==========

README: PHYCAA+ physiological correction scripts (2014/04/10)

==========

Author: Nathan Churchill, University of Toronto
email: nchurchill@research.baycrest.org

==========

This is a data-driven multivariate algorithm, used to remove physiological noise in BOLD fMRI data. It estimates noise directly from the data without requiring any external measurements, and doesn't require manual component selection. The algorithm and model details provided in:

> *Churchill & Strother (2013). "PHYCAA+: An Optimized, Adaptive Procedure for Measuring and Controlling Physiological Noise in BOLD fMRI". NeuroImage 82: 306-325*

Please cite this article if code is used in any publications.

README CONTENTS:

NOTE: there are 2 different ways to run the PHYCAA+ code. Sections III-IV.
explain how to perform denoising if:

(III.) you are doing single-subject denoising, on 4D NIFTI data
(IV.) you are doing multi-subject denoising on 4D NIFTI data (e.g. prior to a group-level analysis), and require a consistent vascular mask across subjects

---

I. ALGORITHM AND SCRIPT OVERVIEW

Matlab scripts:

run_PHYCAA_plus.m : *used to perform individual subject denoising (section III.)*
run_PHYCAA_plus_group.m : *used to perform group-level denoising (section IV.)*

These are wrapper scripts which make use of the following functions:

PHYCAA_plus_step1.m : *first step of the algorithm; masks out high-variance vascular tissues*
PHYCAA_plus_step2.m : *second step of the algorithm; regresses out physiological timeseries*
NN_group_average.m : *for group-level analysis; estimates a consensus maps of vascular tissue*

PHYCAA+ is a 2-stage algorithm that estimates and removes physiological noise. Each stage is implemented in a separate Matlab script ("PHYCAA_plus_step1.m" and "PHYCAA_plus_step2.m" respectively). The general algorithm procedure is outlined below.

STEP-1: Identifies voxels containing probable non-neuronal tissues in the brain (e.g. vasculature, sinuses, ventricles) and produces a brain "weighting map", with values ranging from 1=neuronal tissue, to 0=high likelihood of non-neuronal tissue content. This is used to down-weight, i.e. reduce the signal amplitude of voxels with high non-neuronal tissue content. This model:

1. Sorts voxels by high-frequency power (e.g. f>0.10 Hz), which reflects amount of non-neuronal signal content. This produces a curve of (voxel percentile) vs. (high-frequency power).
2. The curve has a linear region (neuronal tissue), and non-linear region (presence of non-neuronal signal). It identifies the "transition point" between these regions (delta_min) at p<.01.
3. It then identifies the threshold on high-frequency power for "pure" non-neuronal signal ($\boldsymbol{\delta}_{max}$). This is done by either (a) maximizing overlap with an anatomical prior (e.g. a vascular atlas), or (b) choosing the 95th percentile of high-frequency power (threshold identified in Churchill & Strother, 2013)
4. The model weights voxels based on the deviation ($\boldsymbol{\delta}$) from linearity on a range of [0,1] where:

   if($\boldsymbol{\delta} < \boldsymbol{\delta}_{min}$) weight = 1
   if($\boldsymbol{\delta}_{min} < \boldsymbol{\delta} < \boldsymbol{\delta}_{max}$) weight = $(\boldsymbol{\delta}_{max} - \boldsymbol{\delta})/(\boldsymbol{\delta}_{max} - \boldsymbol{\delta}_{min})$
   if($\boldsymbol{\delta} > \boldsymbol{\delta}_{max}$) weight = 0

   Brain voxels are scaled (multiplied) by these weights. Thus, neuronal tissue(=1) is unchanged, while voxels with vascular content show reduced amplitude. Regions with high likelihood of non-neuronal tissue(=0) are effectively masked out.

