On the use of calcium deconvolution algorithms in practical contexts

Mathew H. Evans, Rasmus S. Petersen & Mark D. Humphries

February 15, 2019

Abstract

Fluorescence imaging of somatic calcium (Ca^{2+}) is an increasingly popular technique for recording the activity of large groups of neurons. Recent efforts have seen an explosion in the number of available methods for Ca^{2+} deconvolution and spike inference to improve signal quality, 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, silent cells, tuned cells, dimensionality and correlation distributions vary widely between methods and are 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 deconvolution or spike-inference must be verified across multiple methods and analysis parameters. [Currently ~220 words. Needs to be closer to 150.]

1 Introduction

Imaging is cool and increasingly popular.

However the signal has a number of known artefacts. Background noise and neuropil subtraction. Slow calcium kinetics. Difference in expression across cells. Known experimental artefacts (movement). Deconvolution is a process designed to elevate these problems.

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 community efforts have shown great progress in the speed, scale, and accuracy of deconvolution and spike-inference methods under idealised conditions (Berens et al., 2018), but how well do they work in real-world experiments?

An important aspect of comparing performance of different methods is the chosen metric. Pearson Correlation Coefficient (PCC) is commonly used due to its perceived interpretability (Theis et al., 2016), and common use as a metric of model fit across neuroscience. While different methods give varying results due to analysis and algorithm design choices, they may appear to converge when assessed with PCC (Berens et al., 2018).

OR how much of this convergence is due to using PCC as a metric?

Here we set to assess the performance of state of the art spike inference and deconvolution methods in real-world contexts. On ground truth data, we show that the metric used for fitting parameters affect the ability to recover simple measures of neural activity such as firing rate. On large-scale imaging data we show that ground-truth derived parameters lead to biased estimates of firing rates. Further, different methods disagree on simple measures such as the number of silent or tuned cells, and dimensionality. In addition, artefacts in the data are not automatically removed.

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

First we assessed the performance of three state-of-the-art spike inference methods under ideal conditions. The methods we tested were Suite2P (Pachitariu et al.), MLSpike (Deneux et al., 2016) and an exact ℓ_0 optimization method (Jewell and Witten (2017), dubbed LZero here). These methods were chosen because of their state-of-the-art performance, difference in underlying algorithm, and open source code. Performance of these methods was evaluated on publicly available ground truth datasets - where the spiking activity of a cell is recorded simultaneously with Ca^{2+} imaging using high-signal-to-noise juxtacellular recording techniques (see Figure 1 (a), Chen et al., 2013, `crcns.org`).

In many spike inference methods papers (Brown et al., 2004; Paiva et al., 2010; Theis et al., 2016; Reynolds et al., 2017; Berens et al., 2018) the metric used to assess performance and optimise model parameters is the Pearson Correlation Coefficient between the true and in-

