On the use of calcium deconvolution algorithms in practical contexts

Mathew H. Evans, Mark D. Humphries

September 21, 2018

Abstract

Fluorescence imaging of somatic calcium (Ca^{2+}) is an increasingly popular technique for recording the activity of large groups of neurons. To improve signal quality the fluorescence time-series is typically denoised and deconvolved prior to analysis. Deconvolved Ca^{2+} is an indirect measure of spiking activity, therefore some methods go further and infer the timing and number of spikes underlying the Ca^{2+} trace. Recent efforts have seen an explosion in the number of available methods for Ca^{2+} deconvolution and spike inference, with impressive results. Here we evaluate this progress by comparing the performance of deconvolution algorithms in practical contexts. We find that good estimates of spike rate can be recovered on ground truth data, but only if spike-inference methods are tuned to individual cell properties. We show that a commonly used metric - Pearson Correlation Coefficient - yields widely ranging results with small changes in parameters, and poor estimates of firing rate when compared to a spike-based metric. When analysing large-scale recordings from behaving mice, state-of-the-art methods are inconsistent and perform poorly when using parameters tuned to ground truth data. Estimates of event rate and correlation distributions, of silent cells, tuned cells and dimensionality vary widely between methods and affected by parameter choices. We conclude that to date there are no 'magic bullet' approaches for inferring accurate estimates of neural activity from fluorescence imaging data. We suggest that conclusions of analyses that depend on the deconvolution or spike-inference must be verified across multiple methods and analysis parameters. [Currently $\sim$250 words. Needs to be closer to 150.]

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. misclassify cells as tuned/silent. Poor estimates of data dimensionality.
- artefacts may not be removed after all

Recent community efforts have shown great progress in the speed, scale, and accuracy of spike-inference methods (Berens et al., 2018)

2 Results

2.1 Spike inference methods work well on ground truth data if parameters are fitted using Error Rate instead of Pearson Correlation Coefficient

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-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 increases as estimated firing rate increases. As a consequence, if we choose the model parameters that maximise the correlation coefficient then the recovered spike-train consistently overestimates the ground truth rate of spiking (Fig 1(e,g)). Over the population correlation-coefficient-optimized Suite2P both over and under estimates firing rate, in some cases severely (Figure 1 (g,j)). In addition, correlation coefficient does not change smoothly with gradual changes in the threshold parameter, as can be seen in the lines for individual cells in Figure 1 (e), which could lead to overfit-

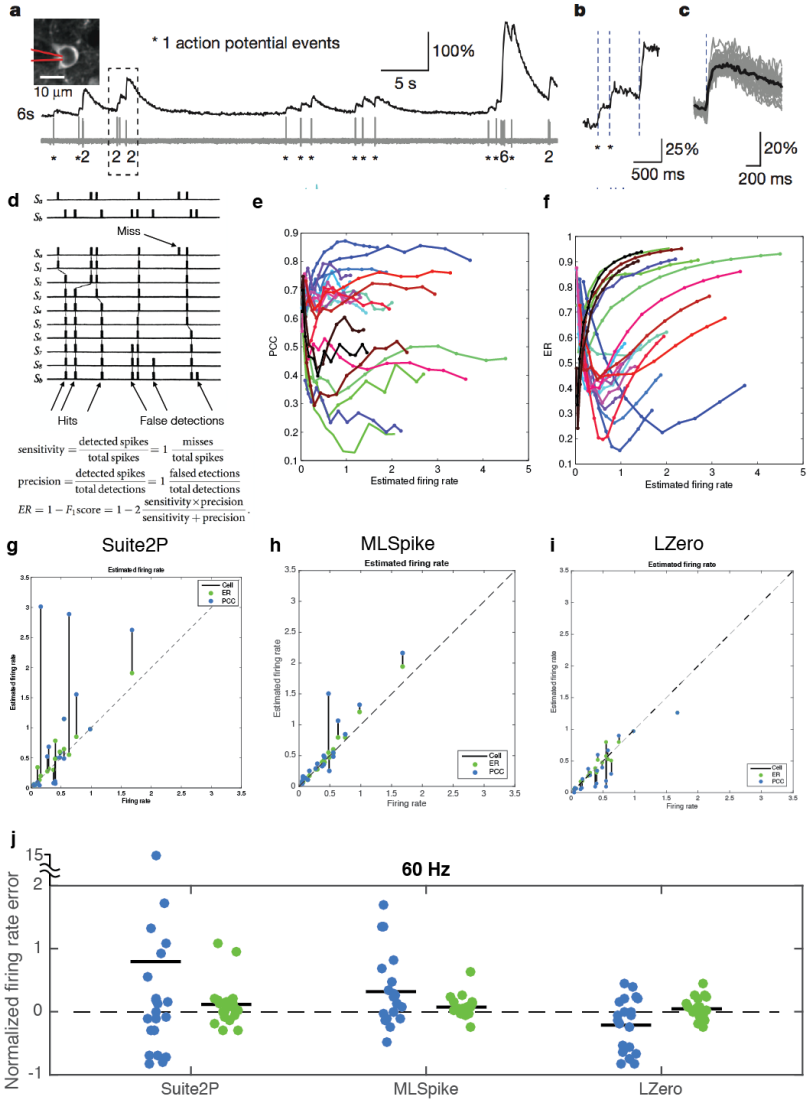

Figure 1: Ground truth data analysis. (a) Ground truth data collection example from Chen et al. (2013). Top left, an example field of view from imaging data. Red lines denote the outline of a juxtacellular recording pipette. Time series shows measured calcium fluorescence (top) and simultaneously recorded 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 here) required to match two spike trains, up to some temporal precision. Bottom: In Deneux 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) Correlation coefficient as a function of estimated firing rate (using Suite2P, Pachitariu et al.). 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 (Deneux 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 across all three methods. Lines are means. (a) reproduced from Chen et al. (2013). (d) reproduced from Victor and Purpura (1996). *MDH: Just need to bear in mind that we'll need a plan for replacing panels a-d with our own plots and schematics*

