.conference-board.org/topics/us-leading-indicators/>) and the Total Economy Database™ (<https://www.conference-board.org/topics/total-economy-database/>), but these are available only to subscribers to their service.

1.1.4 The ICE Dataset

The PCRA data set `ShortDurationCredit` is a `data.frame` with the monthly total return, Yield to Maturity, Weighted Average Maturity, Modified Duration and Option-Adjusted Spread of the ICE BofA 1-3 Year AAA-A US Corporate Index (Index C110) and the ICE BofA 1-3 Year US Treasury Index (Index G102).

1.1.5 The Tradeweb Dataset

The PCRA data set `USTreasuryTradeweb` is a `data.frame` with the daily bid and ask prices and yields of the On-the-run and First Off-the-run 1-Year, 2-Year, 5-Year, 10-Year and 30-Year U.S. Treasury Notes (or Bonds) from January 2, 2008 to December 30, 2021.

2 Basics of the CRAN PCRA Package

Here we focus on a few basic aspects of the `stocksCRSPmonthly` and the `factorsSPGMIr` data sets in the CRAN version of PCRA, and defer more detailed discussion to the vignette “CRSP Stocks and SPGMI Factors in PCRA”. The PCRA package contains many R functions and data sets, a list of which can be viewed by using the `ls` function as follows in the RStudio Console:

```
library(PCRA)
ls("package:PCRA")
```

The data sets `stocksCRSPmonthly` and `factorsSPGMIr` are `data.table` objects, and to view and manipulate them you need to install and load the `data.table` R package. Install the current version of `data.table` in RStudio using the same procedure as for the PCRA package. Then load `data.table` by typing `library(data.table)`, and confirm that `stocksCRSPmonthly` and `factorsSPGMIr` are indeed `data.table` objects by typing:

```
library(data.table)
class(stocksCRSPmonthly)

## [1] "data.table" "data.frame"

class(factorsSPGMIr)

## [1] "data.table" "data.frame"
```

PCRA uses `data.table` objects for the `stocksCRSPmonthly` and `factorsSPGMIr` data sets for two reasons:

1. The `data.table` package has powerful data manipulation tools that do not exist for `data.frame` objects, and
2. The `data.table` package can handle large data sets without its performance becoming degraded.

Note that a `data.table` object is indeed a table object, with dimensions and column names obtainable as follows:

```
dim(stocksCRSPmonthly) # Number of rows and columns

## [1] 81144      14

names(stocksCRSPmonthly) # Names of data items in each column

## [1] "Date"          "Ticker"        "TickerLast"    "Company"       "CapGroup"
## [6] "CapGroupLast" "GICS"          "Sector"        "Return"        "RetExDiv"
## [11] "Price"         "PrcSplitAdj"   "Ret13WkBill"   "MktIndexCRSP"
```

You can discover the structure of `stocksCRSPmonthly` by displaying the first and last five rows of its first, second and ninth columns by typing:

```
stocksCRSPmonthly[, c(1:2, 9)]

##           Date Ticker      Return
##           <Date> <char>      <num>
##    1: 1993-01-31  ARONA  0.28947368
##    2: 1993-01-31   ABM -0.11250000
##    3: 1993-01-31   ABT -0.08971193
##    4: 1993-01-31  ADBE  0.31999999
##    5: 1993-01-31   ADI  0.01538462
##    ---
## 81140: 2015-12-31   WGO -0.11555557
## 81141: 2015-12-31   WHR -0.09629589
## 81142: 2015-12-31   WMT  0.05013595
## 81143: 2015-12-31   WTS -0.10520627
## 81144: 2015-12-31   XOM -0.04543236
```

The results displayed above indicate that the row structure of `stocksCRSPmonthly` consists of blocks of size 294, one block for each of the months from 1993 through 2015. That time period consists of 23 years (or 276 months), and since the block for each month consists of 294 stocks, the total number of rows of `stocksCRSPmonthly` is $276 \times 294 = 81,144$, as was seen by the use of `dim(stocksCRSPmonthly)`.

