Why deconvolve Ca^{2+}?

A practical guide to recovering neural activity from cellular fluorescence imaging

Mathew H. Evans

April 27, 2018

Abstract

Fluorescence imaging of somatic calcium (Ca^{2+} ~~for short~~) or voltage is an increasingly popular technique for recording the activity of large groups of neurons (cite reviews). Given that Ca^{2+} is an indirect measurement of neural activity, and signal quality considerations (background noise, slow kinetics, and heterogeneity in expression across cells) the fluorescence time-series is typically *deconvolved* prior to tuning or correlation analysis. Some methods go further, inferring the timing and number of spikes (*spike inference*). Recent efforts have seen an explosion in the number of available methods for Ca^{2+} deconvolution and spike inference, with impressive results (Theis, spikefinder). Here we reconsider this progress showing (i) a popular metric - Pearson correlation coefficient (PCC) - is poorly conditioned (results change a great deal with small changes in parameters), and results in poor estimates of firing rates (see also Ganmor, Reynolds CosMIC), a problem that remains when inferring spikes in data collected at large-scale-recording temporal resolution (with sampling rates of 2-10Hz). (ii) When applying deconvolution or spike inference to large scale measurements from somatosensory cortex (Peron et al 2015) performance of state-of-the-art methods is inconsistent across methods and poorer than may be expected from published results or ground truth data. We conclude that to-date there are no 'magic bullet' approaches for inferring accurate estimates of neural activity from fluorescence imaging data. We suggest ~~great care must be taken in the application of spike inference/deconvolution methods, and~~ that conclusions of analyses that depend on the deconvolution or spike inference process must be verified across multiple methods or analysis parameters.

1 Introduction

why deconvolve
- improve signal/noise ratio i.e. better estimate what the cell is doing
- increase temporal resolution e.g. to recover tuning and correlations
- normalisation of expression across neurons
- get rid of known artefacts

how to deconvolve
- many methods giving different results (due to analysis/design choices)
- comparing methods involves metrics
- metrics are important as they give different results

the cautionary tale
- if the signal isn't perfect our metrics and methods lead to biased/poor estimates of simple neural properties (FR, tuning, correlations)
- this bias may be worse (more wrong) than the benefits we're supposed to get when deconvolving e.g. miss-classify cells as tuned. Poor estimates of data dimensionality.
- artefacts may not be removed after all

2 Results

Fitting spike inference methods using Pearson correlation coefficient leads to poor estimates of firing rate

The point of this section is that if you are going to do deconvolution, PCC is a biased metric so use something else e.g. ER. This isn't a particularly novel point any more - see Reynolds biorxiv 2017 + Ganmor

Results in brief
- Using PCC as a metric leads to poor estimation of firing rate, both under and over estimation, out by a factor of 1 (double the true firing rate)
- PCC does not vary smoothly with model parameters, potentially leading to poor (over) fitting to training data N.B. we didn't test this - all data was used for fit + evaluation, not separate training/test sets.
- ER is a better choice (but see also CosMIC, Reynolds et al 2017 or Information based methods, Theis)
- This result is true regardless of method used (see repeated analysis on MLSpike + LZero
- This result is true on both high and low frame rate data

To assess the performance of spike inference methods at estimating spike trains from Ca^{2+} signals, different methods can be tested on *ground truth* datasets - where the spiking activity of a cell is recorded simultaneously with Ca^{2+} imaging, ideally using high-signal-to-