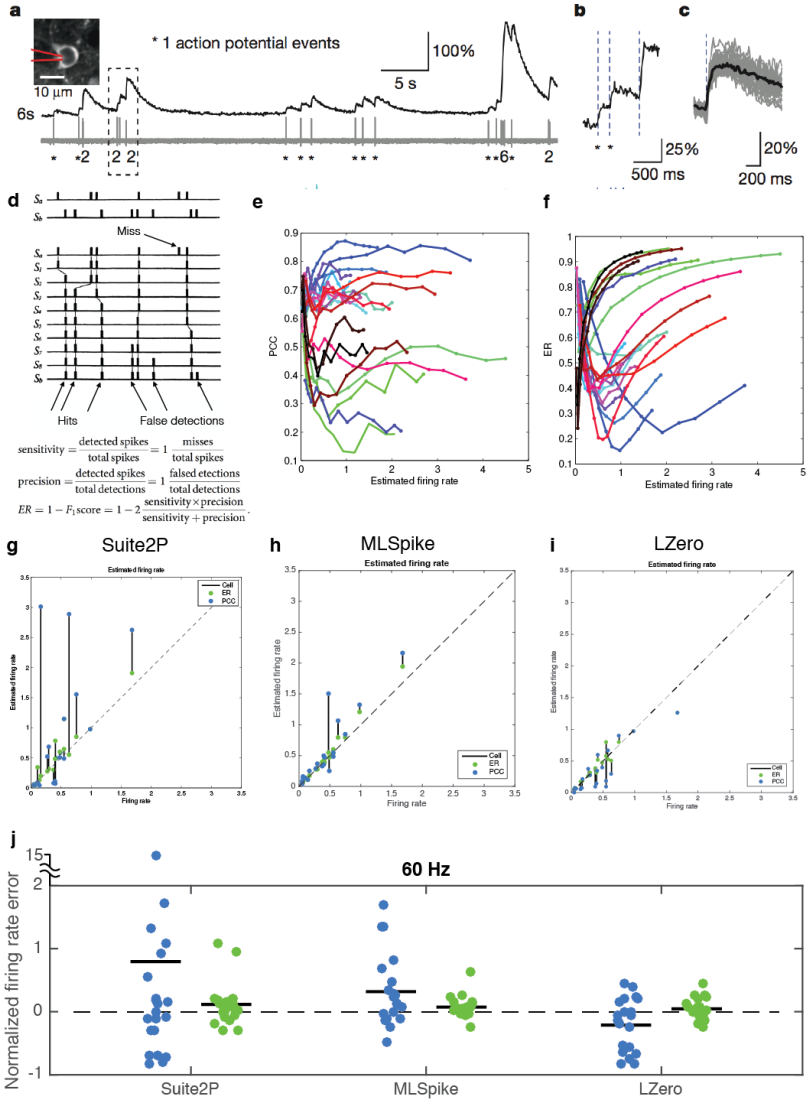

Figure 1: Ground truth data analysis. (a) Ground truth data collection example from Chen et al. (2013). Top left, an example field of view from imaging data. Red lines denote the outline of a juxtacellular recording pipette. Time series shows measured calcium fluorescence (top) and simultaneously recorded 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*

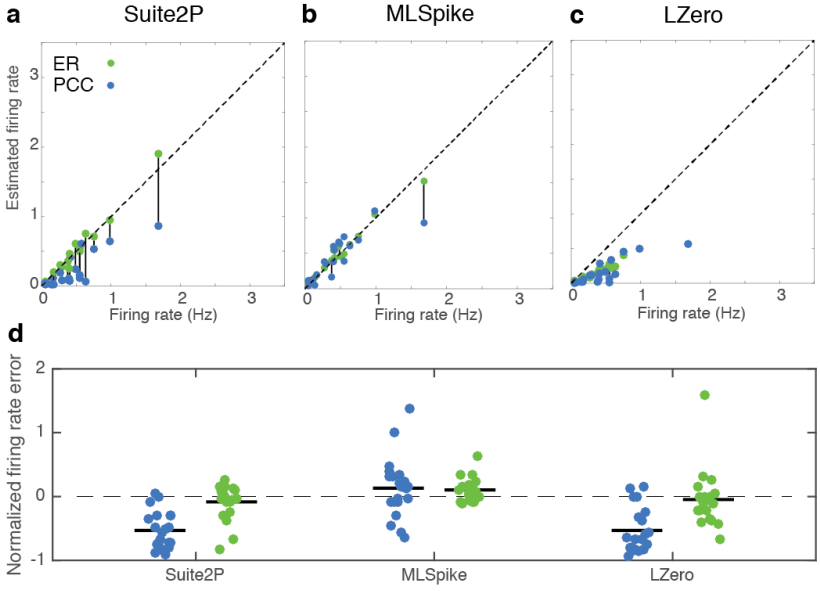

Figure 2: Downsampled ground truth analysis. (a-c) Estimated firing rate for 'best' deconvolution parameters versus real firing rate using Suite2P, MLSpike and LZero applied to ground truth data downsampled to 7Hz to more closely match population imaging experiments. (d) Normalised firing rate error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

ferred spike train, due to it's perceived interpretability and wide use in model assessment across neuroscience. Here we determined how correlation coefficient varies with small changes in model parameters, and how it relates to simple measures of neural activity such as firing rate.

Figure 1 (e) shows the results from inferring spikes from ground truth data with Suite2P (Pachitariu et al.) using a range of an internal threshold parameter which trades off misses vs false detections.

We found that correlation coefficient increases as estimated firing rate increases (Fig. 1(e)). As a consequence, if we choose the model parameters that maximise the correlation coefficient, the recovered spike-train consistently overestimates the ground truth rate of spiking (Fig 1(g), blue dots). Over the population, choosing model parameters that optimise correlation coefficient results in over and under estimates firing rate, in some cases by large margins (Figure 1 (g,j), mean error 79.47% overestimate of firing rate). We found similar results using the two other spike inference methods, with MLSpike and LZero both returning poor estimates of firing rate when optimised using correlation coefficient (Fig 1(h,i,j), blue dots, mean error 31.72% overestimates for MLSpike, 21.14% underestimates for LZero), indicating that it is correlation coefficient and not the methods per se that result in poor performance.

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 overfitting or noisy estimates of the best parameters.

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)). 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.