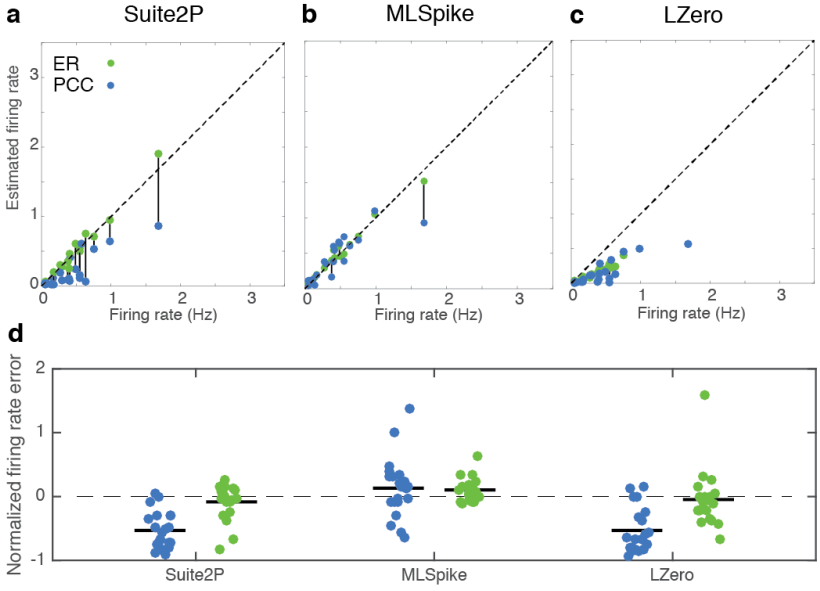

Figure 2: Downsampled ground truth analysis. (a-c) Estimated firing rate for 'best' deconvolution parameters versus real firing rate using Suite2P, MLSpike and LZero applied to ground truth data downsampled to 7Hz to more closely match population imaging experiments. (d) Normalised firing rate error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

ting or noisy estimates of the best parameters. TO DO: PLOT PARAMETER DISTRIBUTIONS SHOWING VARIABILITY IN BEST PARAMS ACROSS CELLS.

To confirm that correlation coefficient is at fault and not the particular spike inference method used we repeated the groud truth data analysis with two other recently proposed methods. Spike inference with MLSpike (Deneux et al., 2016) and an exact L-Zero optimization method (Jewell and Witten (2017), dubbed Zero here) results in poor estimates of firing rate when optimised using correlation coefficient (Fig 1(h,i,j)).

To address the weaknesses of Pearson correlation coefficient, we implemented the Error Rate (ER) spike distance metric of Deneux et al. (2016), a summary statistic based on the distance measure of Victor and 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 correlation-coefficient-optimized approach (mean error $\sim$0.1Hz, compare blue and green dots in 1 (g,j)). In addition, unlike for correlation coefficient, individual cell results in Figure 1 (f) show ER varies smoothly with gradual changes in Suite2P's threshold parameter. ER results in better estimates of firing rate than correlation coefficient when it is used to optimise parameters for the two other spike inference methods tested, MLSpike (Fig 1 (h, j middle)) and LZero (Fig 1 (i, j right)). In sum, our results show that three different spike inference methods can accurately recover firing rate if Error Rate is used to optimise model parameters on ground truth data instead of correlation coefficient.

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 et al., 2015a). Therefore it is important to ask whether the problems of using Pearson correlation coefficient 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 - Error-Rate-optimized methods out performed correlation-coefficient-optimised methods (Fig 2). Interestingly, correlation-coefficient-optimised methods often underestimated the firing rate of cells at 7Hz imaging rate (Fig 2 (d)), where overestimation had been the usual result with 60Hz data (Fig 1).

Together, these results show - as reported comprehensively recently (Berens et al., 2018) - that modern spike-deconvolution methods can accurately recover neural activity, but the choice of metric for evaluation and fitting is critical. Pearson correlation coefficient 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. A spike-train-based method such as Error Rate(Deneux et al., 2016; Victor and Purpura, 1996), or other recently developed methods (Reynolds et al., 2017; Theis et al., 2016) are more appropriate.

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

We have shown that different spike inference methods can recover good estimates of neural activity if parameters are set appropriately. However, it is not possible to fit parameters to representative ground truth data for most experiments. On a real-world example, does spike inference result in better estimates of neural activity than more simple deconvolution or de-noising processes?

To determine the effect of analysis method and parameter choice on simple measures of neuron activity in the absence of ground truth, we compared the results of analysis of df/f Ca^{2+} from a single experiment from Peron et al. 2015a (see Figure 3) to eight different approaches for deconvolution, de-noising and spike inference. We compared the three spike inference methods tested in Section 2.1 - Suite2P, MLSpike and LZero - to a kernel-convolved version of the returned spikes (i.e. de-noised df/f); de-noised Ca^{2+} as reported in the original Peron et al. (2015a) study; and the simple deconvolution approach of Yaksi and Friedrich (2006) (see Methods for implementation details).

