Supplementary Material

Table S1 – Frequently asked questions. The variety of considerations relevant to the topic of mean task-evoked responses potentially inflating task-state FC estimates are substantial. We therefore include this question-driven organization of relevant considerations to provide a more direct means for frequent questions to be addressed.

Questions	Answers
1) Why remove mean evoked responses prior to task-state FC analysis?	So that the amount of inter-node linear interaction can be estimated, above and beyond mere task co-activations. Rather than the ambiguous inference that two regions are "likely active *or* interacting during the task", removing mean evoked responses allows for the more specific inference, "likely interacting during task". Given that most neuroscience studies have focused on mean evoked responses, it is important to remove these responses to reduce the risk of reporting mean evoked responses (mere first-order changes in activity level) as "connectivity" results (second-order effects).
2) What other reasons are there for removing the mean evoked responses prior to task-state FC analysis?	Task-state FC can be conceptualized as an interaction between psychological and physiological factors – a "psychophysiological interaction" (PPI) (Friston et al., 1997; McLaren et al., 2012). As with most interactions, it is important to account for the main effects prior to interpreting the interaction. This ensures the estimated interaction is not just the main effect (here, mean evoked response) masquerading as an interaction. Additionally, task-state FC can be conceptualized in terms of causal inference, in which the main effect of task events is a confound for estimating the interaction between pairs of nodes. This is clearly demonstrated in the "no connectivity zone" example in Figure 3, wherein the main effect produces correlations in nodes with no physical means of interacting.
3) How can the task-timing confounding be corrected?	As with all confounds, holding the confounding factor constant can remove its effect (Pearl, 2009). The most straightforward strategy for holding stimulus/task state constant is to only include time points from that state – accounting for the main effect of task state. Critically, however, neural processing involves temporal autocorrelation, such that the effects of rest-to-task (or task-to-task) state changes extend beyond stimulus presentation (state-transition transients). Both models reported here indicate that the temporal autocorrelation introduced by the fMRI BOLD response is especially problematic in this respect (see Figure 2). Another strategy is therefore necessary, and we advocate subtraction of an estimate of the temporally-extended response to the stimulus. We also show that flexible fitting of the task-evoked response is critical for properly removing the task-timing confound for fMRI analyses.
4) Does removing the mean evoked response remove too little variance (leaving false positives)?	This is unlikely, given that all false positives were removed in the minimal and neural mass models following subtraction of mean evoked responses. This result appears to rely on several prerequisites: 1) stimulus identity and timing are identical across event instances, 2) enough time points are included in the mean evoked response estimate to cover the temporally-extended response to the stimuli, 3) more than one event instance is included in the mean evoked response estimate (to allow separation of each evoked response from the mean response), 4) an accurate estimate of the evoked response shape is used (e.g., using FIR regression), and 5) enough temporal separation between events of different conditions to ensure separate mean evoked response estimates and correlation estimates. Temporal jitter between stimuli can be useful for mean evoked response estimation (especially when linear regression is used) (Miezin et al., 2000), but this is not helpful for correlation estimation. Block designs (with or without temporal jitter) may be most appropriate, given that this allows removal of both transient and sustained evoked responses (Al-Aidroos et al., 2012; Visscher et al., 2003) with enough temporal separation between condition types.