Across all three spike inference methods, optimising parameters with ER results in much better estimates of firing rate compared to optimisation with correlation coefficient (mean error <10% - 0.12% Suite2P, 0.073% MLSpike, 0.05% LZero, compare blue and green dots in 1 (g-j)). 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), as is the case for the ground truth datasets analysed here (60Hz imaging). 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). Given that the ground truth dataset described above was recorded at 60Hz, it is important to ask whether imaging at a lower frame rate improves the ability of Pearson correlation coefficient to recover ground truth spiking, and whether lower frame rates impair the ability of Error Rate to recover accurate spiking estimates.

To assess the effect of sampling rate on spike inference we repeated the ground truth analysis, but with imaging data down-sampled to 7Hz and found similar results to the 60Hz case. Optimising spike inference parameters with Error-Rate leads to better estimates of firing rate than optimising with correlation-coefficient (Fig 2, mean absolute error 39.7% (-53.0,13.3,-52.9) for correlation coefficient, 7.8% (-8.3,10.4,-4.7) for Error Rate). Interestingly, optimising with correlation coefficient leads to underestimates of firing rate at 7Hz (Fig 2 (d)), where overestimation is more typical with 60Hz data (Fig 1).

One goal of ground truth analysis is to find optimal analysis parameters for experiments where ground truth is not available. Figure 3 shows the best parameters for each cell across spike inference methods and sampling rates. Across all conditions, there is substantial variability across cells, with best parameters varying over two orders of magnitude in some cases (left hand panels of Fig. 3). This suggests that the best parameters for one cell may perform poorly for another cell, and optimising spike inference parameters in the absence of ground truth data may be difficult.

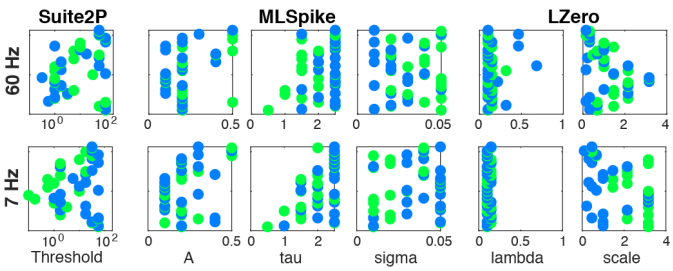

Figure 3: Best spike inference parameters varies across cells. X-axis: parameter value, Y-axis: cell ID (arbitrary but consistent order). Parameters for Suite2P (Threshold), MLSpike (A, tau, sigma) and LZero (lambda, scale) vary significantly across cells and sampling rates (rows), regardless of analysis metric (dot colour, ER: green, PCC: blue).

To further illustrate this point we analyzed how firing rate estimates and Error Rate vary with deviations from the best analysis parameters (for Suite2P applied to downsampled data, as in Fig.2). Fig. 4 (a) shows estimated firing rate departs from the true firing across a range of analysis parameters. Normalizing the results by each cell's true firing rate (Fig. 4 (b)) show that firing rate errors are particularly large for low firing rate neurons - a particular concern for applications of spike inference to data from cortex in awake animals where low firing rates

are the norm (O'Connor et al., 2010; Wohrer et al., 2013). Fig.4 (c) shows how Error Rate changes as parameters are varied from the best parameters in Fig.2. Error Rate increases abruptly with small increases or decreases in the best parameters. Fig.4 (d) shows the same data as in (c) but for a restricted range of analysis parameters, emphasising how Error Rate can increase with even very small changes in analysis parameters.

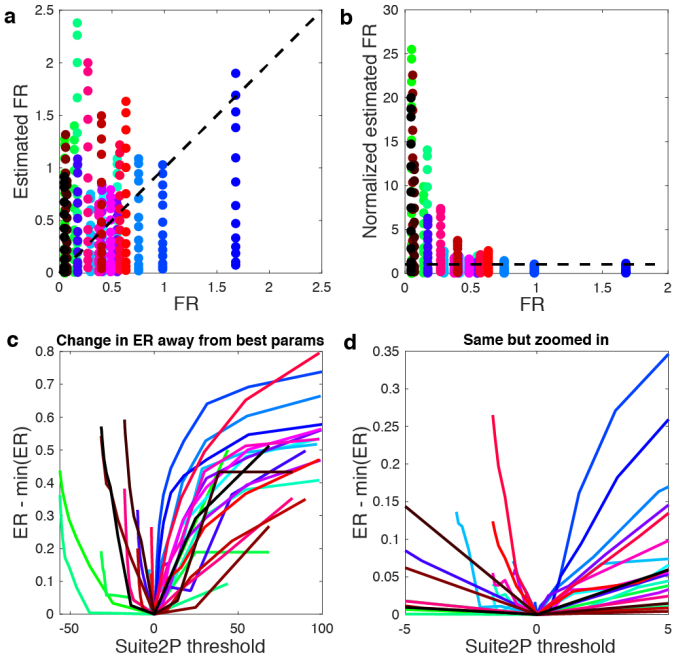

Figure 4: Variability in spike inference performance with changes in analysis parameters. (a) For each cell (colours) the estimated firing rate (y-axis) varies substantially across analysis parameters (dots). Dashed black line indicates correct estimated firing rate. (b) as in (a) for normalized estimated firing rate (Estimated FR/true FR). (c) Error rate increases sharply with small changes away from the best analysis parameters. (d) as in (c) showing a small region of parameter space around the best parameters). All results are from Suite2P applied to downsampled data as in Fig.2.