In Peron et al. 2015a two-photon Ca^{2+} imaging was used to record neural activity from up to $\sim$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 an approximately log normal distribution at the population level (Wohrer et al., 2013). 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

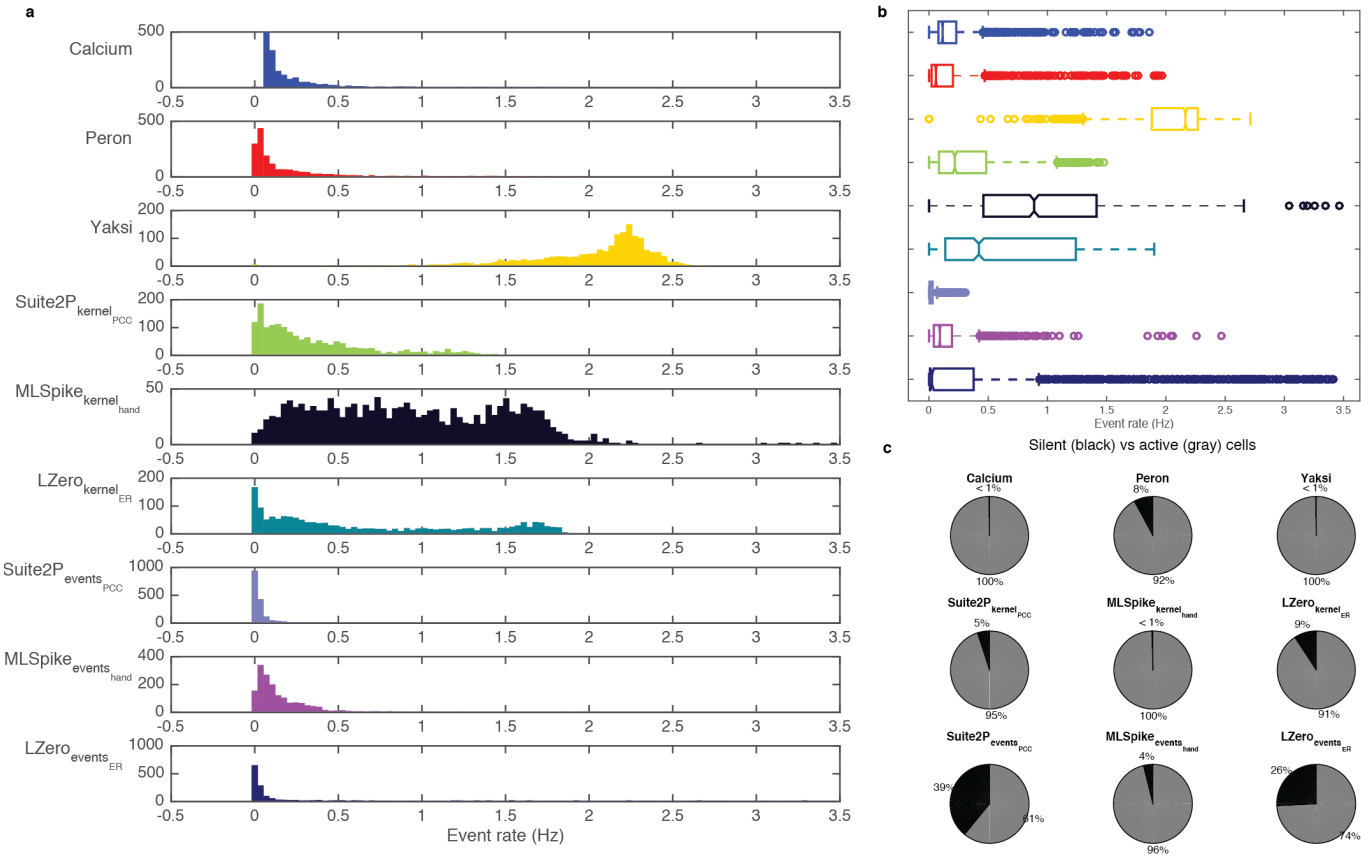

Figure 4: Estimated 'event rate' for all cells in an example session. For the first 6 methods (Calcium - LZero$_{kernel}$), events are detected as fluorescence transients greater in magnitude than 3 std deviations of background noise. Background noise = data - smoothed version of data, to eliminate slow 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.

