README
December 18, 2019
(rev. September 27, 2021)

The CharsAreCov directory contains the following files. They are listed in a "work-flow" order, from data processing to results. The raw data we used was generously given to us in Feburary 2017 by Andreas Neuhierl and subsequent work using these data should reference our paper and Freyberger, Neuhierl, and Weber, "Dissecting Characteristics Nonparametrically," *Review of Financial Studies,* Volume 33, Issue 5, May 2020, pages 2326-2377. In all .m files, file paths should be adjusted as necessary.

- characteristics_data_feb2017.csv, characteristics_data_feb2017.m, characteristics_data_feb2017.mat
    - ".csv" is data received from Andreas Neuhierl
    - ".m" is Matlab code converting it to .mat file. In communication with Andreas (and checking the market cap numbers) we know that the characteristics in the original .csv file are already lagged.
    - ".mat" is the converted data
- IPCA_empirical_datamaker.m, IPCA_empirical_datacall_FNW36.m, IPCADATA_FNW36_RNKDMN_CON.mat
    - "datamaker.m" creates data for estimation. It employs a run statement on line 53
        - many options are given by 0/1 or empty/nonempty are available on lines upto line 42
        - produces an object "Q" that is deprecated; you will see other mentions of "Q" in later files, which are also deprecated
        - Note that time runs columnwise in the eventual data matrices, and in the 3rd dimension for tensors
    - "datacall_FNW36.m" is the Matlab file run on line 53 of "datamaker.m" to deal with the characteristics_data_feb2017.mat data, which we refer to as "FNW36" as an abbreviation for Freyberger, Neuhierl, and Weber (FNW) and the number of characteristics there (36).
    - ".mat" is created by "datamaker.m" with
        - option_rnkdmn_beg = 1
        - option_constant = 1
        - all else zero or empty, whichever means ignore
- IPCA_empirical_GB.m, mySOS.m, num_IPCA_estimate_ALS.m
    - "_GB.m" estimates the "restricted" IPCA model where there is no alpha intercept.
        - Krange on line 4 determines the factor dimensions for which this will be run (define the loop starting on line 9)
        - The data being used is defined as dataname on line 10. Also there are numerical convergence choices there
        - Line 23 gets an initial GammaBeta_initial from the SVD of X

        - GammaBeta, Factor, Lambda are important estimates, on lines 48-50
        - RFITS_GB, XFITS_GB are fitted values for individual assets and managed portfolios, respectively, using actual factor realizations (lines 61, 63, 66, 68). These give Total R2.
        - RFITS_pred_GB, XFITS_pred_GB are fitted values using Lambda, the factors' mean. These give Pred R2.
        - The results are saved on line 85 (see next outer-most bullet)
    - "mySOS.m" is a Frobenius norm that ignores NaNs
    - "ALS.m" is the Alternating Least Squares code that is used to estimate the restricted, unrestricted, and pre-specified factor models
        - Latent factor estimations are in lines 48-51, 54-56
        - Gamma estimation is in lines 62-84
        - Lines 86-100 impose normalization assumptions
        - Lines 102-118 impose normalization assumptions with respect to pre-specified factor estimates (if part of model estimated)
- Results_GB_IPCADATA_FNW36_RNKDMN_CON_K1.mat through "_K6.mat"
    - Saved results from IPCA_empirical_GB.m
- IPCA_empirical_GBGA.m, Results_GBGA_IPCADATA_FNW36_RNKDMN_CON_K1.mat through "_K6.mat"
    - "GBGA.m" estimates the "unrestricted" IPCA model with an alpha intercept
        - Similar to "_GB.m" code
        - The alpha intercept is estimated by sending in a ones vector as a pre-specified factor (line 63, the last argument sent into num_IPCA_estimate_ALS is "ones(1,T)")
    - ".mat" files are the results
- IPCA_empirical_GBGAtest.m, IPCA_empirical_GBGAtest.sh
    - ".m" implements the alpha Wald-like test via bootstrap in lines 123-166
    - ".sh" is an slurm sbatch script that can be used with the ".m" (adjusted for the proper cluster directories, etc). This allows a smaller number of bootsims to be done by each instance of the ".m" file, and then the results can be compiled.
    - If the ".sh" file is not used, then uncomment lines 5-8 and fill with appropriate values
    - To find the boostrapped p-value, do the following.
        - Suppose that boot_GA is a matrix with L columns and as many rows as bootstraps. Suppose that GBGA_GA is the L-vector GammaAlpha estimate (GBGA_GA).
        - Run mean(sum(boot_GA.^2,2) > mySOS(GBGA_GA)) to calculate the bootstrapped p-value
- Observable_Factor_Regressions.m, F-F_Research_Data_5_Factors_2x3_plusMOM.mat, M4.mat, HXZ_q-Factors_monthly.mat, BarillasShanken.mat
    - ".m" estimates FITS_ and FITS_cond from observable factor models employing regession-betas, where FITS_ use the factor realization itself – they give Total R2. FITS_cond use the factor mean – they give Pred R2.