Together, these results show that modern spike-inference methods can accurately recover neural activity, but the choice of metric for evaluation and fitting of parameters are of critical importance. 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 based on information theory (Theis et al., 2016) or fuzzy set theory (Reynolds et al., 2017), are more appropriate. However, while good estimates of neural activity can be achieved with modern spike inference methods the best parameters vary substantially between cells, and small changes in analysis parameters result in poor spike inference performance. This suggests spike inference may not be successful when ground truth data is not available, and parameters cannot be optimised for each cell.

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

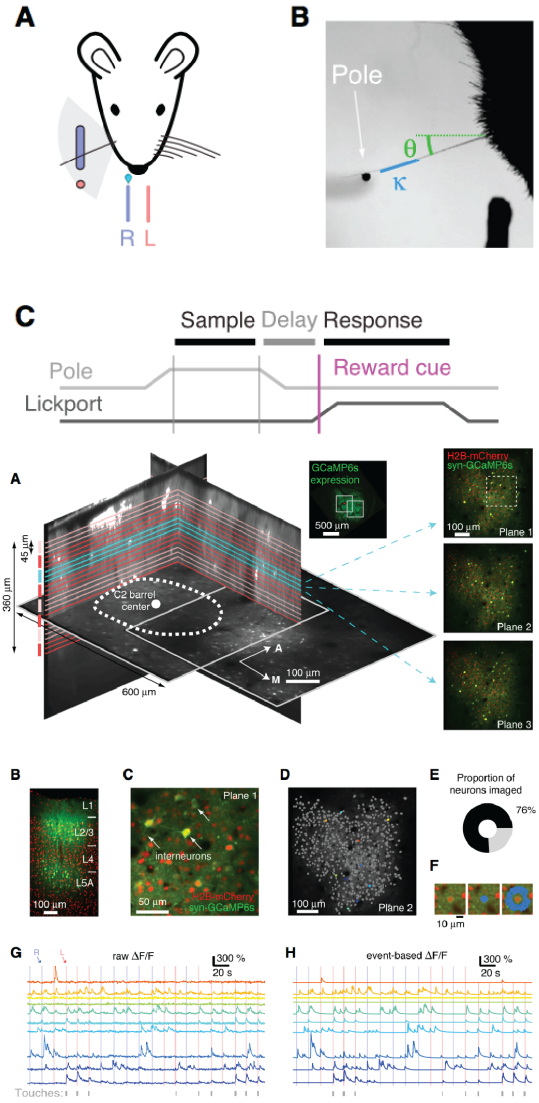

Figure 5: (Peron et al., 2015a) experimental design. TO DO EXPAND description once final figure arrangement is decided *MDH: Again, we need to think about how to remove or replace these figure panels for publication. We don't need the bottom rows of A-F here for example. For our purposes, we only care 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*

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 good estimates of neural activity, and how do these estimates compare to 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 5) 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 (e.g. pairwise correlations are stable, see Fig. S1).

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 (see Methods). Figure 6 (a) and (b) show that no two methods return the same distribution of spike/event rates. While some methods produce qualitatively correct distributions (median near zero, long right skewed tails), they disagree quantitatively (Calcium, Peron, Suite2P$_{events}$, LZero$_{events}$). Other methods appear to overestimate the average firing rate of the population as well as the number of cells with high firing rates (Yaksi, MLSpike$_{events}$). Applying a kernel to inferred spikes affects the result in interesting ways (Suite2P$_{kernel}$, MLSpike$_{kernel}$, LZero$_{kernel}$). Distributions become broader, and mean event rates are increased, suggesting noise in the spike inference process is amplified through the additional processing step of convolution with a kernel and thresholding.

It has been estimated from cell-attached recordings that $\sim$13% of somatosensory neocortical cells are silent during the pole localisation task (cells emitting fewer than one spike every two minutes, O'Connor et al. 2010), a quantity increasing to $\sim$26% in Layer 2/3. For the nine approaches we tested, in six the estimated proportion of silent cells to be below 8%, with wide disagreement between the other three methods (Figure 6 (c)). Even for simple statistics, the choice of deconvolution or spike inference method results in widely different results.

