The Art and Pitfalls of fMRI pre-processing: 
*Introduction and simple theoretical considerations*

Christian Habeck
Cognitive Neuroscience Division
Department of Neurology, Columbia University, NY 10032
Purpose of the course

• Educate fMRI practitioners about
  • **Standard pre-processing tools and software (main focus)**
    • Unresolved issues (raise awareness, but leave solution for later)

• Ideally, enable practitioners to engage in methodological research themselves
Course Schedule

8:00-8:30
Introductory remarks: the problem of pre-processing pipeline dependence for task-based and resting-state fMRI
Christian Habeck, Columbia University, Neurology, New York, NY, United States

8:30-9:00
Temporal Preprocessing (slice-timing, temporal filtering, spike removal)
Yunjie Tong (formerly: Blaise Frederick), McLean Hospital, Belmont, MA, United States

9:00-9:30
Spatial Preprocessing (Spatial Alignment, Normalization, and Smoothing)
Ray Razlighi, PhD., Columbia University, New York, NY, United States

9:30-10:00
Artefact Removal (motion-related)
Christian Windischberger, MR Center, Medical University of Vienna, Vienna, Austria

10:00-10:30
Break

10:30-11:00
Artefact Removal (Physiological)
Rasmus Birn, University of Wisconsin-Madison, Madison, WI, United States

11:00-11:30
FSL Pre-Processing Pipeline
Mark Jenkinson, University of Oxford, Oxford, United Kingdom

11:30-12:00
An SPM Perspective on fMRI Preprocessing
Lars Kasper, University of Zurich & ETH Zurich, Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, Zurich, Switzerland
Why is pre-processing necessary?

The raw fMRI signal is influenced by several factors that are statistical or subject-dependent in nature:

- Temporal offset between slice acquisitions
- Subject motion
- Subject respiration and heart rate
- Subject’s brain size, shape
- Scanner noise and field inhomogeneity

etc.....

making analysis across time or subjects impossible without pre-processing.
Lack of overall variance context

Many reviews, but recent paper gives nice summary of signal contamination and clean-up methods

Murphy, Birn, Bandettini, NeuroImage 80 (2013) 349–359

➤ Interaction between different processing states and relative variance considerations are difficult to perform and mostly lacking
Typical within-subject pre-processing pipeline

Correction for motion (scrubbing/regression)
Temporal band pass filtering
Co-registration to structural
Re-alignment
Slice timing correction

Successive discrete processing stages transforming raw signal \( r(t) \) into pure signal \( s(t) \)

\[
s(t) = F_5 \circ F_4 \circ F_3 \circ F_2 \circ F_1 \circ r(t)
\]

Alternatively:

\[
F_5(F_4(F_3(F_2(F_1(r))))))
\]
Typical within-subject pre-processing pipeline

1. Correction for motion (scrubbing/regression)
2. Temporal band pass filtering
3. Co-registration to structural
4. Re-alignment
5. Slice timing correction

Successive discrete processing stages transforming raw signal \( r(t) \) into pure signal \( s(t) \)

\[
s(t) = F_5 \circ F_4 \circ F_3 \circ F_2 \circ F_1 \circ r(t)
\]

Alternatively:

\[
F_5(F_4(F_3(F_2(F_1(r))))))
\]
Obvious optimization quest

• What are the “best” functions/algorithms $F$?
  • What are the best parameter settings?

➤ This nuance is not very important: for functions that acquire an additional setting of a parameter the difference in the parameter setting can be interpreted as leading to a different function

➤ However, this is enough; the order of processing stages needs to be optimized too
Why is optimization of complete pipeline, including order of stages, strictly necessary?

Successive discrete processing stages

\[ s(t) = F_5 \circ F_4 \circ F_3 \circ F_2 \circ F_1 \circ r(t) \]

Nested mathematical formulations \textit{cannot} be true since:
- Different artifacts interact with one another (e.g. slice timing and motion)
- Artifacts are happening in parallel at all times

\[ \Rightarrow \text{Unless all functions } f(n) \text{ are linear (=ordering irrelevant), the ordering has to be optimized too} \]

\[ \Rightarrow \text{There is no apriori optimal ordering of the tools in the pipeline} \]
Further: pre-processing pipelines for task-based fMRI are no guide for resting fMRI

Outcome of interest for

**Task fMRI:** task activation, i.e. correlation of voxel activity $x$ with task design $d$ (1\textsuperscript{st} order moment)

\[ x' d \]

**Resting fMRI:** functional connectivity, i.e. correlation of between voxel activities $x$ and $y$ (2\textsuperscript{nd} order moment)

\[ x' y \]
Pre-processing pipelines for task-based vs. resting fMRI

Simple pseudo notation for variational derivative with respect to pre-processing functions/parameters $u$ for finding zero gradients

Task-fMRI:

$$\delta_u (x' \ d) = (\delta_u x)' \ d$$

Resting fMRI – product rule applies:

$$\delta_u (x' \ y) = (\delta_u x)' \ y + x' (\delta_u y)$$

→ Different orders (1$^{st}$ vs. 2$^{nd}$) invalidate generalization of optimized task-based pipelines to resting-state fMRI
Recall of these simple theoretical insights

Form the previous considerations, it follows that -

• Breaking out the pipeline into separate modules is strictly speaking incorrect
• No optimal apriori ordering of modules can be derived; optimization has be made empirically
• Optimization for first-order signal moments (=task-related activation) might give different results from higher-order moments (=connectivity)
How to proceed?

Optimization should give answers to the questions:
• Which tools to use?
• With what parameters?
• In what order?

➡️ Before an informed search of prohibitively large search space can be attempted, there is an additional question:
What quality criterion should inform the judgment of good pipeline performance?

**Simulated data** (=gold-standard knowledge exists for judging performance)
- Residual sum of squares of signal in simulated data (minimize)
- False positives, false negatives (minimize)

**Real-world data**
- Replication of activation between data folds (maximize)
- Replication of functional connectivity between data folds (maximize)
- Correlation between replication quality and primary-interest measures (minimize)
- Prediction of subject identity (or task) in one data fold, after training a model in the other data fold (maximize)

→ The questions touch on apriori value judgments and inform, but are not informed by, the tools to be used.

HBM 2015, Educational Course: The Art and Pitfalls of fMRI pre-processing
Some questions to leave you with – some of which will not be answered the end of today

• What is the best pipeline for removing artefacts?
  • The pipelines constitute of procedures/algorithms arranged in a certain order, with certain parameters for each algorithm

• What are the metrics used for judging good performance?
  • The decision of what constitutes good performance might use metrics such as $R^2$ in simulations, and false positives/negatives in real-world data.

• Should some standard tools be left out entirely since they cause more harm than help?
  ➔ Some of these questions have to be answered empirically, with simulations or real-world data, possibly on a case-by-case basis
Some questions to leave you with – some of which will not be answered the end of today

The prime analysis framework with software support for task fMRI is:

**NPAIRS** (Nonparametric, Prediction, Activation, Influence, Reproducibility, re-Sampling)

[https://code.google.com/p/plsnpairs/](https://code.google.com/p/plsnpairs/)

Author: Stephen Strother, Rotman Institute, University of Toronto with students, colleagues
One simple example of deficient pre-processing in resting BOLD

Look at influence of quality of replication of global connectivity between split halves on selected connectivity

• Replication of global connectivity between data folds captured with mean $\eta^2$
• Outcome of interest: connectivity within visual resting-state network

⇒ Computational details are not very important here
Data robustness influences correlation strength, measurement bias
Connectivity in VIS network

Replication between data folds

swP<0.001

Scheiβe!
Take it away speakers!