STEP-2: estimates a set of physiological component timecourses that are present in the brain, using the non-neuronal map from "PHYCAA_plus_step1.m" as a spatial prior. These timecourses are then regressed out at each voxel. This model:

1. Takes in an initial estimate of BOLD activation pattern(s) (SPMs) and regresses out their signal timecourses. Physiological noise is estimated on the "residual" data. This is done to decorrelate physiological noise with the BOLD signal of interest. Otherwise, you risk over-regressing your data, as physiological effects are often partially coupled to neuronal response. ***This is an optional step, but highly recommended!*** *
2. Transforms residual data into Principal Component (PC) space, to optimize data dimensionality for physiological estimation
3. Tests PC subspace sizes 1-K (where K varies up to 50% of maximum data dimensionality). For each K, it identifies significantly autocorrelated components using Canonical Autocorrelation Analysis. It then uses the non-neuronal brain map (of STEP-1), to identify components that explain the greatest median variance in non-neuronal tissues. These components are retained as physiological timeseries. It also measures a brain map of total variance explained by the physio. components. This is done for >=2 runs of the data, and reproducibility measured between the runs' variance maps
4. It then selects the physiological components for PC dimensionality K that maximizes spatial reproducibility, as physiological noise is expected to originate from consistent brain regions.

These components are then regressed out of the data, and voxel values down-weighted, using the weighting map from STEP-1. This step controls for physiological noise effects that are present in neuronal brain tissue (e.g. grey matter).

The wrapper scripts requires Jimmy Shen's tools for loading nifti files
(www.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-analyze-image)
A "NIFTI_tools" folder is included with the latest scripts, just make sure it is in your path if needed.

II. GENERAL REQUIREMENTS

1. FMRI data sets, in NIFTI format (with a .nii extension). All data run through PHYCAA+ must be in the same anatomical coordinate space, i.e. the same spatial normalization
2. Binary brain mask, in NIFTI format. Must be a single, common mask for all datasets being denoised.
3. Vascular/CSF mask (**OPTIONAL**), in NIFTI format. Can be used to improve the accuracy of data-driven vascular maps (STEP-1 of PHYCAA+). Must be a binary map, where 1=vascular and/or CSF tissues; 0=neuronal tissue
4. Activation SPMs (**OPTIONAL**), in NIFTI format. Protects against over-regression of task-linked signal in STEP-2. Can be results of any analysis model of interest (e.g. GLM, LDA, SVM, PLS…)