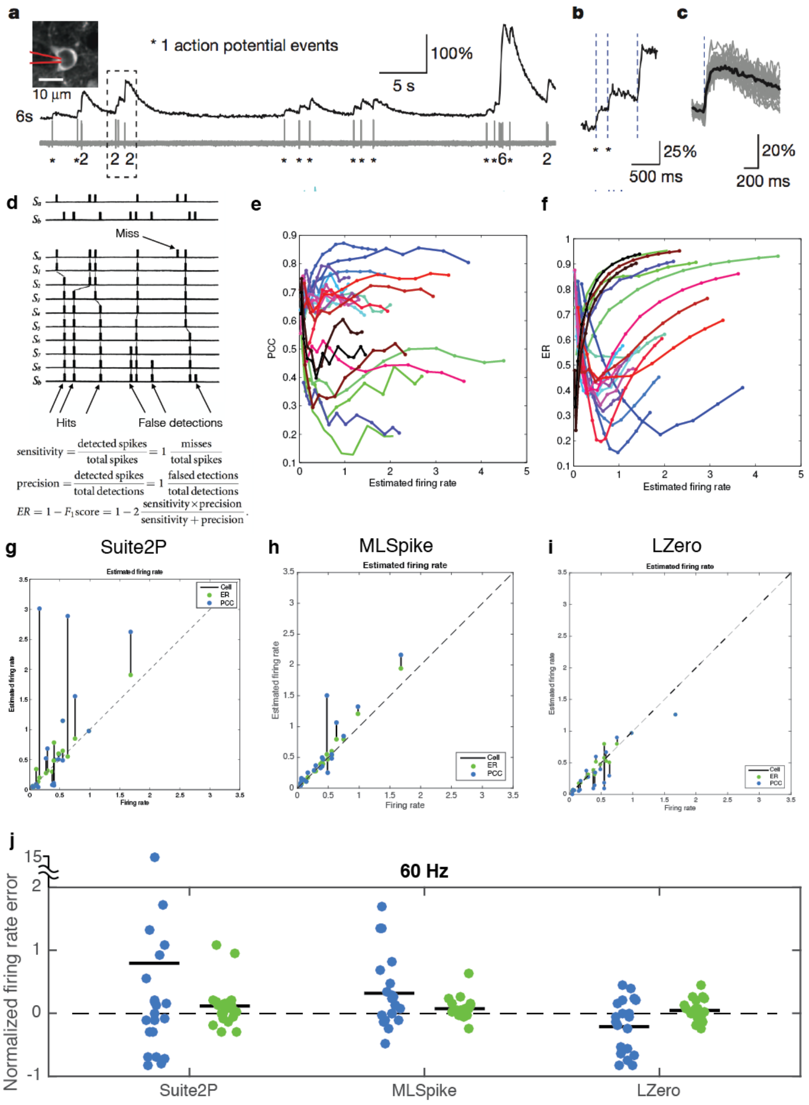

Figure 1: Ground truth data analysis. (a) Ground truth data collection example from Chen et al. Nature 2013. Top left, an example field of view from imaging data. Red lines denote the outline of a juxtacellular recording pipette. Time series shows measured calcium fluorescence (top) and simultaneously recorded voltage (below). Spikes are marked with asterisks. (b) Single spikes influence the calcium trace. (c) raw data (grey) and average (black) of single spike induced changes in fluorescence. (d) top: Victor and Purpura (1996) proposed a spike metric to compare spike trains. This metric is generated by determining the number of elementary operations (shift, addition, or deletion of individual spikes - depicted as rows in here) required to match two spike trains, up to some temporal precision. Bottom: In Deneaux et al 2016 the Error Rate (ER) is similarly computed as a normalised ratio of sensitivity vs precision in spike detection. Detections are counted to within 0.5s. (e) PCC as a function of estimated firing rate (using Suite2P, Pachitariu et al 2017). Colours are different cells.(f) as in (e) but with ER as a metric. (g) Estimated firing rate for 'best' deconvolution parameters versus real firing rate. Best parameters are taken as the highest or lowest points in (e) and (f), respectively. (h) as in (g) but using MLSpike (Deneaux et al 2016). (i) as in (g) but using LZero (Jewell and Witten 2017). (j) Normalised firing rate error (estimated FR - true FR / true FR) for all cells and across methods. Lines are means. (a) reproduced from Chen et al. 2013.

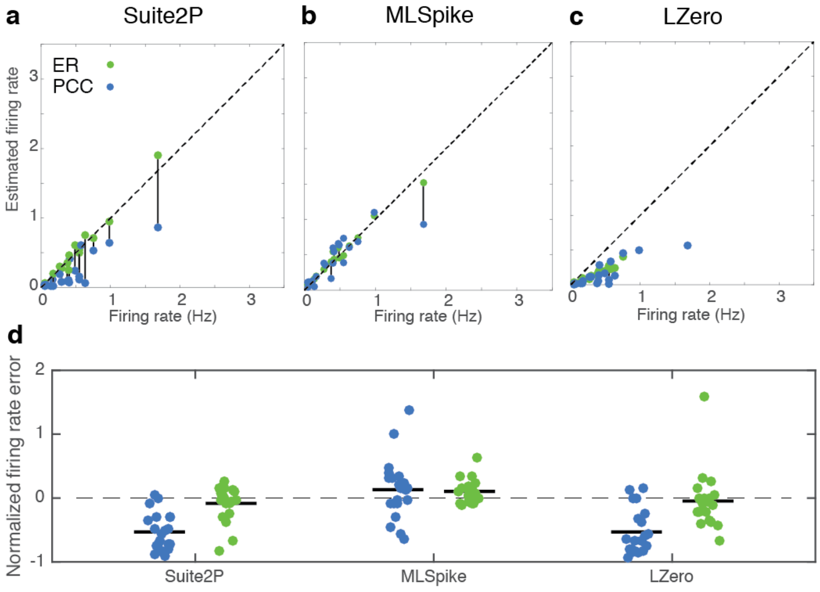

Figure 2: Downsampled ground truth analysis. (a-c) Estimated firing rate for 'best' deconvolution parameters versus real firing rate using Suite2P, MLSpike and LZero applied to ground truth data downsampled to 7Hz to more closely match population imaging experiments. (d) Normalised firing rate error (estimated FR - true FR / true FR) for all cells and across methods. Lines are means.

noise techniques such as juxtacellular recording (see Figure 1 (a)).

Figure 1 (e) shows the results from inferring spikes from ground truth data with Suite2P (Pachitariu et al) using a range of a threshold parameter which trades off misses vs false detections. In general Pearson correlation coefficient (PCC) increases as estimated firing rate increases. As a consequence, if we choose the model parameters that maximise PCC then the recovered spike-train consistently overestimates the ground truth rate of spiking (Fig 1(e,g)). Over the population PCC-optimized Suite2P overestimates firing rate, in some cases severely (Figure 1 (g,j)). In addition, PCC does not change smoothly with gradual parameter changes, as can be seen in the lines for individual cells in Figure 1 (e), which could lead to overfitting/noisy estimates of the best parameters.

To address the weaknesses of PCC, we implemented the Error Rate (ER) spike distance metric of Deneaux et al 2016, a summary measure based on the distance measure of Victor & Purpura (1996). ER (outlined in Figure 1 (d)) returns a normalised score which is 0 for a perfect match between two spike trains, and 1 when all the spikes are missed. When evaluating the same inferred spike trains from 1(e), ER is best (lowest) for intermediate estimated firing rates, suggesting that estimates closer to the true firing rate are rewarded with good scores (Figure 1 (f)). This intuition is shown to be true when comparing the best estimate of firing rate to the true firing rate for each cell (green dots in Figure 1 (g, j left)). Though ER-optimized Suite2P overestimates firing rate, it does so to a much smaller degree than the PCC-optimized approach (mean error ∼0.1Hz, compare blue and green dots in 1 (g,j)). In addition, unlike for PCC, individual cell results in Figure 1 (f) show ER varies smoothly with parameter changes. ER results in better estimates of firing rate than PCC when it is used to optimise parameters for two other spike inference methods, MLSpike (Deneaux et al 2016, Fig 1 (h, j middle)) and LZero (Jewell & Witten 2017, Fig 1 (i, j right)).

All light microscopy experiments (including two photon imaging) have a 'photon budget' (set by the microscope and sample) which can be deployed by the experimenters to achieve certain goals. Higher signal to noise can be achieved with high frame rates and zoomed in imaging (more pixels per cell). If the goal is to record from large numbers of neurons the overall photon budget must be spread more thinly per neuron, with smaller numbers of pixels per cell and lower frame rates, resulting in lower signal to noise ratios (Peron, Chen & Svoboda COiN 2015). Therefore it is important to ask whether the problems of using PCC as a metric are alleviated or compounded in a low frame rate regime. To assess this we repeated the ground truth analysis, but with imaging data down-sampled to 7Hz and found similar results - ER-optimized methods out performed PCC-optimised methods (Fig 2). Interestingly, PCC-optimised methods often underestimated the firing rate of cells at 7Hz imaging rate (Fig 2 (d)).

Together, these results show that when optimising or comparing spike inference methods the choice of metric is critical. PCC is a poor choice of metric as it returns inconsistent results with small changes in algorithm parameters, and leads to poor estimates of simple measures such as firing rate when used across methods and sampling rates.