5) Does removing the mean evoked response remove too much variance (inducing false negatives)?	This is unlikely, given that the neural mass computational model indicated a *reduction* in false negatives following removal of the mean evoked responses. Yet it is technically possible that two nodes could have a highly stereotyped interaction time-locked to the stimuli, producing covariance that is included in the mean evoked response. In this case, removing the mean evoked response from both nodes would inappropriately reduce the detected covariance. However, since neural populations are known to always have variability across event instances, moment-to-moment and event-to-event covariance would remain, such that the proper level of covariance would likely be detected. Put another way, for a signal to be removed by subtracting the mean evoked response that signal would need to be time-locked to the task and 100% consistent in amplitude with the mean evoked response across events. This appears unlikely given the inherent variability of neural responses. Supporting this conclusion, we found that ~90% of inter-region covariance remained in fMRI data after removing the mean evoked response from all nodes (see Supplementary Materials section, pg. 8). In other words, if two nodes are truly interacting, evidence for such interaction is very likely to be present in the moment-to-moment and event-to-event variance left over after removal of the mean evoked responses.
6) Do other sources of confounding remain after removal of mean evoked responses?	Yes. Any time a node influences two or more other nodes there is likely an inflation of the correlation between those downstream nodes. While this is problematic, we focus here on the confounding influence of external stimuli. It will be essential for future research to resolve the more general issue of confounding in FC analysis. Some progress toward removing such confounds have been made, typically via partial correlation or multiple regression (Cole et al., 2016a; Ramsey et al., 2010; Smith, 2012). Note that partial correlation and multiple regression FC approaches are likely only minimally impacted by the task-timing confound, due to partialling out of common signals across time series (mean evoked responses).
7) Is it problematic that removing the mean evoked response leaves some evoked variance?	No. Subtracting the mean evoked response removes the first-order (main) effect of the task stimuli, leaving second-order event-to-event variability in the evoked response that can be used to assess state-dependent changes in the statistical relationship between nodes (i.e., task-state FC). Testing for correlations using such event-to-event variability is very similar to the "beta series" task-state FC approach, which estimates correlations between event-to-event variation in estimated evoked responses (Rissman et al., 2004).

Details of task-state vs. resting-state FC empirical comparisons

The effects reported in **Figure 5B** were relatively consistent across the seven tasks performed by each participant, despite differences in timing, duration, and cognitive processes across the tasks. The percentage of connections with task-state FC increases from resting-state FC (false discovery rate corrected for multiple comparisons, $p<0.05$) for each of the seven tasks without task-regression preprocessing was: 2.0% (emotion task), 4.3% (gambling task), 8.9% (language task), 4.6% (motor task), 7.9% (reasoning task), 14.7% (social task), and 7.7% (working memory task). In contrast, for the FIR approach the rate of task-state FC increases were: 0.8% (emotion task), 2.2% (gambling task), 3.0% (language task), 1.9% (motor task), 3.5% (reasoning task), 2.1% (social task), and 3.1% (working memory task). Thus, there were fewer task-state FC increases for every task when using the FIR approach, demonstrating consistency in this result.

There were also effects of the FIR approach on task-state FC decreases from resting-state FC. Consistent with task-state FC being inflated positively without correction, the FIR approach identified a somewhat larger (but overall similar) number of task-state FC decreases from resting-state FC. This was apparent from the shift from 20.03% of FC decreases without task-regression preprocessing to 21.45% with FIR task-regression preprocessing. Similar results were found when using the constrained basis set approach, though with even more task-state FC decreases (27.39% on average). Focusing on the FIR approach (given its greater flexibility for fitting HRF shape), these results suggest that task timing regression results in a similar number of task-state FC decreases from resting-state FC compared to when no task regression is used.

Validating the FC inflation estimates using inter-subject correlations as a proxy for false positives

Al-Aidroos et al. (2012) performed an analysis involving inter-subject correlations to estimate the effectiveness of FIR task regression in reducing possible false positives in task-state FC analysis. We performed a similar analysis here. The reasoning behind this analysis is that there cannot be neural interaction between the brains of separate subjects, such that any time series correlations across subjects is unlikely to be driven by true neural interactions. Instead, such inter-subject correlations would likely be driven primarily by the task timing that all participants are subjected to. This intuition is compatible with the notion (see Figure 1) that cross-even mean evoked responses can lead to task-state FC false positives by task timing acting as a confounding "third" variable. We repeated the Al-Aidroos et al. (2012) analysis with the HCP N-back task data, observing inter-subject correlations for each of the four primary preprocessing strategies tested here (**Figure S1**). Note that we restricted the time series going into the inter-subject correlations to within-block time points (as with the other analyses), already reducing the effect of task timing on the resulting correlations.