III. RUNNING PHYCAA+ ON INDIVIDUAL SUBJECTS

Takes in data from a single subject (and possibly multiple runs), and performs physiological denoising. Input/output syntax is shown below. See "PSEUDO-CODE EXAMPLE" afterwards for an example of how to run the scripts:

```
%========================================================================
%  RUN_PHYCAA_PLUS: wrapper script that runs the 2-stage PHYCAA+ algorithm,
%  used to correct for physiological noise. Takes in names of 4D fMRI data
%  in NIFTI format, and outputs denoised NIFTI data.
%========================================================================
%
% SYNTAX:
%
%   run_PHYCAA_plus( input_cell, mask_name, prior_name, task_SPM_names, dataInfo, num_steps,
out_prefix )
%
% INPUT:
%
%   input_cell     = cell array of strings, each giving path/name of a run
%                    of fMRI data
%                     e.g.  input_cell{1}= 'my_directory/subject1_run1.nii'
%                           input_cell{2}= 'my_directory/subject1_run2.nii'
%                             ....
%                    If you only have 1 run, it will do split-half noise
%                    estimation (e.g. slightly different than multi-run)
%                    Multiple runs do not require the same number of
```

```
%                     timepoints (but they should be close to ensure stability)
%   mask_name       = string specifying path/name of binary brain mask used
%                     to remove non-brain tissue - required to run PHYCAA+.
%   prior_name      = Parameter of STEP-1. OPTIONAL string specifying
%                     path/name of binary brain mask with probable non-
%                     neuronal tissue locations (e.g. a CSF atlas). If
%                     included, this step selects a threshold for masking
%                     out voxels (voxel weight =0) that maximizes overlap
%                     with the prior mask. Otherwise, the model masks out
%                     the top 5% of voxels (average prior-based threshold
%                     idenfified in Churchill & Strother 2013). If not
%                     included, leave this entry empty (e.g. prior_name=[])
%   task_SPM_names = Parameter of STEP-2. An OPTIONAL single string or cell
%                     array of strings; each giving the path/name of an
%                     analysis SPM, used to estimate the residual subspace
%                     in Step-2 of PHYCAA+, before physio. component estimation.
%                       e.g.   task_SPM_names    = 'my_directory/spm.nii'
%
%                         or    task_SPM_names{1} = 'my_directory/spmA.nii'
%                               task_SPM_names{2} = 'my_directory/spmB.nii'
%                                ...
%                     if you don't want to estimate the residual subspace,
%                     leave this entry empty (e.g. task_SPM_names = []).
%    dataInfo     =  a structure with the following fields. Only "TR" must
%                     be specified, the rest are optional:
%
%                dataInfo.TR  : Parameter of STEP-1. Rate of data acquisition (in sec.)
%          dataInfo.FreqCut    : Parameter of STEP-1. High-frequency threshold in Hz, for which
%                                f > FreqCut is primarily  physiological noise. If not given
%                                DEFAULT value is FreqCut=0.10.
%          dataInfo.comp_crit  : Parameter of STEP-2. Determines how conservative physiological
%                                component selection is. Value can range (0 <= comp_crit < 1),
%                                where larger comp_crit indicates more conservative selection,
%                                i.e. variance must be more concentrated in non-neuronal tissue
%                                DEFAULT value is comp_crit=0, which gives robust results.
%          dataInfo.keepmean   : Parameter of STEP-2. PHYCAA+ subtracts voxel mean of each run,
%                                before component estimation.
%                                         0= voxel means are discarded
%                                         1= voxel means re-added after noise regression
%                                If not specified, DEFAULT value is keepmean=0.
%          dataInfo.make_output: A general output parameter.
%                                  0= do NOT produced preprocessed data (just physio maps etc.)
%                                  1= preprocessed data as well (down-weight and/or regressed)
%                                If not specified, DEFAULT value is make_output=1; preprocessed
%                                data are automatically created
%
%   out_prefix    = prefix for physio output. If empty (out_prefix=[]),
%                   DEFAULT prefix is 'PHYCAA_new'
%   num_steps     = integer indicates how many steps of PHYCAA+ to perform:
%                     1= do STEP1 only,  output (a) non-neuronal map, (b) downweighted data
%                     2= do STEP1+STEP2, output (a) non-neuronal map, (b) physio component
%                        map, (c) downweighted+regressed data
%
% OUTPUT:
%         if( num_steps==1 )
%            (1) IF( dataInfo.out_format==1 ), downweighted fMRI datasets, denoted
%                [input_cell{r}(1:end-4),'_PHYCAA_step1.nii']
%            (2) non-neuronal weighting map, as a NIFTI volume [out_prefix,'_NN_map.nii']
%            (3) matfile [out_prefix,'_PHYCAA_step1.mat'], containing output of PHYCAA+ step-1
%
%         if( num_steps==2 )
%          (1) IF( dataInfo.out_format==1 ), downweighted+regressed fMRI datasets, denoted
%              [input_cell{r}(1:end-4),'_PHYCAA_step1+2.nii']
```

```
%          (2) non-neuronal weighting map, as a NIFTI volume [out_prefix,'_NN_map.nii']
%          (3) Z-scored map of physiological component variance, as NIFTI volume
%              [out_prefix,'_Physio_Zmap.nii']
%          (4) matfile [out_prefix,'_PHYCAA_step1+2.mat'] containing output of PHYCAA+steps1,2
%
```