2.1 Spike inference and deconvolution methods disagree on estimates of simple neural statistics

Results in brief

- we want to see what effect deconvolution/spike inference method choice has on data analysis

- methods (within and between 'class' i.e. deconvolution vs spike inference) disagree on event rate of cells - some methods are qualitatively wrong, others have quantitative disagreement

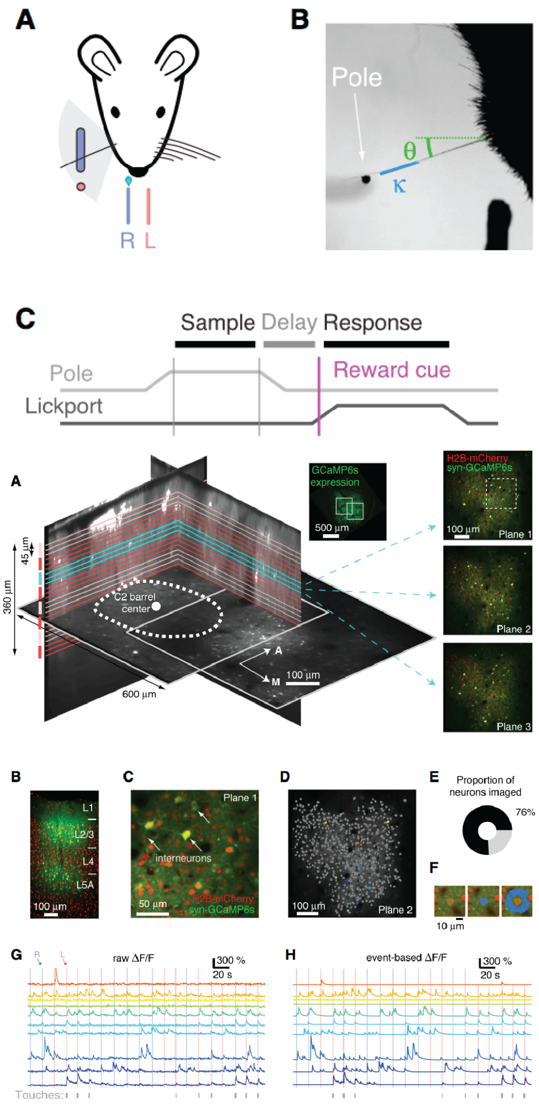

Figure 3: Peron et al. 2015 experimental design. TO DO EXPAND description once final figure arrangement is decided

To determine the effect of deconvolution or spike inference on simple measures of neuron activity we applied nine different approaches (see Methods) for Ca^{2+} deconvolution to data from a single experiment from Peron et al 2015 (see Figure 3).

In Peron et al 2015 two-photon Ca^{2+} imaging was used to record neural activity from up to ∼2000 neurons simultaneously at 7Hz from superficial barrel (Layer 2/3 somatosensory) cortex as mice performed a head-fixed tactile localisation task with their whiskers. For the results presented here 1552 neurons were recorded for a total of just over 56 minutes (23559 time points). This relatively long recording ensured good estimation of the measured parameters (i.e. pairwise correlations are stable see Fig. 12).

The most basic analysis of neural activity is to determine the mean firing rate of each cell in the recording - a quantity that is known to follow a log normal distribution at the population level (Wohrer et al 2012). We determined the mean spike/event rate per cell for all approaches. Figure 4 (a) and (b) show that no two methods return the same distribution of spike/event rates. The deconvolution methods (Yaksi - LZero$_{kernel}$) appear to overestimate the average firing rate of the population as well as the number of cells with high firing rates. Spike inference methods can be tuned to produce qualitatively correct distributions (median near zero, long right skewed tails), they disagree quantitatively. **TO DO: Add back in 'straw man' results based on best ground truth parameters?**.

It has been estimated from cell-attached recordings that ∼13% of somatosensory neocortical cells are silent during the pole localisation task (fewer than one spike every two minutes, O'Connor et al 2010), a quantity increasing to ∼26% in Layer 2/3. For the nine approaches we tested, seven estimated the proportion of silent cells to be below 10%, with wide disagreement between the other two methods (Figure 4 (c)). Even for simple statistics, the choice of deconvolution or spike inference method results in widely different results.

2.2 Different spike inference methods lead to different estimates of task related neurons

Results in brief

- Absolute firing rate may not be important, so what about task related activity?
- Different methods disagree about the number of task-tuned cells
- Different methods disagree on the identity of task-tuned cells
- Cells classed as tuned by one method lacks clear task related activity when processed with another method
- This disagreement means either (a) some tuned cells are missed (b) some un-tuned cells are classed as tuned (c) both.
- Spike inference/ deconvolution leads to fewer neurons classified as task tuned (vs calcium), perhaps due to removal of task-related changes
- Agreement between methods may be an indicator of robust tuning

For many analyses, it is the relative activity of a cell and not its exact firing rate that is important. A common analysis is to ask whether a neuron's activity is task related - does a cell respond more during a specific epoch of the task than would be expected from a random process. Such task tuning may then imply that a given cell or region of the brain is causally involved in the task, and serve as a target for manipulation studies. We quantified the proportion of task related neurons in our dataset following the approach of Peron et al 2015. Calcium/instantaneous firing rate/ events for each cell is shuffled in time before a trial-averaged PSTH is generated, and the largest peak in the PSTH is recorded. This is repeated 10,000 times.

A distribution of shuffled PSTH peak magnitudes is generated, and if the peak of the true (data) PSTH is larger than the 95%ile of the shuffled distribution, in either the Left or Right (Go, No Go) trials, that cell is considered 'tuned'. Firstly, each method estimates a different proportion of tuned vs untuned cells, both in comparison to estimates from the raw Ca^{2+}, and in comparison to one another. Secondly, the methods only agree on the tuned status of individual neurons for ∼50 cells (from a range of 50 - 250 tuned cells TO DO GET ACCURATE NUMBER).

Figure 6 shows the trial-averaged activity for cells classified as tuned when looking at raw Calcium data only (Fig 6 (a)), or where multiple methods agree that the cells are tuned (Fig 6 (b-d)). It is unclear whether peaks in Calcium activity in (a) are artifactual or real, as they are often eliminated in the other methods. However, in cells classified as task-related by 6 methods (Fig 6 (c)) clear peaks of activity can be seen across methods, with those peaks lasting for multiple time frames. Comparison across methods could prove a powerful approach increase the reliability of fluorescence imaging analysis.