2.3 Estimates of the number of task related neurons are affected by analysis method

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 a task or experiment 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 further study. We quantified the proportion of task related neurons in our dataset following the approach of Peron et al. 2015a. Processed Calcium fluorescence for each cell was shuffled in time before trial-averaged activity was calculated, and the largest peak in each shuffled average was recorded. This is repeated 10,000 times.

A distribution of shuffled peak trail-average magnitudes is generated, and if the peak of the true (data) trial-average is larger than the 95%ile of the shuffled distribution - in either the 'lick left' or 'lick right' 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+} fluorescence, and in comparison to one another (Fig. 7 (a)). Secondly, the methods only agree on the tuned status of individual neurons for 21 cells (from a range of 44 - 734 tuned cells Fig. 7 (b)).

There is still substantial disagreement when comparing methods within class: whether looking at deconvolution and de-noising methods ('continuous methods', Fig. 7 (c), left) or spike inference methods (Fig. 7 (c), right), all methods only agree on the tuned status of $<$ 50 cells (38 for continuous methods, 25 for spike inference methods). This result is problematic, as this disagreement could mean either (i) some tuned cells are missed (ii) some un-tuned cells are classed as tuned (iii) both. Looking at each cell's tuned status in more detail (Fig. 7 (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, suggesting agreement between methods may a robust cue to tuning.

Figure 8 shows the normalised trial-averaged activity (rows) for cells classified as tuned when looking at Calcium fluorescence data only (Fig 8 (a)), or where multiple methods agree that the cells are tuned (Fig 8 (b-d)). It is unclear whether peaks in Calcium activity in (a) are artefactual or real, as they are often missing in the trial-averaged activity from other methods (compare panels in Fig 8 (a)). However, in cells classified as task-related by 6 or more methods (Fig 8 (c,d)) 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 to increase the reliability of fluorescence imaging analysis.