PSEUDO-CODE EXAMPLE

Because PHYCAA+ estimates physiological noise that is decorrelated with BOLD signal in Step-2, it is dependent on your choice of analysis model. In the pseudo-code example, we assume a single subject, with two runs of data. We also assume you have already run some analysis model on the data, and produced an activation map "signal_SPM.nii". You can use any analysis model that produces a map (or set of maps) of brain activation (e.g. GLM, linear discriminant, SVM, etc.).

```
% (1) define file names %
input_cell{1} =  'mydata/sub1_run1.nii'; % data run1, subject1
input_cell{2} =  'mydata/sub1_run2.nii'; % data run2, subject1
mask_name     = 'mydata/brain_mask.nii'; % 3d binary brain mask
task_SPM_names{1} = 'signal_SPM.nii'; % analysis SPM, without denoising
out_prefix    =        'mydata/sub_1'; % prefix for physio. data output

% only necessary to specify "TR" parameter of dataInfo
dataInfo.TR        = 2.0;

% (2) run PHYCAA+ steps 1-2: obtain downweighted + regressed fMRI data
%     NB: we are running with 'signal_SPM.nii' added, and num_steps=2

run_PHYCAA_plus( input_cell, mask_name, [], task_SPM_names, dataInfo, 2, out_prefix );

% output:
% . down-weighted+regressed data,
%                     "mydata/sub1_run1_PHYCAA_step1+2.nii"
%                     "mydata/sub1_run2_PHYCAA_step1+2.nii"

% (3) Data is now de-noised! Re-analyze the fullly de-noised data
%
%                     "mydata/sub1_run1_PHYCAA_step1+2.nii"
%                     "mydata/sub1_run2_PHYCAA_step1+2.nii"
%    with analysis model X:

   (... re-run your analysis model ...)

% output:
% . improved estimate of brain activation volume "signal_SPM.nii"
```