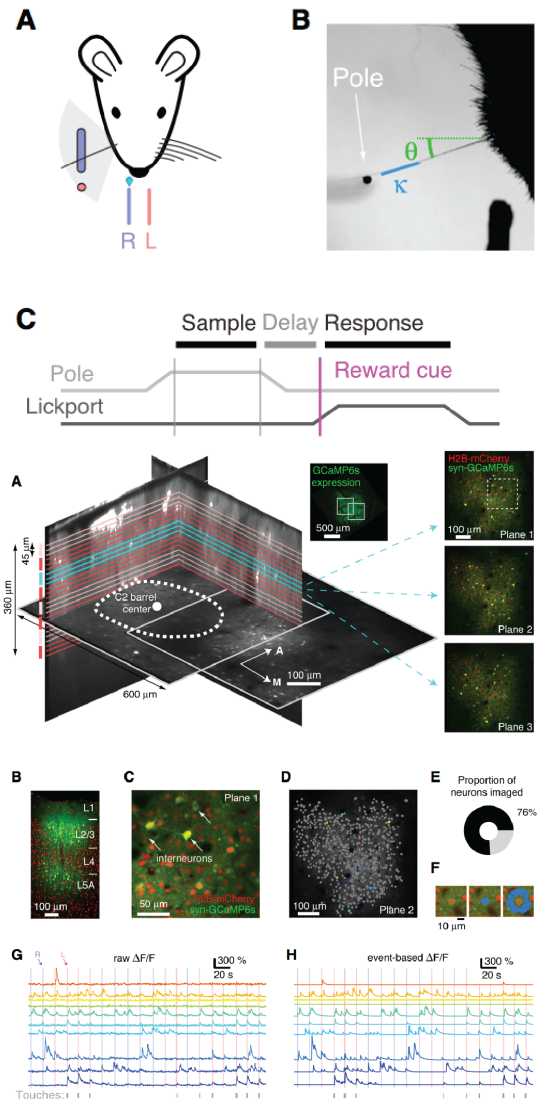

Figure 3: (Peron et al., 2015a) experimental design. TO DO EXPAND description once final figure arrangement is decided *MDH: Again, we need to think about how to remove or replace these figure panels for publication. We don't need the bottom rows of A-F here for example. For our purposes, we only care that there is a single recording of many neurons during a task. So versions of the bottom row G and H from the actual session we use would be good for the final version. We can redraw the top row A and C; For the parameters in top panel B, as far as I know these aren't used here (just "touch"), so an be eliminated too. If we need them, then the redrawn A panel can have the angle and curvature parameters from B*

tails), but they disagree quantitatively. **TO DO: Add back in 'straw man' results based on best ground truth parameters?**. *Need to clarify 'hand tuning' and put the GT derived params first, then hand tuned results.*

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, in seven the estimated 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.

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

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 involved in the task, and serve as a target for future causal or manipulation studies. We quantified the proportion of task related neurons in our dataset following the approach of Peron et al. 2015a. Calcium/instantaneous firing rate/ events for each cell are 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 (Fig. 5 (a)). Secondly, the methods only agree on the tuned status of individual neurons for $<$50 cells (from a range of 50 - 250 tuned cells Fig. 5 (b)). **TO DO get precise number?**.

There is still substantial disagreement when comparing methods within class: whether looking at deconvolution and de-noising methods (Fig. 5 (e), left) or spike inference methods (Fig. 5 (e), right), all methods only agree on the tuned status of $<$ 50 cells. This result is problematic, as this disagreement could mean either (a) some tuned cells are missed (b) some un-tuned cells are classed as tuned (c) both. Looking at each cell's tuned status in more detail (Fig. 5 (d)) it becomes clear that while some cells are only classified as tuned by one or two methods, there is wider agreement between methods about a larger group of cells **TO DO get proportion**, suggesting agreement between methods may a robust cue to tuning.

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.

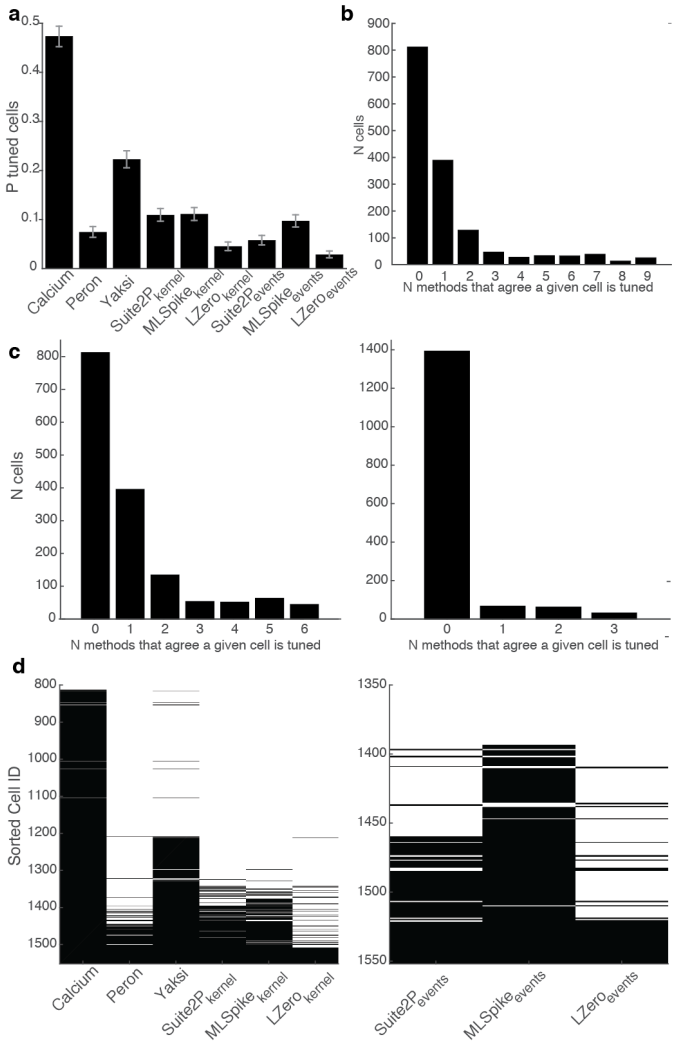

Figure 5: Tuned cells. Tuned cells were determined through shuffle tests (following Peron et al 2015, see Methods). (a) Number of tuned cells per deconvolution method. Error bars are 95% binomial confidence intervals (Jeffreys interval) (b) Agreement between methods. Bars show total number of cells classified as tuned by N methods. (c) Agreement between continuous signal methods (left) and spike inference methods (right). **TO DO add method class as title**. (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).