2.3 Deconvolution and spike inference may degrade the ability to recover precisely timed responses

Results in brief

- Experimental manipulations or task events may result in precisely timed (+ brief, small amplitude) responses in L2/3 of cortex
- Deconvolution/spike inference is often employed to increase the ability to detect such responses/cells, by reducing background noise and removing the slow kinetics of Calcium changes.
- For clearly tuned cells, this approach appears valid (example PSTH shows temporally sharp peak)
- However, as for task tuning, different methods disagree on the number of tuned cells.

To determine whether deconvolution improves the temporal precision of analyses such as tuning curves we computed the touch-triggered average for all cells in the example dataset. In the somatosensory system, touch onset is a salient sensory signal known to drive a subset of neurons with short latency and low jitter (O'Connor et al 2010, Hires et al 2014). To determine touch tuning for each cell we found the peak in the touch triggered

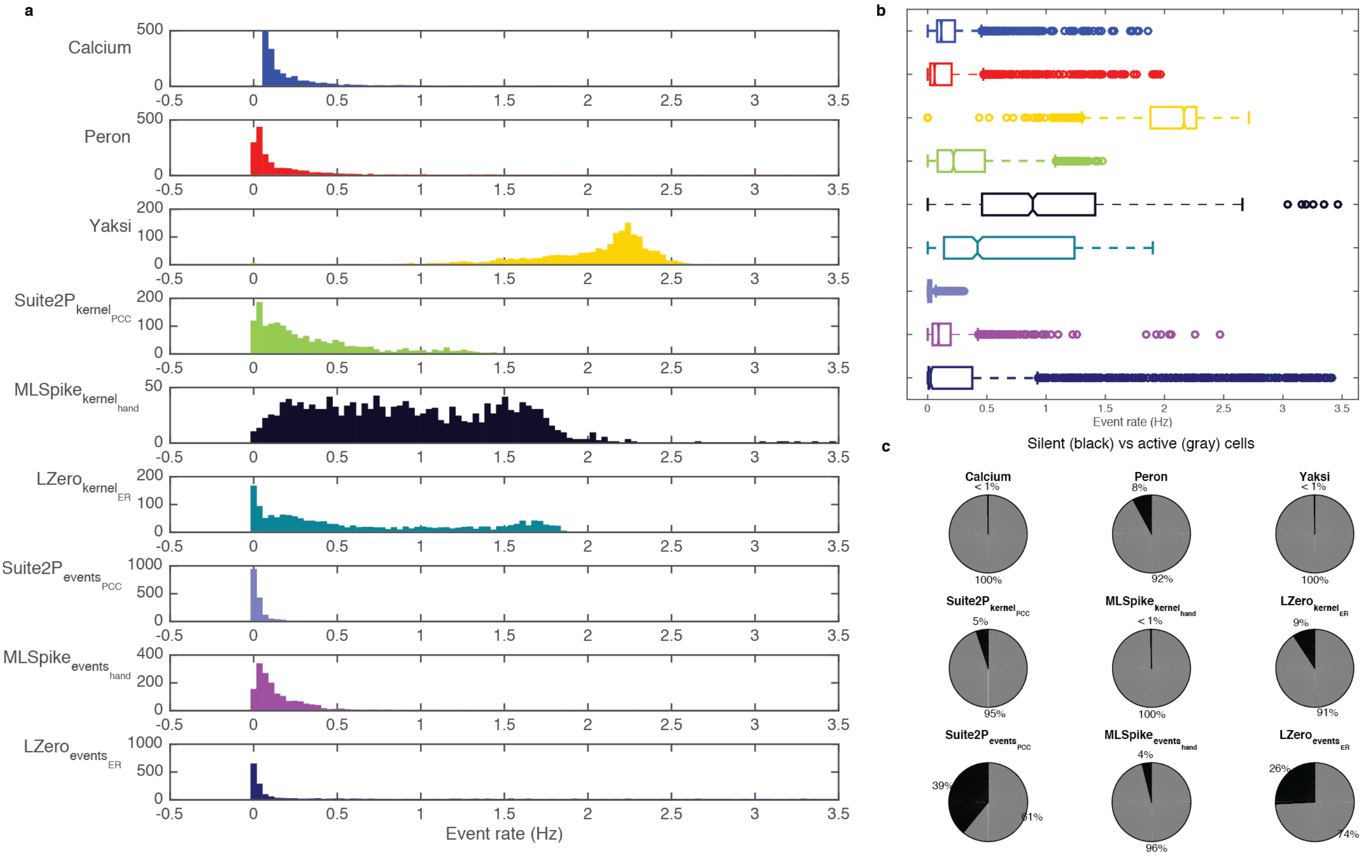

Figure 4: Estimated ‘event rate’ for all cells in an example session. For the first 6 methods (Calcium - LZero$_{kernel}$), events are detected as fluorescence transients greater in magnitude than 3 std deviations of background noise. Background noise = data - smoothed version of data, to eliminate slow transients. Methods 7-9 (Suite2P$_{events}$ - LZero$_{events}$) return a spike count per time bin. (a) Histograms of event rate per cell for each method. (b) Same data as in (a) but plotted as box and whisker plots. Notch = median, box limits = 25th and 75th percentile, whiskers = extent of data up to 1.5 IQR (c) Proportion of active (gray) vs silent (black) cells for each method. Silent = event rate < 0.0083Hz.