Without task regression there a variety of high inter-subject correlations, primarily in the visual, dorsal attention, and frontoparietal networks. These correlations had a maximum of r=0.50. HRF regression reduced the maximum to r=0.28, basis set regression to r=0.20, and FIR regression to r=0.14. To the extent that inter-subject correlations reflect task-state FC false positives, these results further confirm the conclusion that FIR regression was the most effective of the tested approaches for reducing false positives. However, the non-zero correlations remaining for FIR regression suggests that some subtle effects of evoked responses remains even after removal of event-averaged evoked responses. One prominent possibility is that sequential effects – such as due to practice with the N-back task across blocks – systematically alters the evoked response amplitudes over time. This is consistent with studies showing consistent reductions of evoked activations with practice in a variety of brain regions (Chein and Schneider, 2005). The consistency of these practice effects across subjects could, in theory, drive inter-subject correlations. It would be important for future research to assess this possibility by investigating practice effects. Further, it

could be informative for future studies to also remove event-to-event variance, as was performed by Al-Aidroos et al. (2012) in another supplementary analysis.

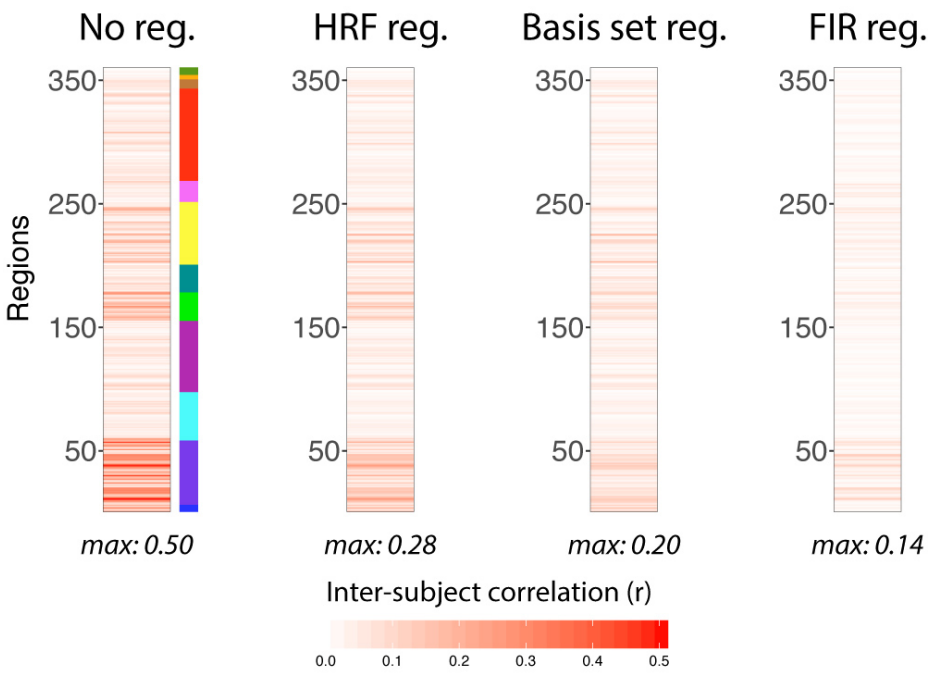

Figure S1 – Inter-subject correlation reduction as a function of preprocessing strategy. Results of inter-subject correlation of N-back task time series are shown for: No regression, canonical HRF regression, basis set regression (5 regressors), and FIR regression. The color bar next to the no regression results reflect the network assignments for each of the 360 regions shown in Figure 5B. As with the FC analyses, the time series were isolated to within-block time points (i.e., the inter-block rest periods were removed).