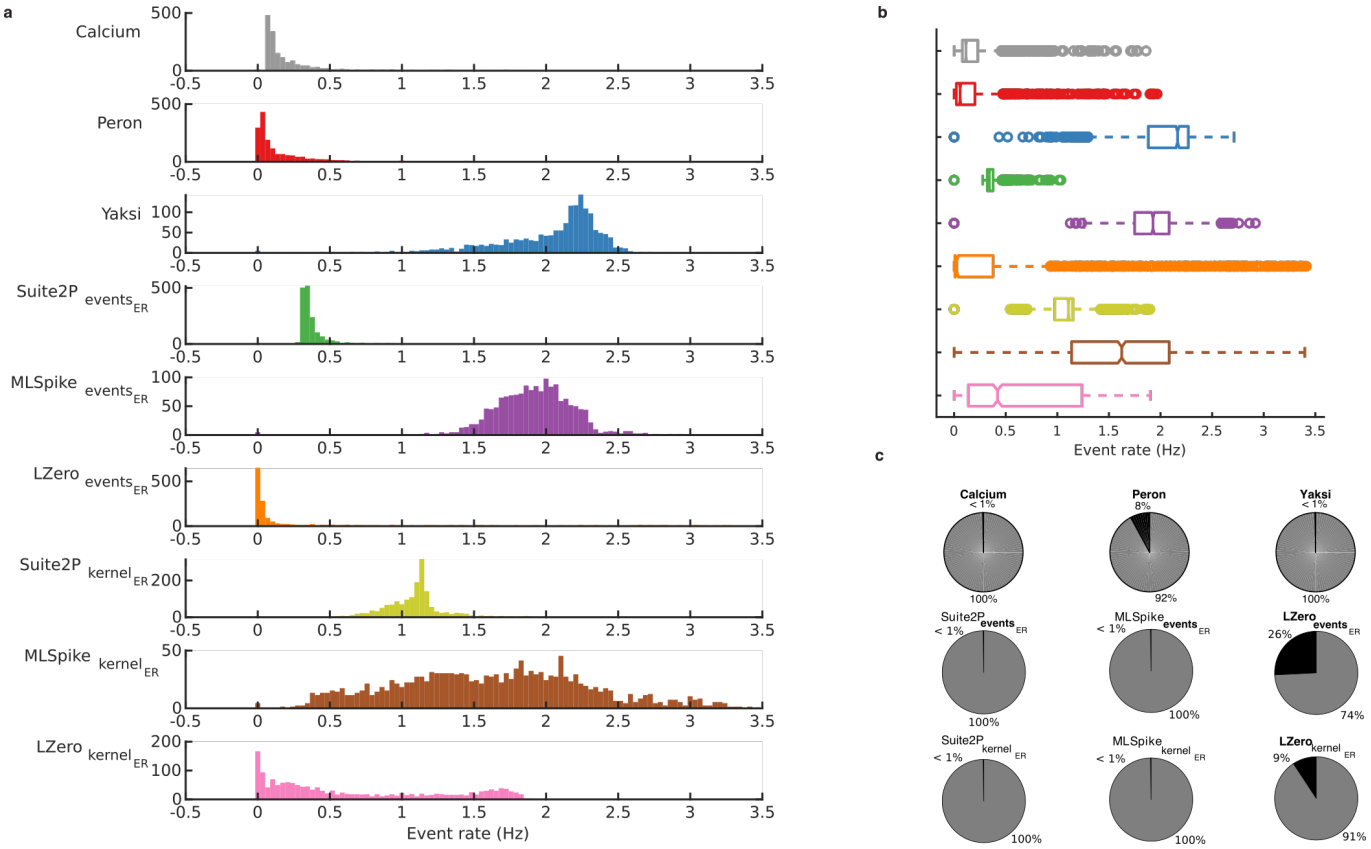

Figure 6: Estimated 'event rate' for all cells in an example session. For the first 6 methods (Calcium - LZero$_{kernel}$), events are detected as fluorescence transients greater in magnitude than 3 std deviations of background noise. Background noise = data - smoothed version of data, to eliminate slow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

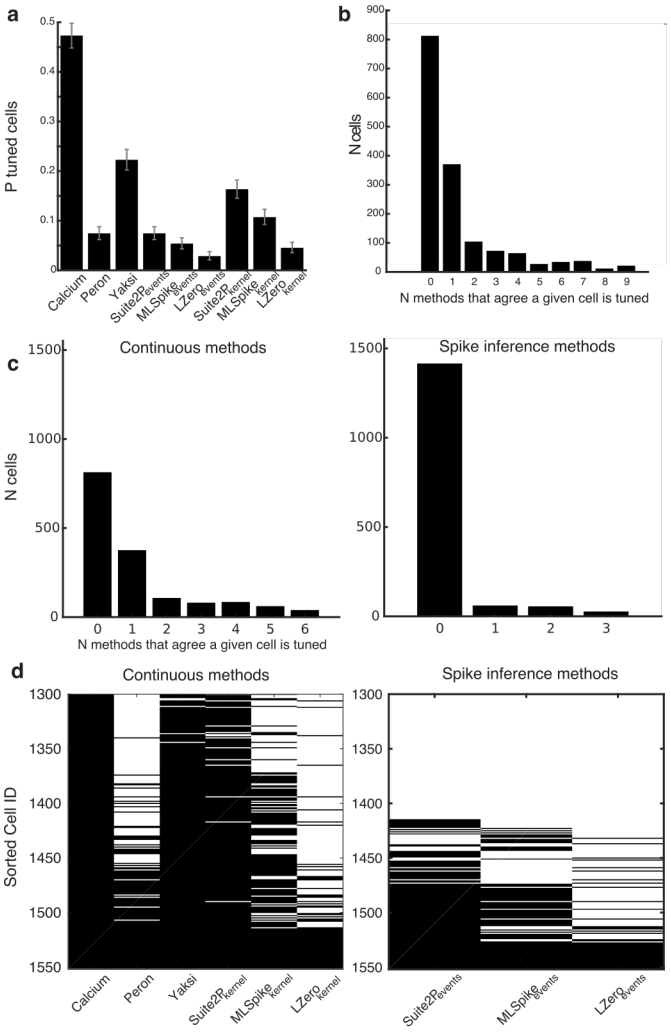

Figure 7: Tuned cells. Tuned cells were determined through shuffle tests (following Peron et al 2015, see Methods). (a) Number of tuned cells per deconvolution method. Error bars are 95% binomial confidence intervals (Jeffreys interval) (b) Agreement between methods. Bars show total number of cells classified as tuned by N methods. (c) Agreement between continuous signal methods (left) and spike inference methods (right). (d) Arrays 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 (i.e. agreement, as plotted in c).