Similarly, you can see the dimensions and column names of `factorsSPGMIr` with:

```

dim(factorsSPGMir) # Number of rows and columns

## [1] 81144      22

names(factorsSPGMir) # Names of data items in each column

## [1] "Date"          "Ticker"          "TickerLast"      "Company"
## [5] "CapGroup"      "CapGroupLast"    "GICS"            "Sector"
## [9] "AnnVol12M"     "Beta60M"         "BP"              "EP"
## [13] "LogMktCap"     "PM12M1M"         "AccrualRatioCF"  "AstAdjChg1YOCF"
## [17] "CFROIC"        "Chg1YAstTo"      "EBITDAEV"        "FCFP"
## [21] "PM1M"          "SEV"

```

2.1 Data Table Objects Versus Data Frame Objects

You may have wondered why `class(stocksCRSPmonthly)` and `class(factorsSPGMir)` return both “data.table” and “data.frame”, the first of which is indeed the class of each of these two data sets. The appearance of “data.frame” after “data.table” indicates that the `data.table` object *inherits* all the properties of a `data.frame` object and adds further features to enhance both its syntax and its performance. Given this inheritance, a `data.table` object can easily be converted to a `data.frame` by applying the `data.frame` function to it:

```

dat.df <- data.frame(stocksCRSPmonthly)
class(dat.df)

## [1] "data.frame"

names(dat.df)

## [1] "Date"          "Ticker"          "TickerLast"      "Company"          "CapGroup"
## [6] "CapGroupLast" "GICS"            "Sector"          "Return"           "RetExDiv"
## [11] "Price"         "PrcSplitAdj"     "Ret13WkBill"     "MktIndexCRSP"

```

Users who are familiar with manipulating `data.frame` objects will find the above conversion method handy.

2.2 The Rounded SPGMI Factors Data

The rounded data in `factorsSPGMIr` was obtained by rounding the data in `factorsSPGMI` to 4 significant digits, and then compressing it with level 9 “xz” compression. The results of these steps are displayed in Table 2.2. The compression reduces the 6.452 MB `factorsSPGMI` data set by 72.3% to about 1.787 MB for `factorsSPGMIr`. This overall compression process was required so that the PCRA package — which also contains the 1 MB `stocksCRSPmonthly` data and other miscellaneous data sets totaling about 2 MB — satisfies CRAN’s maximum package data size of 5 MB.

Table 2: Reduction in size of the `factorsSPGMI` data relative to the original set from rounding and data compression.

factorsSPGMI	Size(KB)	Reduction(%)
Original	6452	0.0
Rounded to 4 Sig. Digits	2702	58.1
Level 9 ‘xz’ Compressed	1787	72.3

Rounding factors data to four significant digits turns out to make little difference in the precision of `factorsSPGMIr` data relative to the `factorsSPGMI`. This claim is supported by Figure 1, which shows that the fractional differences in all the factors is always less than about 1% (it is less than 0.5% for four of the fourteen factors). We expect these minor differences in `factorsSPGMIr` data relative to `factorsSPGMI` data to have a negligible impact on factor model fits.