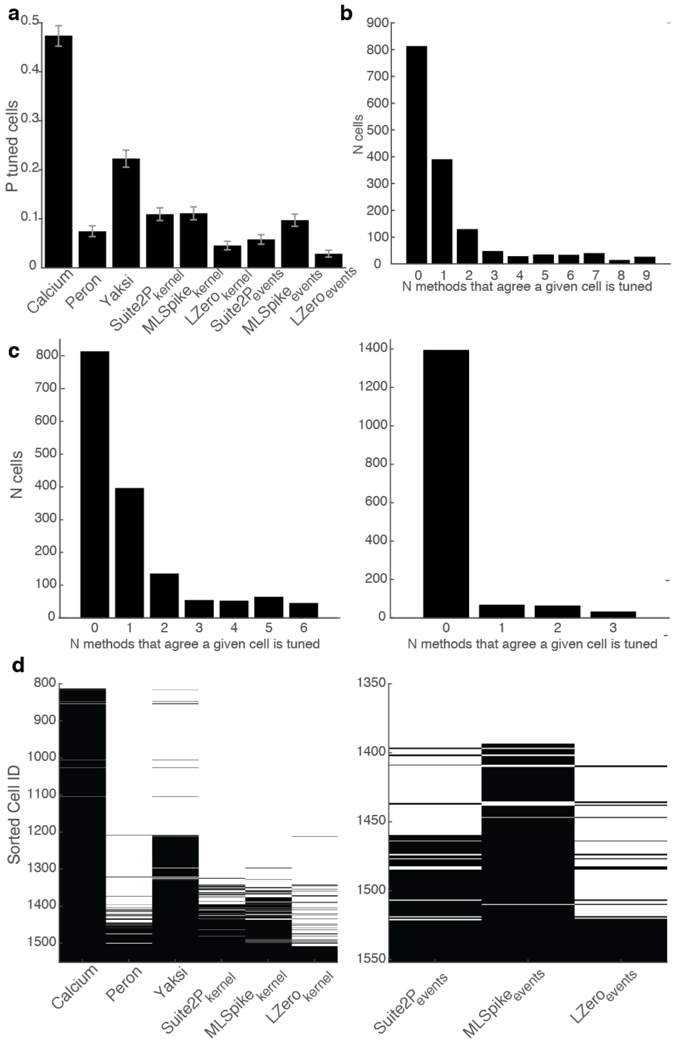

Figure 5: Tuned cells. Tuned cells were determined through shuffle tests (following Peron et al 2015, see Methods). (a) Number of tuned cells per deconvolution method. Error bars are 95% binomial confidence intervals (Jeffreys interval) (b) Agreement between methods. Bars show total number of cells classified as tuned by N methods. (c) Agreement between continuous signal methods (left) and spike inference methods (right). (d) Array of tuned cell identities, separately for continuous signal methods (left) and spike inference methods (right). Black = tuned, white = not tuned. Rows are cells, ordered by the number of methods that classify that cell as tuned (agreement, as plotted in c).

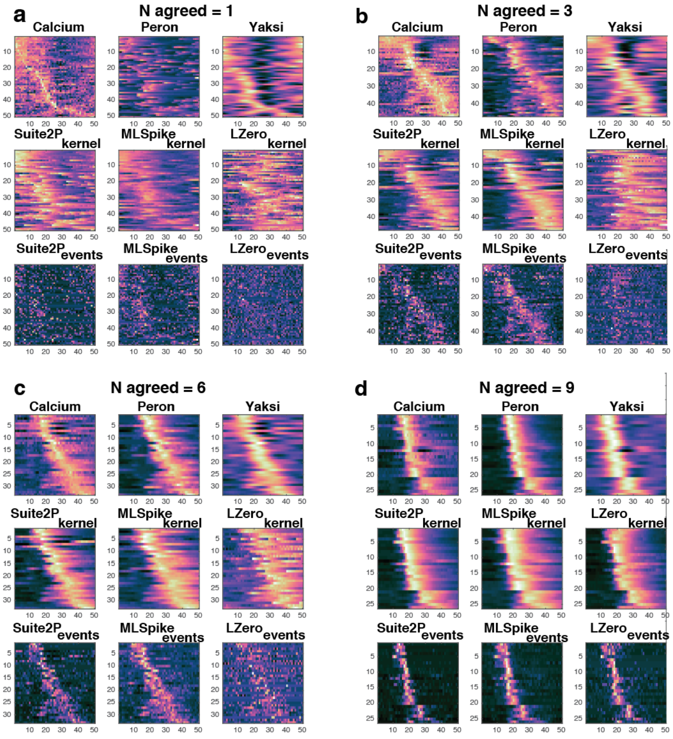

Figure 6: Degree of agreement between methods can identify strongly tuned cells. (a) Example normalised (z-score) trial-average histograms for 50 cells (rows) classified as tuned in an analysis of raw Calcium data. Each subsequent panel shows trial-average histograms for the same cells, but following processing by each of the eight deconvolution/spike inference methods. (b) - (d) as in (a) but showing trial-average data for cells classed as tuned by 3, 6 and all 9 methods. TO DO: add axis labels. Add marker to show which methods classify which cells as tuned?

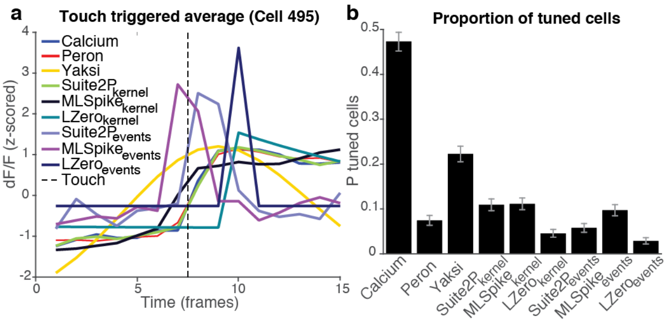

Figure 7: Touch-related responses. (a) Comparing touch-triggered average (mean deconvolved FR per imaging frame) from different deconvolution methods for one example cell. Touch occurs during frame 7 (dotted line). (b) Number of touch tuned cells varies across methods. A cell is classed as touch tuned if peak touch-triggered activity is significantly greater than shuffled data (Mann-Whitney U test, Benjamini Hochberg corrected). Error bars are Jeffreys intervals

average, and compared the data distribution (one data point per touch) at this time point to a shuffled data distribution (Benjamini Hochberg corrected Mann-Whitney U test). As for task tuning, data processed with different methods result in different estimates of the number of touch-tuned neurons.

TO DO? Comparison of estimates when deconvolving/inferring spikes vs not

- signal to noise

- temporal resolution (rise/decay time)

2.4 Deconvolution and spike inference results in different estimates of the dimensionality of population recordings

Results in brief

- Dimensionality reduction techniques such as PCA allow researchers to make sense of large scale neuroscience data

- Often performed as a pre-processing stage ahead of visualisation or clustering, PCA provides an estimate of the dimensionality of data - the number of orthogonally separable sources of variance in the data

- Dimensionality estimates are affected by deconvolution/spike inference choices

- Depending on the approach, the same dataset can appear low or high dimensional

- TO DO: add back in (in supplement) results when deconvolutino/spike inference parameters are different?