Method: Non-parametric shuffling procedure to correct for task-timing-driven FC false positives

We utilized a non-parametric shuffling procedure as an alternative approach to correct for task-timing-driven FC false positives (Averbeck et al., 2006; Grün, 2009) (**Figure S2**). This involved identifying time points for each block of the working memory HCP task, then swapping the two blocks for each of the eight task conditions (0-back and 2-back for separate blocks of face, place, tool, and body part stimuli). This swapping was done for all regions' time series except for the "seed" region when constructing the shuffled FC matrix, ensuring a pairwise mismatch between block identities when computing the pairwise correlations. The resulting false positive FC estimates were then converted to Fisher's z-transformed correlation values prior to averaging across subjects (then converted back to r-values for reporting).

Validating the FC inflation estimates using a non-parametric event-shuffling procedure

Several methods have been developed to remove evoked activity in the context of spike train correlations from invasive neural recordings (Brody, 1999; Grün, 2009). We reasoned that these non-parametric methods could be applied to fMRI data to further validate our false positive estimates, which are based on a parametric FIR GLM. The non-parametric methods, which involve fewer assumptions than parametric GLM approaches, involve shuffling events for each neural time series relative to each other (Grün, 2009) (**Figure S2A**). This breaks the moment-to-moment (and event-to-event) relationship between the time series while keeping the time-averaged evoked relationship. The time-averaged evoked relationship is maintained because the shuffling is between activity events that share the same task timing (e.g., trials with the same onset times between shuffled blocks). For simplicity, we shuffled the events by shifting each region's time series in time by one block (of the same condition) relative to each other region. Any remaining correlations between the regions' time series could not be driven by moment-to-moment (or event-to-event) covariance, but rather from task-timing-locked evoked activity that is common across task blocks. We consider such task-timing-driven correlations to be false positives (based on the computational model results).

We applied this shuffling procedure to the empirical fMRI data. We focused on the working memory task because the block durations were more uniform than most of the other tasks in the dataset (facilitating block exchangeability during shuffling), and because the working memory task included the most data per subject. See Methods for details regarding this analysis. We found that the overall pattern of false positive estimates was quite similar ($r=0.83$, $p<0.00001$) between the shuffle-based and FIR-based false positive estimates (**Figure S2B**). The percentage of significant false positives (based on t-tests comparing uncorrected vs. corrected FC, $p<0.05$ FDR) was also very similar to the FIR results: 38% (37% with FIR). Further, the false-positive-corrected task-state FC matrices were highly similar across both methods ($r=0.97$, $p<0.00001$). This convergence with a highly distinct approach provides additional validation of the FIR-based false positive reduction approach.