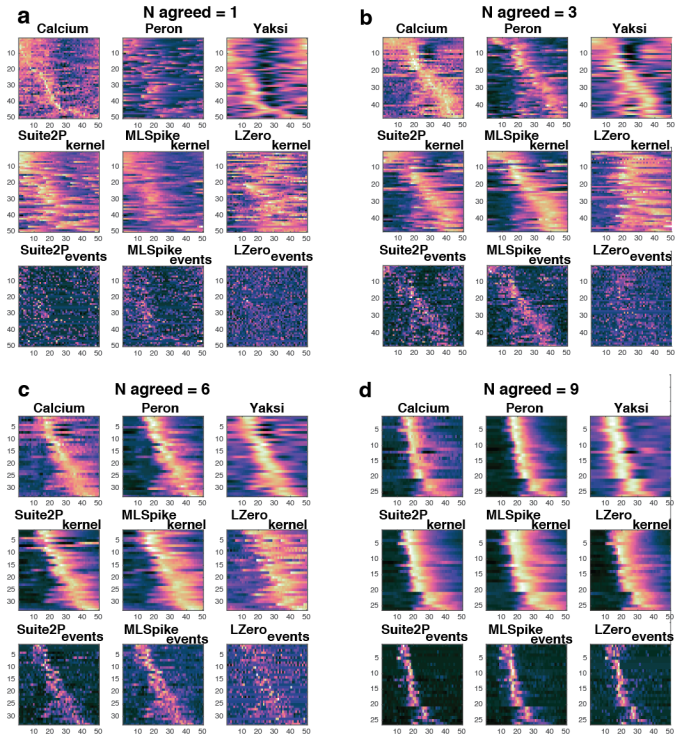

Figure 6: Degree of agreement between methods can identify strongly tuned cells. (a) Example normalised (z-score) trial-average histograms for 50 cells (rows) classified as tuned in an analysis of raw Calcium data. Each subsequent panel shows trial-average histograms for the same cells, but following 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?**

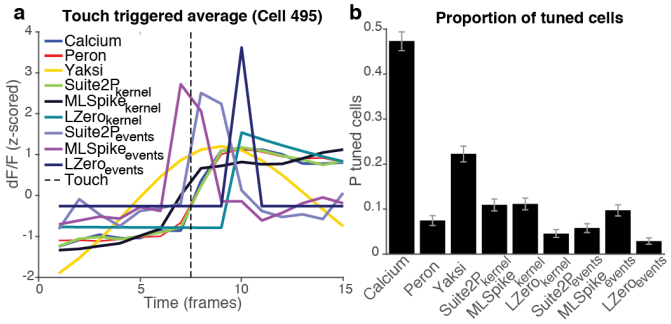

Figure 7: Touch-related responses. (a) Comparing touch-triggered average (mean deconvolved FR per imaging frame) from different deconvolution methods for one example cell. Touch occurs during frame 7 (dotted line). (b) Number of touch tuned cells varies across methods. A cell is classed as touch tuned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

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

Experimental manipulations or task events often result in precisely timed responses in some neurons. To determine whether deconvolution and spike inference improve 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 to spike with short latency and low jitter (O'Connor et al., 2010; Hires et al., 2015). To determine touch tuning for each cell we found the peak in the touch triggered 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).

Spike inference or deconvolution is often employed to increase the ability to detect temporally sharp responses by reducing background noise and removing the slow kinetics of Ca^{2+} changes. For clearly tuned cells this approach appears valid - in Fig 7 (a) the touch-triggered averages for an example cell show a temporally sharp peak around touch for the spike inference methods. However, as for task tuning, different methods disagree on the number of touch-tuned cells. Fig 7 (b) shows the proportion of cells classed as touch-tuned after processing signals with each method. Of particular note, the spike inference methods disagree substantially on this score, with one method (MLSpike) estimating 300% more touch-tuned cells than another (LZero). Without more analysis it is unclear which - if any - is correct.

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

- signal to noise

- temporal resolution (rise/decay time)

2.5 Pairwise correlation distributions are affected by spike inference and deconvolution

A goal of many Ca^{2+} imaging experiments is to record from populations of neurons, and then perform clustering or dimensionality reduction. These analyses often rely on estimates of pairwise correlations (CITATION NEEDED). The methods tested in this study are designed to remove noise and sharpen temporal responses, which would lead to more accurate estimates of pairwise correlation. Figure 8 (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.

Description Deconvolution/de-noising methods (Yaksi, Suite2P$_{kernel}$, MLSpike$_{kernel}$, LZero$_{kernel}$) have broad distributions more similar to that resulting from smoothing the raw Ca^{2+}. Suite2P$_{kernel}$ and Peron have median PCCs below zero, suggesting these methods are actively decorrelating the data (choosing parameters that penalise false-positives). Yaksi's distribution is symmetric (like all the noise surrogates) suggesting dirt has been added to the data. LZero$_{kernel}$ 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.