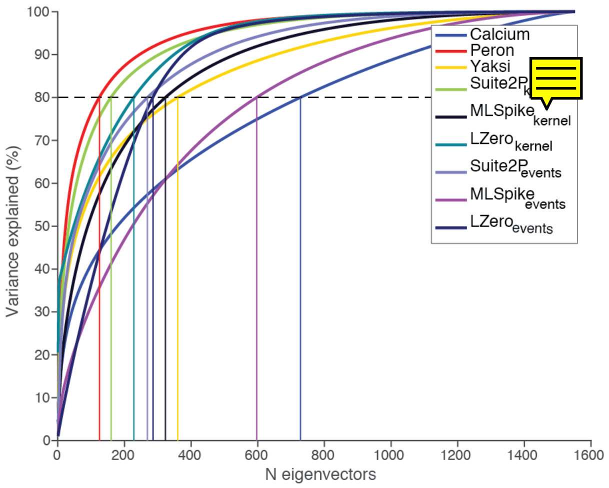

Figure 8: Cumulative variance explained by N eigenvectors following Principal Components Analysis. Different spike inference/deconvolution methods result in different estimates of the dimensionality of the data. For example, 80% of the variance can be explained by 120 or 720 eigenvectors (orthogonal dimensions) depending on the processing method used.

2.5 Pairwise correlation distributions are affected by spike inference and deconvolution

Results in brief

- Many population analyses involve interpretation of pairwise correlations between cells

- Deconvolution is designed to remove noise, while spike inference will sharpen temporal responses, which will lead to more accurate estimates of pairwise correlation

- In our analysis, different methods result in different estimates of pairwise correlation.

- Correlations are not just scaled - some pairs that appear correlated from processing by one method are uncorrelated when processed with another method

- Some differences in pairwise correlation distributions may be caused by introduction of noise (more symmetric distributions) or smoothing (broader distributions).

- Specifics:

- Some methods agree with each other (Suite2P$_{kernel}$,MLSpike$_{kernel}$), some actively decorrelate (Peron, Spike inference methods)

- Interpretation:

- Deconvolution/spike inference is always a trade-off between false positives and misses - meaning you get both - resulting in altered pairwise correlations, and their distributions

- Deconvolution, by eliminating photonics shot noise (smoothing the time series), increases the temporal correlations in the data, leading to stronger correlations

- Choosing analysis parameters that result in long tailed firing rate distributions inevitably lead to more miss errors, and therefore actively decorrelate.

- Miss real spikes + overestimate background rates (see also Ganmor), therefore correlation estimates are noisier (due to false positives) or biased (due to

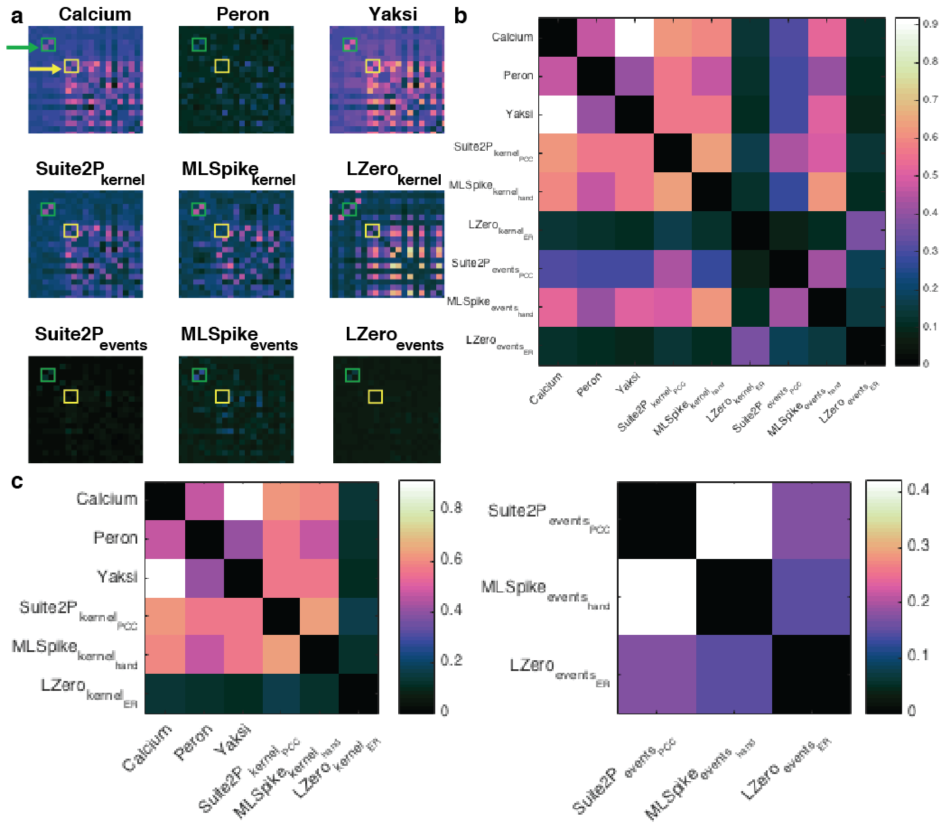

Figure 9: Pairwise correlations. (a) Example pairwise correlations for 50 cells. Some pairs of cells appear correlated when the data is processed by different methods (green arrow and boxes). Other pairs appear correlated when processed with one method but not with others (yellow arrow and boxes). (b) Correlation between pairwise correlation matrices for each method. Some methods result in similar correlation matrices (e.g. Yaksi and Calcium), while others generate distinct correlation matrices (LZero methods). (c) as in (b) but split to show continuous methods (left) or spike inference methods (right).