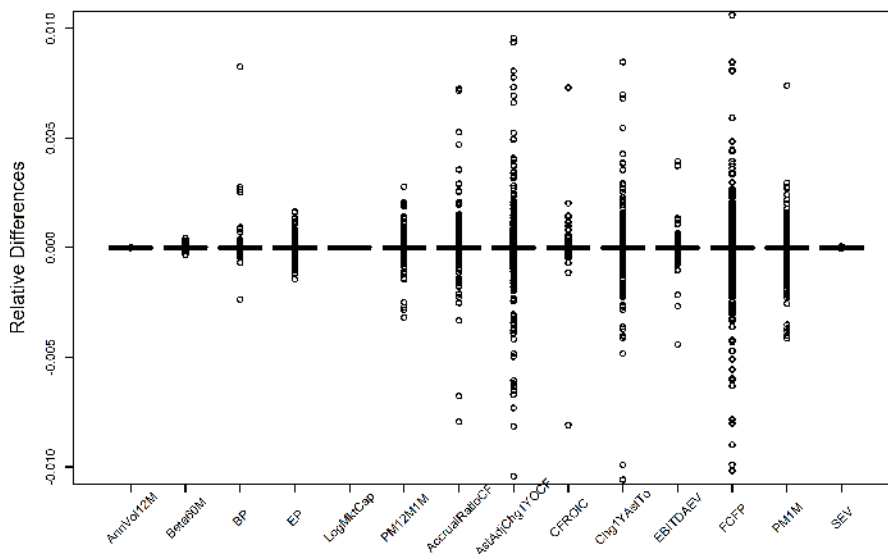


Figure 1: Fractional differences in `factorsSPGMIr` relative to `factorsSPGMI` for each of the 14 factors.

As is explained in the vignette “Introduction to CRSP Stocks and SPGMI Factors in PCRA”, users of the CRAN PCRA-1.3.0 package actually have access to the non-rounded `factorsSPGMI` data in a special `DataPlus` folder of the PCRA Github repository at <https://github.com/robustport/PCRA/>. Our recommendation to users of the CRAN version of PCRA is to first fit factor models using `factorsSPGMIr`, and then, if there is any concern about the adverse impact of rounding, to check the model fit by refitting it using `factorsSPGMI`.

2.3 PCRA Data Sets Are Not Open Source

While the PCRA package contains a number of open source R functions and scripts, the `stocksCRSPmonthly`, `stocksCRSPweekly`, `stocksCRSPdaily`, `factorsSPGMIr`, `factorsSPGMI`, `SP500`, `SP500from1967to2007`, `SPIndustrials`, `SP400Industrials`, `SP425Industrials`, `ChoeOptionStrategies`, `ConferenceBoardETI`, `ShortDurationCredit` and `USTreasuryTradeweb` data sets supplied with it *are not open source*, and *may not be redistributed in any way or form, or used for any commercial (i.e., non-educational) purpose*.

The authors wish to express their gratitude to CRSP®, SPGMI, S&P Dow Jones Indices, CBOE LiveVol LLC, ICE Data Indices LLC and Tradeweb Markets LLC for permitting inclusion of these data sets in the PCRA package. Their generosity will greatly enhance students’ understanding of portfolio construction and risk analysis methods by enabling reproduction of the book’s examples, completion of its computational exercises, and engagement with problems assigned by course instructors.

3 Development Version of PCRA

The development version of PCRA (currently PCRA-1.3.1) can be installed from its Github repository (repo) at <https://github.com/robustport/PCRA/>. It contains new features that will be included in the next PCRA CRAN release (currently PCRA-1.4). Furthermore, an important convenience feature of PCRA development versions is that they contain all the CRSP® stocks data and SPGMI factors data sets, namely:

- `stocksCRSPmonthly`
- `stocksCRSPweekly`
- `stocksCRSPdaily`
- `factorsSPGMI`
- `factorsSPGMIr`

However, users of the PCRA development version will need the `factorsSPGMIr` rounded data only if they wish to confirm that its use in portfolio construction does not produce significant differences relative to results obtained with `factorsSPGMI`.

In order to install the development version of PCRA one needs to have installed the R package `devtools`, which can be done in the RStudio console following the procedure used to install for example the `data.table` and `PCRA` packages. Once `devtools` has been installed, the development version of PCRA can be loaded using the following commands in the RStudio console:

```
library(devtools)
devtools::install_github("robustport/PCRA")
```

The main advantages of the development version of PCRA are:

- It contains new features not yet available in the CRAN version of PCRA
- It contains all the CRSP® and SPGMI data sets.

The primary disadvantage of using the PCRA development version is that some new features or data sets may not yet be fully tested or documented.

4 Concluding Comment

If the reader has not already done so, she should read the “CRSP® Stocks and SPGMI Factors in PCRA” vignette and the “PCRA Reproducibility” vignette, which are available at the RPCRA website (<https://robustport.github.io/RPCRAwebsite/>) and at the PCRA CRAN website (<https://cran.r-project.org/package=PCRA>).