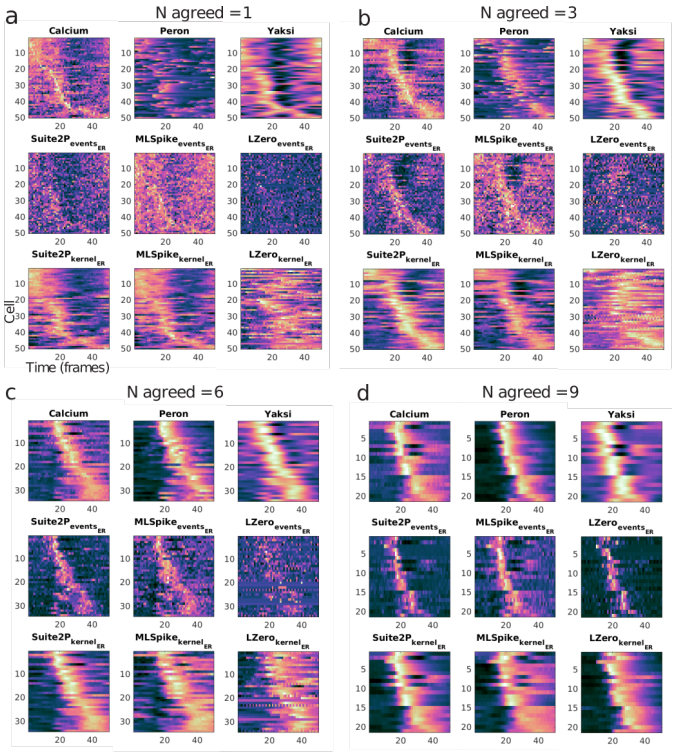

Figure 8: Degree of agreement between methods can identify strongly tuned cells. (a) Example normalised (z-score) trial-average histograms for 50 cells (rows) classified as tuned in an analysis of raw Calcium data, ordered by the time of peak activity. Each subsequent panel shows trial-average histograms 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. *NB: Add marker to show which methods classify which cells as tuned?*

2.4 Spike inference allows recovery of precisely timed responses at the cost of sensitivity

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 activity (15 imaging frames peri-touch-time, 133 touches), and compared the data distribution (one data point per touch) at this time point to a matched 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 9 (a) the touch-triggered average for an example cell shows a temporally sharp peak around touch time for the spike inference methods (MLSpike$_{events}$, Suite2P$_{events}$, LZero$_{events}$, purple, green and orange lines respectivey). However, as for task-tuning (Fig 8), different methods disagree on the number of touch-tuned cells. Fig 9 (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$_{events}$) estimating 4500% more touch-tuned cells than another (LZero$_{events}$) *i.e. 45 tuned cells vs 1 tuned cell*. Therefore, while spike inference results in recovery of temporally sharp neural responses in robustly tuned neurons, this comes at the cost of robust detection of tuned cells.

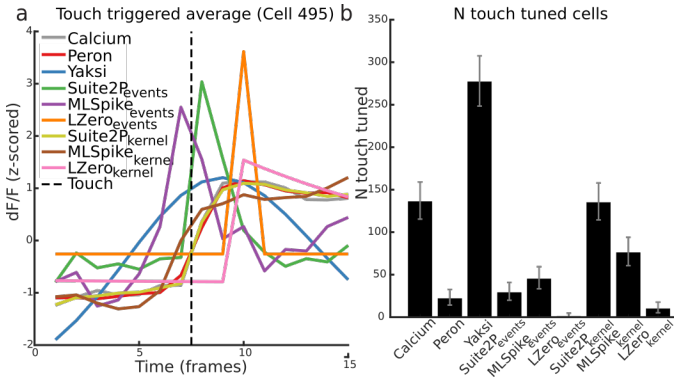

Figure 9: Touch-related responses. (a) Comparing touch-triggered average (mean activity per imaging frame across touch events) from different deconvolution methods for one example cell. Touch occurs during frame seven (dotted line). (b) Number of touch-tuned cells varies across methods. A cell is classed 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

2.5 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 correlation or covariance (Okun et al., 2015; Cunningham and Yu, 2014). The deconvolution and spike inference methods tested in this study are designed to remove noise, sharpen temporal responses and lead to improved estimates of neural activity, steps which could lead to more accurate estimates of pairwise correlation. Figure 10 shows 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 (see methods, *Fig in supplement?*).

Each method changes the pairwise correlation distribution in a different way. Most continuous methods (Yaksi, Suite2P$_{kernel}$, MLSpike$_{kernel}$, LZero$_{kernel}$) have broad distributions similar to those resulting from smoothing the raw Ca^{2+}. Peron has median PCC below zero, suggesting this method is actively decorrelating the data, perhaps due to choosing parameters that penalise false-positive spike detection. Yaksi's distribution is symmetric (like all the noise surrogates) suggesting noise has been added to the data. LZero resulted in very sparse time series, so the long tail of positive correlations in LZero$_{kernel}$ are likely to be the large group of almost silent cells. Spike inference methods all have sharply peaked distributions but medians varying from slightly negative to to slightly positive.