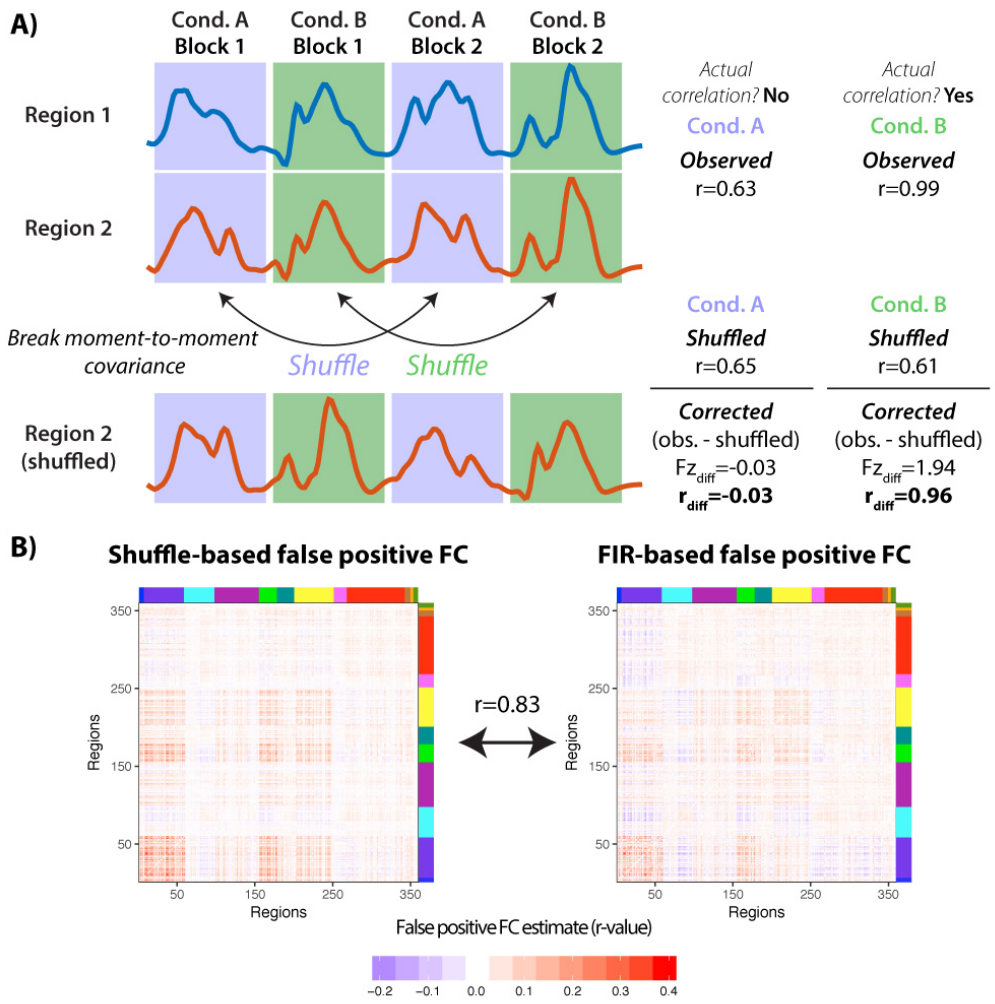

Figure S2 – A non-parametric shuffling procedure to estimate task-state FC false positives. A) The procedure used to estimate task-state FC false positives (and correct for those false positives) is illustrated. This approach is often used to correct for task-timing-induced false positives with spike correlations in multi-unit neural recordings (Brody, 1999; Grün, 2009). There was no inter-region correlation due to the moment-to-moment variance in condition A, but there was a strong correlation in condition B. The amount of FC inflation is estimated by shuffling blocks/events of the same condition in the second region, breaking the moment-to-moment relationship (which is of primary interest) but leaving the time-averaged evoked relationship. Thus, the post-shuffle correlation can be used to correct for task-timing-induced FC inflation. The Fisher's z-transform (Fz) was used to subtract the r-values without the bias introduced by the restricted range of r-values. **B**) The shuffling procedure was applied to the empirical fMRI data, indicating a similar pattern of false positives as the FIR regression approach. Results for the working memory task are shown.

Testing an alternative FC estimation method: PPI

We next tested whether an alternate FC estimation method is similarly affected by fMRI-induced inflation. As we have shown previously, covariance is the common statistical measure underlying a variety of FC measures (Cole et al., 2016b): Pearson correlation, Spearman rank correlation, multiple regression, and PPI are all forms of normalizing/modifying simple covariance. Specifically, Pearson correlation normalizes covariance by dividing by (a transform on) the standard deviations of the time series, while Spearman rank correlation is equivalent to calculating Pearson correlation on the rank orders of the time series values. PPI is the simple pairwise regression between time series along with some nuisance regressors (Cole et al., 2013). Notably, simple pairwise regression (as used by PPI) is equivalent to the covariance divided by the variance of the source (in a source-target pair) time series.