Interpretation (SPECULATIVE - DISCUSS WITH MARK):

- 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/ de-noising, 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 firing rate distributions peaked close to zero (reducing false positives) inevitably lead to more miss errors, and therefore actively decorrelate.
- Spike inference is never going to be perfect. Miss real spikes + overestimate background rates (spike inference is additive, so adding background spikes is inevitable. See also Ganmor et al. 2016 on this point), therefore correlation estimates are noisier (due to false positives/misses) and biased (due to misses therefore decorrelation, or false positives therefore higher mean correlations)

Looking in more detail at a subset of 50 cells, Fig 9 (a) shows that the disagreement between methods can be seen at the level of individual pairs of neurons. Correlations are not just scaled - some pairs that appear correlated following processing by one method (yellow box and arrows) are uncorrelated when processed with another method. Other pairs are consistently correlated across methods (green box and arrows).

Fig 9 (b) shows the correlation between correlation matrices for different methods, showing how similar the correlation matrices are across methods. In Fig 9 (c) the same data is shown but separating the spike inference methods (right) from the others for better comparison. Though there are some exceptions (Fig 9 (a)), overall there is broad agreement within method classes on which cells are more or less correlated. In particular, Suite2P$_{kernel}$ and MLSpike$_{kernel}$ are correlated with one another, as are Suite2P$_{events}$ and MLSpike$_{events}$. LZero appears to form unique correlation matrices, correlating only with itself (LZero$_{kernel}$ and LZero$_{events}$ correlate only with one another). Yaksi correlates highly with the raw Ca^{2+}, reflecting the fact that Yaksi changes the time series the least of the pre-processing methods.

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

Dimensionality reduction techniques such as eigendecomposition allow researchers to make sense of large scale neuroscience data. Often performed as a pre-processing stage ahead of visualisation or clustering, eigendecomposition provides an estimate of the dimensionality of data - the number of orthogonally separable sources of variance in the data. We applied eigendecomposition to the example data from Peron et al. 2015b processed by the eight different de-noising, deconvolution and spike-inference methods. Fig 10 shows the cumulative variance explained with increasing eigenvectors (dimensions) for each method. The number of dimensions required to explain 80% of the data varies dramatically across methods from 120 (Peron) to 720 (Calcium) **N.B. eyeballed estimates - get real numbers!**. This result indicates that the same dataset can appear low dimensional (<10% dimensions required to explain 80% of the variance) to high-dimensional (~50% of dimensions required).

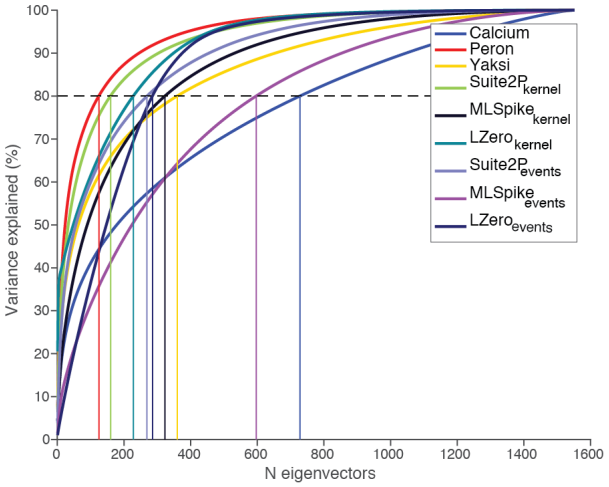

Figure 10: Cumulative variance explained by N eigenvectors following Principal Components Analysis. Different spike inference/deconvolution methods result in different estimates of the dimensionality of the data. For example, 80% of the variance can be explained by 120 or 720 eigenvectors (orthogonal dimensions) depending on the processing method used. *FIX COLOURS: Can't tell which one is the calcium, which the $MLzero_{kernel}$, and which the $LZero_{events}$. Use grey for the raw Ca2+. MDH: To give some idea of how this tunes between "low" and "high" dimensions, replot these "N" as a strip-plot (1D scatter) of the percentage of the population. The N=720 means we can't describe the data in much less than half the available dimensions = very high dimensional. Whereas the 120 = 10% of the available dimensions = low dimensional.*

TO DO: add back in (in supplement) results when deconvolution//spike inference parameters are different? In the first attempt at this analysis there was another group of methods closer to the Pachitariu result of 'full dimensionality'

Spike inference does not automatically remove experimental artefacts

MDH: Not sure we're going to include this - perhaps first need to see the proper version of the figure with the dip clearly visible in the deconvolved and inferred spike time-series 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