IV. RUNNING PHYCAA+ ON GROUPS

Takes in data from a multiple subjects (and possibly multiple runs), and performs physiological denoising on all of them at once. Input/output syntax is shown below. See "PSEUDO-CODE EXAMPLE" afterwards for an example of how to run the scripts:

```
%========================================================================
%  RUN_PHYCAA_PLUS_GROUP: wrapper script that runs the 2-stage PHYCAA+
%  algorithm to correct for physiological noise, for multiple subjects
%  simultaneously. Estimates a group average non-neuronal map, for cases
%  where a single, consistent vascular mask is required for all subjects
%========================================================================
%
% SYNTAX:
%
%   run_PHYCAA_plus_group( input_cell, mask_name, prior_name, task_SPM_names, dataInfo,
num_steps, out_prefix )
%
% INPUT:
%
%   input_cell    =  2-level cell array of strings. Each cell at first level
%                    corresponds to a subject, and every subject cell contains
%                    an array, indicating path+name of a run of fMRI data.
%                     e.g.   input_cell{1}{1}=  'my_directory/subject1_run1.nii'
%                            input_cell{1}{2} = 'my_directory/subject1_run2.nii'
%                                ...
%                            input_cell{2}{1} = 'my_directory/subject2_run1.nii'
%                            input_cell{2}{2} = 'my_directory/subject2_run2.nii'
%                                ...
%                            input_cell{3}{1} = 'my_directory/subject3_run1.nii'
%                            input_cell{3}{2} = 'my_directory/subject3_run2.nii'
%                                ...
%                    requires AT LEAST 2 subjects to run. Does NOT require
%                    the same number of timepoints across subjects/runs
%                    (though they should be close to ensure stability)
%   mask_name      = string specifying path/name of binary brain mask used
%                    to remove non-brain tissue - required to run PHYCAA+.
%                    ** NB: one brain mask for all subjects/runs;
%                           everything should be in same coordinate space!
%   prior_name     = Parameter of STEP-1. OPTIONAL string specifying
%                    path/name of binary brain mask with probable non-
%                    neuronal tissue locations (e.g. a CSF atlas). If
%                    included, this step selects a threshold for masking
%                    out voxels (voxel weight =0) that maximizes overlap
%                    with the prior mask. Otherwise, the model masks out
%                    the top 5% of voxels (average prior-based threshold
%                    idenfified in Churchill & Strother 2013). If not
%                    included, leave this entry empty (e.g. prior_name=[])
%   task_SPM_names = Parameter of STEP-2. An OPTIONAL single string or cell
%                    array of strings; each giving the path/name of an
%                    analysis SPM, used to estimate the residual subspace
%                    in Step-2 of PHYCAA+, before physio. component estimation.
%                      e.g.   task_SPM_names    = 'my_directory/spm.nii'
%
%                        or    task_SPM_names{1} = 'my_directory/spmA.nii'
%                              task_SPM_names{2} = 'my_directory/spmB.nii'
%                               ...
```

```
%                    if you don't want to estimate the residual subspace,
%                    leave this entry empty (e.g. task_SPM_names = []).
%    dataInfo     =  a structure with the following fields. Only "TR" must
%                    be specified, the rest are optional:
%
%      dataInfo.TR         : Parameter of STEP-1. Rate of data acquisition (in sec.)
%      dataInfo.FreqCut    : Parameter of STEP-1. High-frequency threshold in Hz, for which
%                            f > FreqCut is primarily  physiological noise. If not specified,
%                            DEFAULT value is FreqCut=0.10.
%      dataInfo.comp_crit  : Parameter of STEP-2. Determines how conservative physiological
%                            component selection is. Value can range (0 <= comp_crit < 1),
%                            where a larger comp_crit indicates more conservative selection,
%                            i.e. variance must be more concentrated in non-neuronal tissue.
%                            DEFAULT value is comp_crit=0, which gives robust results.
%      dataInfo.keepmean   : Parameter of STEP-2. PHYCAA+ subtracts voxel means of each run,
%                            before component estimation.
%                                    0= voxel means are discarded
%                                    1= voxel means re-added after noise regression
%                            If not specified, DEFAULT value is keepmean=0.
%     dataInfo.make_output: A general output parameter.
%                             0= do NOT produced preprocessed data (just physio. maps etc.)
%                             1= produce preprocessed data too (down-weight and/or regressed)
%                           If not specified, DEFAULT value is make_output=1; preprocessed
%                           data are automatically created
%
%   out_prefix    = prefix for physio output. If empty (out_prefix=[]),
%                   DEFAULT prefix is 'PHYCAA_group_new'
%   num_steps     = integer indicating how many steps of PHYCAA+ to perform:
%                     1= do STEP1 only,  output (a) non-neuronal map, (b) downweighted data
%                     2= do STEP1+STEP2, output (a) non-neuronal map, (b) physio component
%                        map, (c) downweighted+regressed data
%
% OUTPUT:
%        if( num_steps==1 )
%     (1) IF( dataInfo.out_format==1 ), downweighted fMRI datasets, denoted
%                          [input_cell{r}(1:end-4),'_PHYCAA_step1_group.nii']
%     (2) non-neuronal weighting map, as a NIFTI volume [out_prefix,'_NN_map_avg.nii']
%     (3) matfile [out_prefix,'_PHYCAA_step1_group.mat'], containing output of PHYCAA+ step-1
%
%        if( num_steps==2 )
%   (1) IF( dataInfo.out_format==1 ), downweighted+regressed fMRI datasets, denoted
%       [input_cell{r}(1:end-4),'_PHYCAA_step1+2_group.nii']
%   (2) non-neuronal weighting map, as a NIFTI volume [out_prefix,'_NN_map_avg.nii']
%   (3) Z-scored map of physiological component variance, as NIFTI volume
%       [out_prefix,'_Physio_Zmap_avg.nii']
%   (4) matfile [out_prefix,'_PHYCAA_step1+2_group.mat']containing output of PHYCAA+ steps1,2
```