Given that covariance underlies two common task-state FC methods used with fMRI – Pearson correlation and PPI – we expected that PPI would be similarly affected by task co-activations as compared to what we found with Pearson correlations. We tested this by calculating PPI using either no task regression, canonical-HRF regression (as used with standard PPI), or FIR regression. PPI involves a task-regression step that assumes the canonical HRF, such that comparison to the canonical-HRF condition will be the most relevant to existing PPI approaches. Note that we used a version of generalized PPI (McLaren et al., 2012), wherein the "psychological" variables are block-level boxcar regressors (Cole et al., 2013). This aids with interpretation (and comparison to the Pearson correlation results), since the interaction term in the PPI calculation is not influenced by the chosen HRF shape. Also note that, unlike the original PPI approach (Friston et al., 1997), generalized PPI is calculated for each task condition separately (rather than using condition contrasts only) with contrasts calculated as subtraction of PPI estimates (McLaren et al., 2012). Another difference from typical PPI approaches was that the task activation regression occurred prior to (rather than simultaneous with) FC estimation. We did this primarily to make the PPI approach (slightly) more conservative, with as much variance as could be accounted for by the task regressors being taken out prior to PPI estimation. Thus, if anything, the approach used here should reduce the chance of false positives relative to typical PPI approaches.

We began by comparing no-regression to FIR-regression with PPI (i.e., with regression rather than Pearson correlation). As with Pearson correlation, we found that all seven tasks involved statistically significant ($p<0.05$, FDR corrected) changes in FC estimates. The percentage of connections with significant ($p<0.05$, FDR corrected) differences for each task were, respectively (increases/decreases): 8.6/6.3, 28.6/7.8, 25.0/22.9, 24.7/4.1, 35.0/12.1, 43.1/15.3, 23.4/8.2. These results demonstrate that task-timing regression matters for PPI analyses, as it significantly alters PPI estimates across a broad variety of brain regions across a broad variety of task manipulations.

We next tested the extent to which PPI results – which exclusively assume the standard HRF shape – likely include task-evoked activation-based FC inflation. This was quantified by comparing PPI calculated using canonical-HRF task regression versus PPI calculated using FIR task regression. Consistent with the Pearson correlation results, canonical-HRF regression resulted in significantly distinct PPI

estimates relative to when FIR regression was used. The percentage of connections with significant ($p<0.05$, FDR corrected) differences for each task are reported in **Table S2.** Notably, the percentage of changed connections tended to be smaller here than the no-regression vs. FIR regression case. This suggests that the canonical-HRF regression typically used with PPI likely helps reduce activation-induced FC inflation. However, given that a large number of significant differences remained when comparing canonical-HRF with FIR regression, the typical PPI approach appears to not be as effective as FIR regression.

Table S2 – Amount of likely activation-induced PPI inflation. Comparison of canonical-HRF vs. FIR regression approaches with PPI estimates, listed for each of the empirical fMRI tasks.

Task name	% connections increased with canonical-HRF PPI	% connections decreased with canonical-HRF PPI
Emotion	3.3%	4.1%
Gambling	25.7%	1.3%
Language	30.8%	9.9%
Motor	6.4%	0.4%
Social	29.8%	0.8%
Reasoning	36.8%	10.6%
Working memory	23.2%	1.1%

Empirical fMRI data: When mean task-evoked variance is not removed, task-state FC is not primarily driven by mean task-evoked variance

A common concern with the task timing regression approach is that it might remove the very cause of task-state FC that is of interest. The neural mass model already suggested that this is not the case, since removing cross-event mean activity did not induce false negatives (**Figure 4F**). Nonetheless, there might be some concern that the task timing regression approach removes the primary source of task-state FC effects in empirical fMRI data. We tested this possibility by comparing the amount of task-state-FC-driving variance removed by task timing regression. We expected that most of the task-state-FC-driving variance would remain after this preprocessing step, consistent with the primary driver of task-state FC being moment-to-moment (rather than cross-event mean) fluctuations. Critically, however, removing this cross-mean variance would still be important, since the relatively small amount of cross-event mean variance was shown in previous sections to cause (false positive) statistically significant effects.