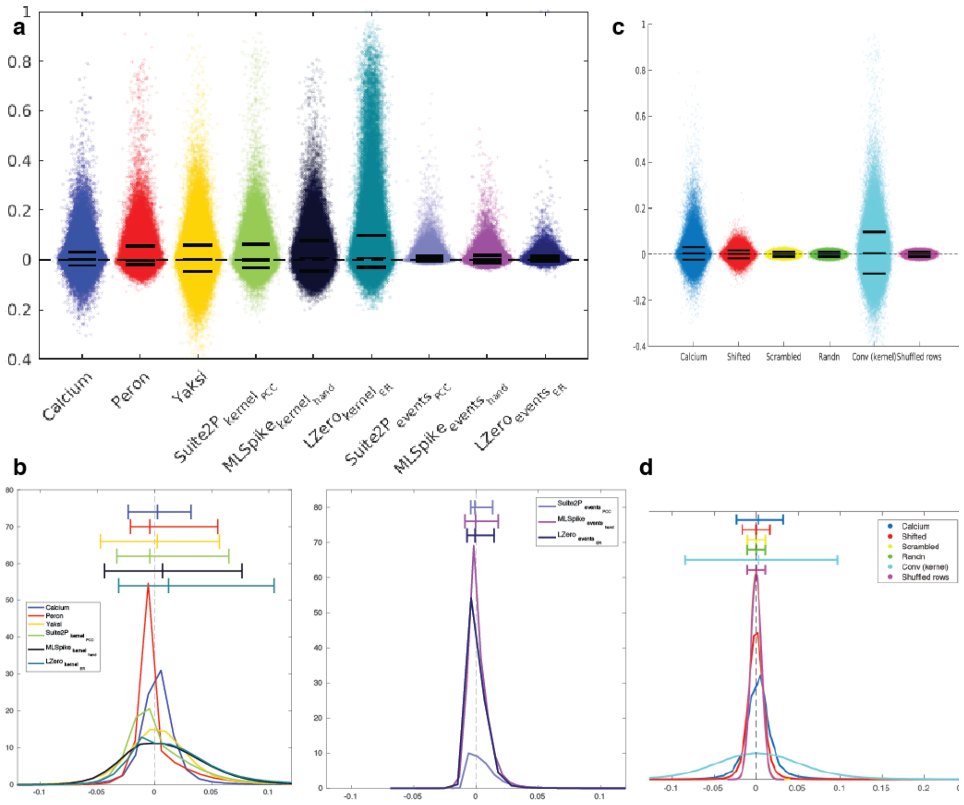

Figure 10: Pairwise correlation distributions. (a) Pairwise correlations between all cells (y-axis) following processing with all deconvolution/spike inference methods (x-axis jitter added for clarity). (b) Pairwise correlation kernel density functions for all methods. (c) and (d) as in (a) and (b) for different randomised versions of the data, to aid interpretation of distribution changes in (a) and (b)

misses/decorrelation)
- MLSpike$_{kernel}$ and LZero$_{kernel}$ have broader distributions more similar to that resulting from smoothing the raw Ca^{2+}. The Peron events have a large number of PCCs below zero, suggesting that choosing parameters that penalise false-positives may be actively decorrelating the data.

A goal of many Ca^{2+} imaging experiments is to record from populations of neurons, and then perform clustering or dimensionality reduction. These analyses rely on estimates of pairwise correlations. Figure 10 (a) and (b) show the distributions of pairwise correlation coefficients computed separately for each method. To aid interpretation of these results we also computed pairwise correlations for five different data surrogates:

- Shifted: the fluorescence time series for each cell was randomly shifted in time (using Matlab's circshift function) by up to 10000 frames. LOGIC: to preserve each cell's autocorrelation

- Scrambled - elements of the original N x T data matrix were sampled randomly (without replacement) to generate a new data matrix. LOGIC: keep true data distribution but randomize everything else

- Randn - pseudorandom values drawn from a normal distribution. LOGIC: Totally random. N.B. Key here is the scrambled data is identical, as PCC doesn't care about the data distribution per se, only the covariability in the data i.e. they both go up, regardless of whether it's a twofold or a tenfold increase

- Conv (kernel) - original data convolved with an exponentially decaying kernel as is used in the MLSpike and Suite2P deconvolution methods. LOGIC: to show the effect that smoothing has on the correlation distribution

- Shuffled rows - like the 'scrambled' data, but shuffling was done separately for each cell (rows of the data matrix). LOGIC: to preserve differences in event rate across neurons. Again, this shouldn't be any different to the scrambled data as PCC is invariant to affine transformations i.e. same tuning but larger changes in firing rate.

MLSpike and LZero have broader distributions more similar to that resulting from smoothing the raw Ca^{2+}. Suite2P and Peron events have a large number of PCCs below zero, suggesting that choosing parameters that penalise false-positives is actively decorrelating the data.

Yaksi's distribution is symmetric (like all the noise surrogates) suggesting dirt has been added to the data.

LZero resulted in very sparse time series, so the long tail of positive values are likely to be the large group of almost silent cells.

Spike inference methods all have sharp peaks just below zero. Not sure why

Spike inference does not automatically remove experimental artefacts

Apart from improving estimates of neural activity, spike inference is also used to remove artefacts from the data such as slow drifts in fluorescence across experiments, or differences in single-spike fluorescence across cells. In head fixed imaging experiments, licking behaviours can lead to a dip in Ca^{2+} fluorescence (Simon Peron, personal communication. Pachitariu COSYNE poster on correcting drift FIGSHARE LINK TO POSTER). However, Figure 11 shows that this dip is also present in deconvolved traces (TO DO ADD BETTER FIG).

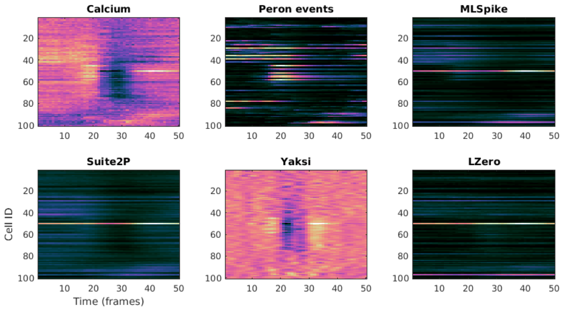

Figure 11: PLACEHOLDER FIGURE. Lick induced dip in Ca^{2+} fluorescence seen in the raw Calcium is also seen in deconvolved data.

3 Discussion

MDH NOTES: Add a Discussion to collect notes on what the conclusions and recommendations are e.g. (1) Don't use PCC; use ER or something similar (full ROC)

(1b) Deconvolution methods trade-off FNs vs FPs (hence need to use metric that captures both)

PCC is invariant to affine transformations of the data (noted also by Theis et al). Specifically, PCC will not change between two cells if the firing rate is doubled or halved. Therefore neither false positives nor false negatives are penalised per se, and spike inference results that maximise PCC between real and inferred spikes cannot be interpreted in terms of spike rate. If the goal of an analysis is to estimate the true firing rate or spike timing of the cell, PCC is not an appropriate metric to use in spike inference optimisation. Instead, a metric such as ER - which explicitly penalises both FPs and FNs, giving better scores to inferred spike trains that are closer to the true spike train in terms of spike count and timing - are a better choice.