PSEUDO-CODE EXAMPLE

In this example, dataset consists of 10 subjects, with 2 data runs each. Because PHYCAA+ estimates physiological noise that is decorrelated with BOLD signal in Step-2, it is dependent on your choice of analysis model. For the pseudo-code, we assume you have already run some analysis model on the data and produced an activation map "group_signal_SPM". You can use any analysis model that produces a map (or set of maps) of brain activation (e.g. GLM, linear discriminant, SVM, etc.):

```
% (1) define file names %

% path for subject 1 data %
input_cell{1}{1}  =  'mydata/sub1_run1.nii'; % data run1, subject1
input_cell{1}{2}  =  'mydata/sub1_run2.nii'; % data run2, subject1
% path for subject 2 data %
input_cell{2}{1}  =  'mydata/sub2_run1.nii'; % data run1, subject2
input_cell{2}{2}  =  'mydata/sub2_run2.nii'; % data run2, subject2
 ...
% path for subject 10 data %
input_cell{10}{1} =  'mydata/sub10_run1.nii'; % data run1, subject10
input_cell{10}{2} =  'mydata/sub10_run2.nii'; % data run2, subject10

task_SPM_names{1} = 'group_signal_SPM.nii'; % analysis SPM, without denoising
mask_name    =  'mydata/brain_mask.nii'; % define group 3d brain mask
out_prefix   =    'mydata/group_run12'; % prefix for physio. data output

% only necessary to specify "TR" parameter of dataInfo
dataInfo.TR        = 2.0;

% (2) run PHYCAA+ steps 1-2: obtain downweighted + regressed fMRI data
%    NB: we are running with 'group_signal_SPM.nii' added, and num_steps=2

run_PHYCAA_plus_group( input_cell, mask_name, [],task_SPM_names, dataInfo, 2, out_prefix );

%  output:
%  . downweighted+regressed data,
%                       "mydata/sub1_run1_PHYCAA_step1+2_group.nii"
%                       "mydata/sub1_run2_PHYCAA_step1+2_group.nii"
%                         ...

% (3) Data is now de-noised! Re-analyze the full set of de-noised data
%
%                       "mydata/sub1_run1_PHYCAA_step1+2_group.nii"
%                       "mydata/sub1_run2_PHYCAA_step1+2_group.nii"
%                         ....
%    with the chosen analysis model X:

   (... re-run multi-subject analysis model...)

% output:
% . improved estimate of brain activation volume "group_signal_SPM_group.nii"
```

V. USEFUL NOTES

A few points to consider when running PHYCAA+:

1. As previously noted, it is *highly* recommended that you do an initial analysis on whatever you think your task signal is and pass the resulting SPMs to PHYCAA+, so that you can get an estimate of signal subspace. This gives a much "cleaner" separation of signal and physiological noise

2. If your initial SPM estimates are highly noise-confounded, with lots of vascular signal, we recommend that you do an initial analysis with down-weighting only, i.e. run_PHYCAA_plus, on the setting "num_steps=1". You can re-analyze your data to get cleaner SPMs, which can be used as input for the full PHYCAA+ denoising procedure.

3. PHYCAA+ was originally designed as an *iterative* algorithm. This means that you can rerun your data multiple times through PHYCAA+ to get improved noise correction. The only requirement is that you re-analyze your data for each iteration; you will need to get a new estimate of the "improved" signal subspace each time. You can iterate " run phycaa+" / "run analysis" steps until the algorithm can no longer identify physiological components (it will give 0 components in the output and terminate). Please refer to our paper (Churchill & Strother 2013) if more information is needed.

4. If you have any concerns about the quality of noise estimation, we recommend looking at the output physiological maps, including the vascular mask ([out_prefix,'_NN_map.nii']) and the Z-scored map of physiological component variance ([out_prefix,'_Physio_Zmap.nii']). If the quality of the maps looks poor across subjects, this may indicate either (a) you need to do some extra preprocessing before this step, e.g. better motion correction, or (b) your initial signal estimates (SPMs) are too noisy – see point (2) above for a possible solution.