We tested this hypothesis by quantifying the change in between-region shared variance before versus after FIR task regression. We found that 89.39% of the shared variance across all pairwise connections (across all 7 tasks) was preserved after FIR regression. This was computed after converting the r-values representing task-state FC to r-squared values (i.e., percent shared linear variance), then averaging across subjects, tasks, and connections. This revealed that mean shared variance during task went from $r^2=0.066$ without task regression to $r^2=0.059$ with FIR task regression on average. This small change indicated that 89.39% of the shared linear variance was

preserved after FIR regression on average. A similar result was obtained when restricting to only connections that were statistically significant ($p<0.05$, FDR corrected) relative to 0 prior to task regression: 89.53% of the shared linear variance was preserved after FIR regression on average. This result confirms our hypothesis that, while critical for reducing the chance of a false positive for any single result, the FIR regression step removed only a small amount (less than 11%) of the variance driving task-state FC effects. This, in turn, demonstrates that task-state FC estimates (even when not performing task regression) are primarily driven by moment-to-moment (and event-to-event) variance rather than the cross-event mean variance removed by FIR regression.

Supplementary References

Al-Aidroos, N., Said, C.P., Turk-Browne, N.B., 2012. Top-down attention switches coupling between low-level and high-level areas of human visual cortex. Proceedings of the National Academy of Sciences 109, 14675–14680. doi:10.1073/pnas.1202095109

Averbeck, B.B., Latham, P.E., Pouget, A., 2006. Neural correlations, population coding and computation. Nat Rev Neurosci 7, 358–366. doi:10.1038/nrn1888

Brody, C., 1999. Correlations without synchrony. Neural Comput 11, 1537–1551.

Chein, J., Schneider, W., 2005. Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. Brain Res Cogn Brain Res 25, 607–623.

Cole, M.W., Ito, T., Bassett, D.S., Schultz, D.H., 2016a. Activity flow over resting-state networks shapes cognitive task activations. Nat Neurosci 19, 1718–1726. doi:10.1038/nn.4406

Cole, M.W., Reynolds, J.R., Power, J.D., Repovs, G., Anticevic, A., Braver, T.S., 2013. Multi-task connectivity reveals flexible hubs for adaptive task control. Nat Neurosci 16, 1348–1355. doi:10.1038/nn.3470

Cole, M.W., Yang, G.J., Murray, J.D., Repovs, G., Anticevic, A., 2016b. Functional connectivity change as shared signal dynamics. J Neurosci Methods 259, 22–39. doi:10.1016/j.jneumeth.2015.11.011

Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6, 218–229. doi:10.1006/nimg.1997.0291

Grün, S., 2009. Data-Driven Significance Estimation for Precise Spike Correlation. J Neurophysiol 101, 1126–1140. doi:10.1152/jn.00093.2008

McLaren, D.G., Ries, M.L., Xu, G., Johnson, S.C., 2012. A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. NeuroImage 61, 1277–1286. doi:10.1016/j.neuroimage.2012.03.068

Miezin, F., Maccotta, L., Ollinger, J., Petersen, S., Buckner, R., 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. NeuroImage 11, 735–759.

Pearl, J., 2009. Causal inference in statistics: An overview. Statist. Surv. 3, 96–146. doi:10.1214/09-SS057
Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference from fMRI. NeuroImage 49, 1545–1558. doi:10.1016/j.neuroimage.2009.08.065
Rissman, J., Gazzaley, A., D'Esposito, M., 2004. Measuring functional connectivity during distinct stages of a cognitive task. NeuroImage 23, 752–763. doi:10.1016/j.neuroimage.2004.06.035
Smith, S.M., 2012. The future of FMRI connectivity. NeuroImage 62, 1257–1266. doi:10.1016/j.neuroimage.2012.01.022
Visscher, K., Miezin, F., Kelly, J., Buckner, R., Donaldson, D., McAvoy, M., Bhalodia, V., Petersen, S., 2003. Mixed blocked/event-related designs separate transient and sustained activity in fMRI. NeuroImage 19, 1694–1708.