(2) Choice of deconvolution method will change inferences taken from all analyses that follow. So use either (a) raw Ca2+ and deconvolution/spike inference OR (b) two different deconvolution/spike inference methods. [NB this links with ideas of robust inference: that obtaining the same result in the face of wide variation increases its reliability]

Point to Figure 6

*(3) Message is *not* abandon deconvolution; message is: get it solved. We need these problems solved: when*

we move to very high frame rate imaging and faster Ca2 sensors, then we will want to look at neural coding at spike resolution. So we will need deconvolution to be properly reliable...

Many questions do not require spike timing (see short discussion in Harris et al 2016 NN Review 'Improving data quality in neuronal population recordings' - *When neurons fire sparsely, for example, neuronal responses can be characterized by how the calcium response itself depends on stimulus or behavioral-related factors. The results of such analyses will not be numerically identical to analyses computed from actual counts (for example when computing correlations among neurons), but if interpreted correctly, this can avoid biases introduced by explicit spike estimation.*).

(4) Deconvolution and spike inference, and the parameters of the methods used, will affect the signal in predictable ways e.g. more/less FPs/FNs depending on whether you are trying to explain every wrinkle in the Calcium trace vs match empirical firing rate distributions. Correlation distributions will be broader if you've smoothed the signal/ removed noise. So build this understanding into your interpretation. For example, the dimensionality of the data depends on where you set your spike detection threshold (sparse vs fuller signal), so conclusions about dimensionality need to reflect this.

4 Methods

Spike train metrics

Pearson correlation coefficient - down sampled or gaussian convolved. Deneux implementation of normalised error rate, derived from Victor and Purpura 1996 Error Rate.

List of deconvolution methods

Suite2P

Suite2P (https://github.com/cortex-lab/Suite2P) is actively developed by Marius Pachitariu and members of the cortexlab (Kenneth Harris and Matteo Carandini) at UCL. Suite2P's USP is it's application to large scale 2-photon imaging analysis, with an emphasis on end-to-end processing (images to neural event time series) and speed. A preprint describing the toolbox is available here:

http://biorxiv.org/content/early/2016/06/30/061507,

and our own notes on the spike detection algorithm are here:

https://drive.google.com/open?id=1NeQhmoRpS-x8R0e84w3TqkUR1PNMXiem6ZljJta-U7A.

MLSpike

MLSpike (https://github.com/mlspike) was developed by Thomas Deneux at INT, CRNS Marseille, France. A model-based probabilistic approach, MLSpike was developed to recover spike trains in calcium imaging data by taking baseline fluctuations and cellular properties into account. A comprehensive explanation of the algorithm and its benefits can be found in the paper:

Deneux, Thomas, Attila Kaszas, Gergely Szalay, Gergely Katona, Tamás Lakner, Amiram Grinvald, Balázs Rózsa, and Ivo Vanzetta. "Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo." Nature Communications 7 (2016).

Link: https://www.nature.com/articles/ncomms12190

MLSpike can return a maximum a posteriori spike train, or a spike probability per time step. We show results for both denoted MLSpike$_{events}$ and MLSpike$_{pspike}$

LZero

The method we refer to as LZero was developed by Sean Jewell and Daniela Witten from U.Washington, Seatle, USA. The goal for this implementation was to cast spike detection as a change-point detection problem, which could be solved with an existing l_0 optimization algorithm. In their paper Jewell and Witten show that the l_0 solution is better than previously implemented l_2 solutions, with results much closer to the real spike train (l_2 solutions tend to overestimate the true firing rate). Details can be found in the paper:

Jewell, Sean, and Daniela Witten. "Exact Spike Train Inference Via l_0 Optimization." arXiv preprint arXiv:1703.08644 (2017).

Link: https://arxiv.org/abs/1703.08644

Yaksi

Yaksi refers to the 'vanilla' deconvolution of Yaksi and Friedrich (2006). This is to be used as a baseline for comparison with more sophisticated methods. ~~***NOTE 8.6.17*** my implementation results in signals that are more temporally smooth (as opposed to more temporally sharp) than the calcium signal, indicating the filtering has not been performed properly.~~

The method is detailed in the paper:
Yaksi, Emre, and Rainer W. Friedrich. "Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging." Nature Methods 3, no. 5 (2006): 377-383.

Peron events

Peron events refer to the extracted events detailed in the original *Peron et al. 2015* paper. It is a version of the 'peeling' algorithm tuned to generate a low number of false positive detections (a rate of 0.01Hz) on ground truth data, leading to a hit rate of 54%
Peron, Simon P., Jeremy Freeman, Vijay Iyer, Caiying Guo, and Karel Svoboda. "A cellular resolution map of barrel cortex activity during tactile behavior." Neuron 86, no. 3 (2015): 783-799.

Events + kernel versions

Where a spike inference method returns spike rates per time point, these are plotted as Method$_{events}$. To compare to other methods that return a de-noised df/f or firing rate estimate, these events are convolved with a calcium kernel and plotted as Method$_{kernel}$.

5 Supplemental

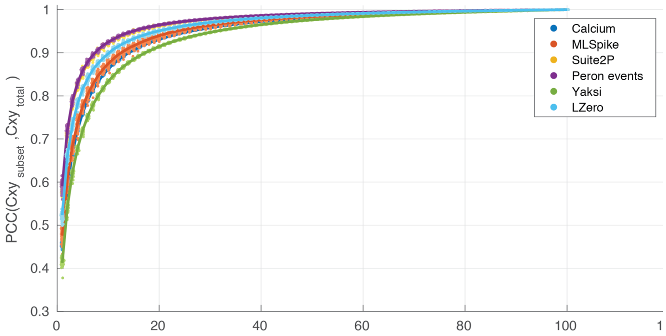

Figure 12: Example datasets are long enough to generate stable correlation estimates. Correlation between the pairwise correlation matrix for a given method, and an equivalent correlation matrix for subsets of the data. For each datapoint in the figure a subset (1%-100%) of the full dataset is extracted at random without replacement and a matrix of pairwise correlations is generated. These correlations are then compared to the matching pairwise correlations in the full dataset. In all instances 20% of the data is sufficient to recover correlations of 0.9, though there is substantial variation between methods.