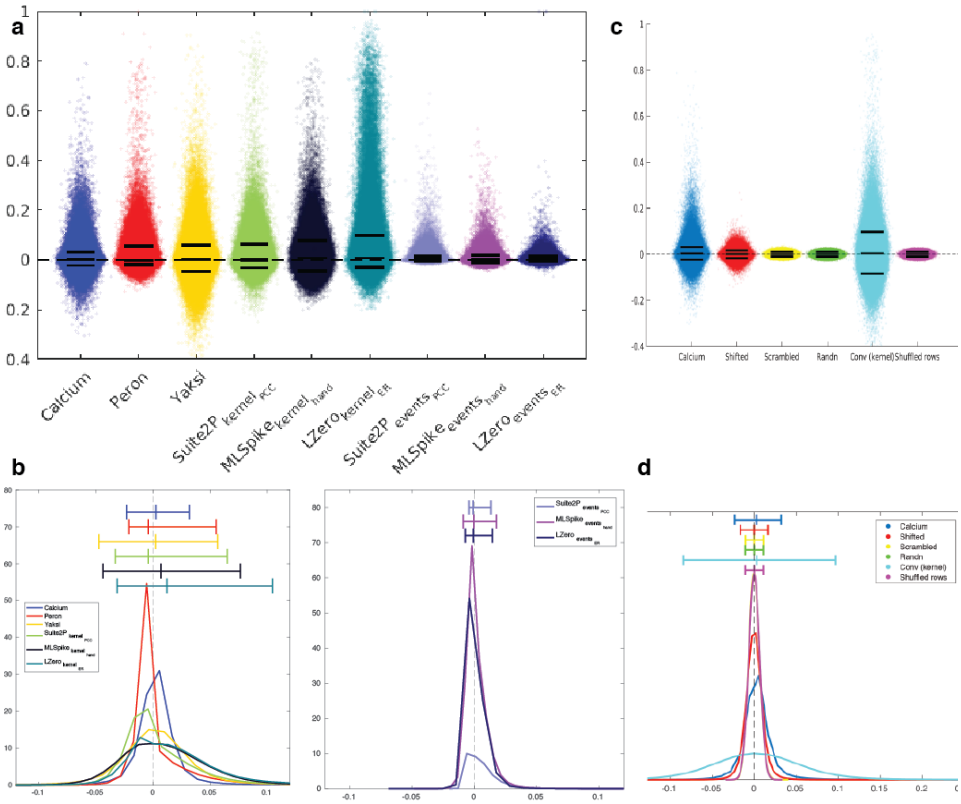

Figure 8: Pairwise correlation distributions. (a) Pairwise correlations between all cells (y-axis) following processing with all deconvolution/spike inference methods (x-axis jitter added for clarity). Solid black lines are 5th, 50th and 95th percentiles. (b) Pairwise correlation kernel density functions for all methods. Vertical lines are medians and 5th, 50th and 95th percentiles. (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).

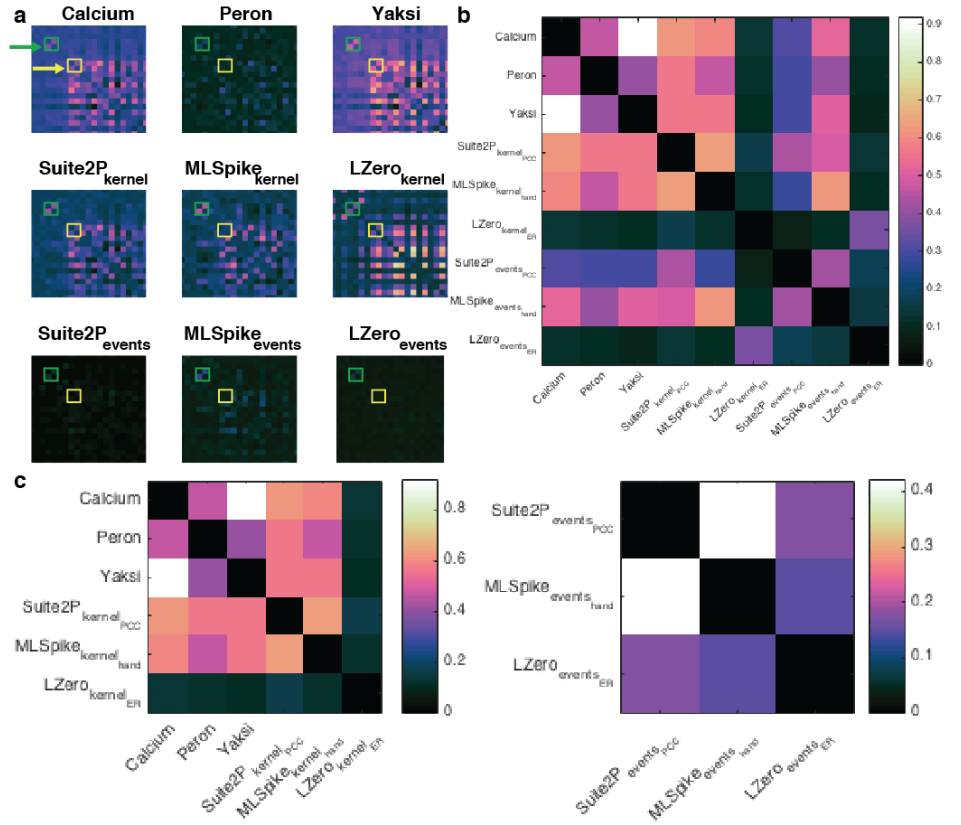

Figure 9: Pairwise correlations. (a) Example pairwise correlations for 50 cells. Some pairs of cells are consistently correlated across different methods (green arrow and boxes). Other pairs appear correlated when processed with one method but not with others (yellow arrow and boxes). (b) Correlation 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). **add label to c saying 'spike inference' and 'continuous/de-noising'**

BETTER FIG).

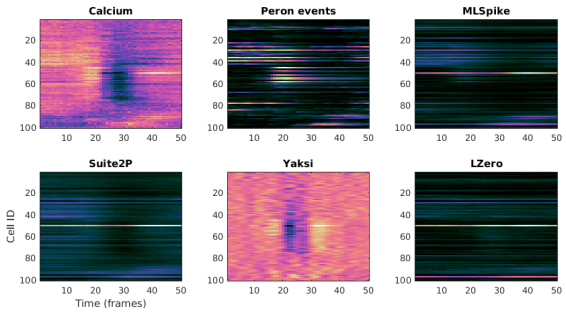

Figure 11: PLACEHOLDER FIGURE. Lick induced dip in Ca^{2+} fluorescence seen in the raw Calcium is also seen in deconvolved data.

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.