        - dataname on line 7 is use to bring in the asset returns, LOC matrix, and date vector. It can remain as is, if QXorR = 'R' on line 9. This generates individual asset results.
        - oosflag = true on line 8 creates out-of-sample results; oosflag = false creates in-sample results
        - annualize = false should be left to generate monthly results
        - ObsChoice on line 11 chooses which observable factors to use
        - The out-of-sample results can use recursive (expanding) windows or rolling windows. We explain using the Fama-French factor code in lines 86-103. stind = 1 on line 91 creates a recursive window, whereas stind = t-60 would set up a 60-month rolling window. oossuffix on line 103 should be changed to appropriately name the specification
    - The various ".mat" files contain the observable factors
- IPCA_empirical_GBGDFF.m
    - This produces results using only latent factors (GB), using latent factors and observable factors (GBGD), and using only observable factors (GD)
- IPCA_empirical_GB_outofsample.m, tanptf.m tanptfnext.m, Results_GB_outofsample_IPCADATA_FNW36_RNKDMN_CON_K1 through "_K6.mat", Results_ObsFactRegROOS_Results_GB_IPCADATA_FNW36_RNKDMN_CON_K1_rec_60_60.mat, Results_ObsFactRegXOOS_Results_GB_IPCADATA_FNW36_RNKDMN_CON_K1_rec_60_60.mat, Table_OutOfSample_Fits.m,
    - "_outofsample.m" creates out-of-sample predictive fits for asset and managed portfolios (lines 71-72), as well as realized "out-of-sample" factors and tangency portfolio returns (lines 75-76) and asset and managed portfolio returns (lines 78-79). The predictive fits OOSRFITS_pred_GB and OOSXFITS_pred_GB (lines 71-72) provide the out-of-sample predictive R2 in Table 5, while the out-of-sample fits OOSRFITS_GB and OOSXFITS_GB (lines 78-79) provide the total R2. The tangency portfolio in line 76 gives the returns whose Sharpe ratio is reported in Table 6 for IPCA.
    - tanptf.m and tanptfnext.m constructs tangency portfolio weights from a time series matrix of returns; the latter file then applies those weights to a separate vector of returns.
    - "K1-K6.mat" are the out-of-sample results from the OOS IPCA procedure
    - "Results_Obs*.mat" files are out-of-sample results created by Observable_Factor_Regressions.m as given in the directory. Notice that this uses the "rec_60_60" out-of-sample specification, which produces recursive/expanding window forecasts that require at least 60 obs to be included
    - Table_OutOfSample_Fits.m creates the OOS R2 results in Table 5 and OOS SR results in Table 6
- Table_pure_alpha.m, Results_GBGA_outofsample_IPCADATA_FNW36_RNKDMN_CON_K1 through "_K6.mat"
    - Produces Table 7 results, from out-of-sample estimates of the extended model with both Gamma_beta and Gamma_alpha

20210927: Thank you Chosen Zhou for pointing out a mistake in the ordering of the Gammadelta identification code in num_IPCA_estimate_ALS.m (in the loop starting at line 103) wherein Gammadelta was orthogonalized with respect to Gammabeta BEFORE being used to rotate the F_New estimate.