RE: Pachitariu et al biorxiv 2017 Robustness of spike deconvolution for calcium imaging of neural spiking. (a) Pachitariu et al 2017 show that PCC between inferred and true spikes can be improved with small modifications to the output of simple non-negative deconvolution algorithms. Shifting spike times by a fixed amount and smoothing the spike count with a gaussian kernel improved PCC - and therefore measured model performance - with no changes to the algorithm. This again suggests PCC is a poor metric for assessing spike inference methods. (b) Pachitariu et al 2017 suggest a novel metric for assessing spike inference/deconvolution methods in the absence of ground truth. In Pachitariu et al 2017's experiments, an ensemble of stimuli are repeated at least twice, allowing the comparison of deconvolved calcium across stimulus repeats. Algorithms that result in consistent deconvolution traces (similar results on both trials, measured with Spearman's correlation) are rated higher. This approach cannot be applied in many studies. In the Peron dataset described here, and any other studies with unconstrained behaviour of any kind, trial conditions are not precisely repeated. Therefore, consistent deconvolution/ spike inference on different trials are not meaningfully related to better algorithm performance.

In addition, Pachitariu et al's approach assumes that cells respond consistently on separate trials. This may well be the case in the sensory periphery e.g. Bale et al J.Neuroscience 2015, numerous studies have shown that this is not generally the case e.g. motor studies.

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

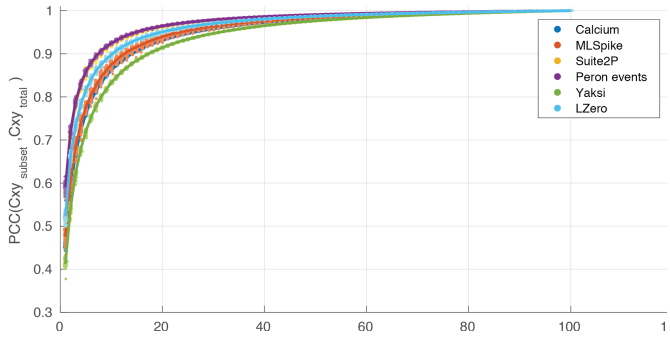

Figure 12: Example datasets are long enough to generate stable correlation estimates. Correlation between the pairwise correlation matrix for a given method, and an equivalent correlation matrix for subsets of the data. For each datapoint in the figure a subset (1%-100%) of the full dataset is extracted 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.

References

Philipp Berens, Jeremy Freeman, Thomas Deneux, Nicolay Chenkov, Thomas McColgan, Artur Speiser, Jakob H Macke, Srinivas C Turaga, Patrick Mineault, Peter Rupprecht, Stephan Gerhard, Rainer W Friedrich, Johannes Friedrich, Liam Paninski, Marius Pachitariu, Kenneth D Harris, Ben Bolte, Timothy A Machado, Dario Ringach, Jasmine Stone, Luke E Rogerson, Nicolas J Sofroniew, Jacob Reimer, Emmanouil Froudarakis, Thomas Euler, Miroslav Román Rosón, Lucas Theis, Andreas S Tolias, and Matthias Bethge. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.*, 14(5):e1006157, May 2018.

Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, Loren L Looger, Karel Svoboda, and Douglas S Kim. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, July 2013.

Thomas Deneux, 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. *Nat. Commun.*, 7:12190, July 2016.

Elad Ganmor, Michael Krumin, Luigi F Rossi, Matteo Carandini, and Eero P Simoncelli. Direct estimation of firing rates from calcium imaging data. January 2016.

Samuel Andrew Hires, Diego A Gutnisky, Jianing Yu, Daniel H O'Connor, and Karel Svoboda. Low-noise encoding of active touch by layer 4 in the somatosensory cortex. *Elife*, 4, August 2015.

Sean Jewell and Daniela Witten. Exact spike train inference via ℓ_0 optimization. March 2017.

Daniel H O'Connor, Simon P Peron, Daniel Huber, and Karel Svoboda. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. *Neuron*, 67(6):1048–1061, September 2010.

Marius Pachitariu, Carsen Stringer, Sylvia Schröder, Mario Dipoppa, L Federico Rossi, Matteo Carandini, and Kenneth D Harris. Suite2p: beyond 10, 000 neurons with standard two-photon microscopy.

Simon Peron, Tsai-Wen Chen, and Karel Svoboda. Comprehensive imaging of cortical networks. *Curr. Opin. Neurobiol.*, 32:115–123, June 2015a.

Simon P Peron, Jeremy Freeman, Vijay Iyer, Caiying Guo, and Karel Svoboda. A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron*, 86(3): 783–799, May 2015b.

Stephanie Reynolds, Simon R Schultz, and Pier Luigi Dragotti. CosMIC: A consistent metric for spike inference from calcium imaging. December 2017.

Lucas Theis, Philipp Berens, Emmanouil Froudarakis, Jacob Reimer, Miroslav Román Rosón, Tom Baden, Thomas Euler, Andreas S Tolias, and Matthias Bethge. Benchmarking spike rate inference in population calcium imaging. *Neuron*, 90(3):471–482, May 2016.

J D Victor and K P Purpura. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J. Neurophysiol.*, 76(2):1310–1326, August 1996.

Adrien Wohrer, Mark D Humphries, and Christian K Machens. Population-wide distributions of neural activity during perceptual decision-making. *Prog. Neurobiol.*, 103:156–193, April 2013.

Emre Yaksi and Rainer W Friedrich. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved ca2+ imaging. *Nat. Methods*, 3(5): 377–383, May 2006.