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 photonic shot noise (smoothing the time series) these methods increase the temporal correlations in the data, leading to stronger correlations (long tails on the distributions)
- 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 11

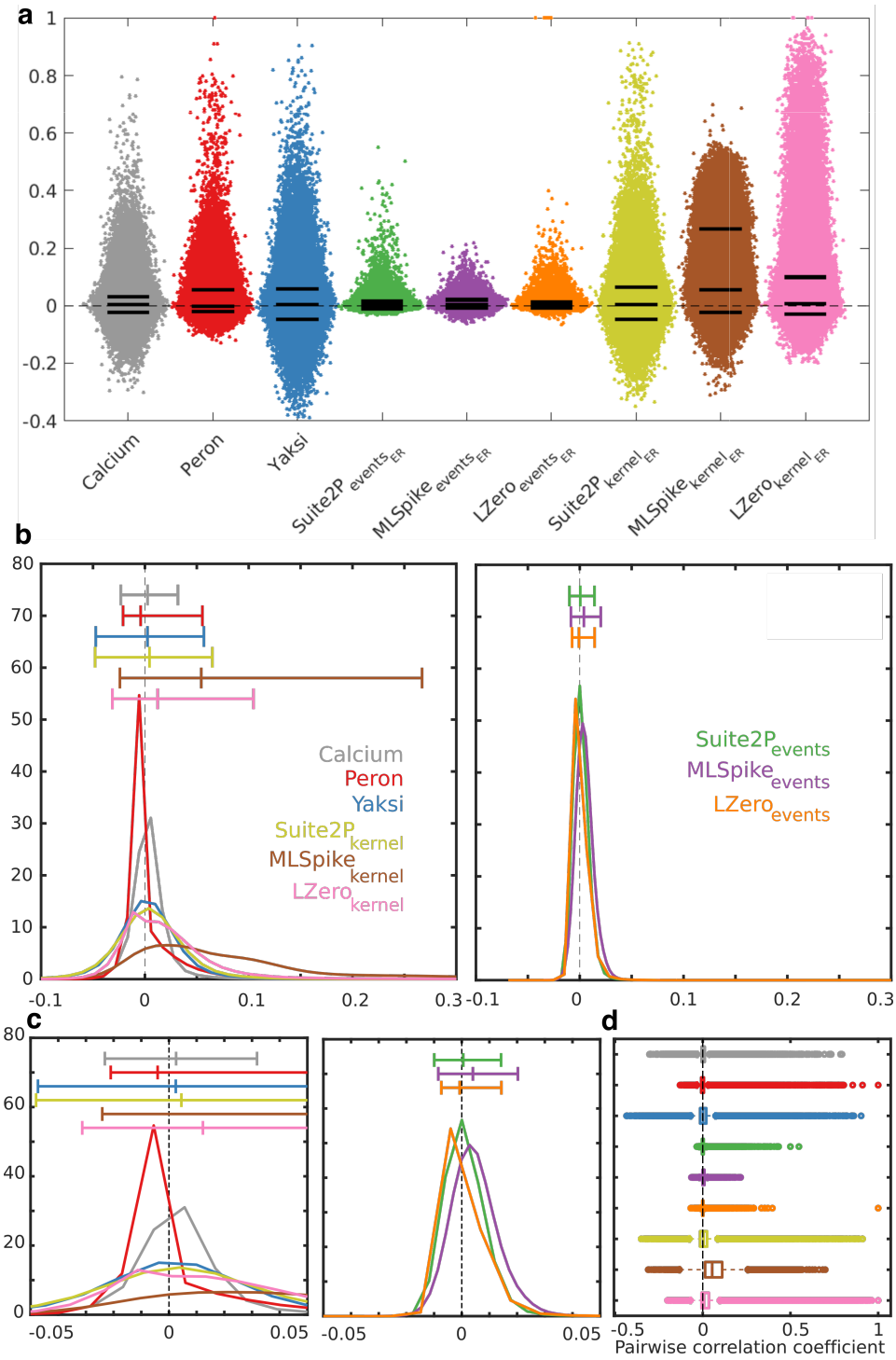

Figure 10: Pairwise correlation distributions. (a) Pairwise correlations between all cells (y-axis) following processing with all deconvolution, de-noising and spike inference methods (x-axis jitter added for clarity). Solid black lines are 5th, 50th and 95th percentiles. (b) Pairwise correlation kernel density functions for continuous (left) and spike inference (right) methods. Axes are shared between figures. Horizontal and vertical bars show medians and 5th, 50th and 95th percentiles. (c) Same data as (b) but showing a restricted range of the data. (d) Boxplot version of (a) for clarity

(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 boxes and arrow) are uncorrelated when processed with another method. Other pairs are consistently correlated across methods (green boxes and arrow).

Fig 11 (b) shows the correlation between correlation matrices for the nine different approaches, showing how similar the correlation matrices are across some methods. In Fig 11 (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 11 (a)), overall there is broad agreement within method classes on which cells are more or less correlated. In particular, Calcium, Yaksi and Suite2P$_{kernel}$ 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.

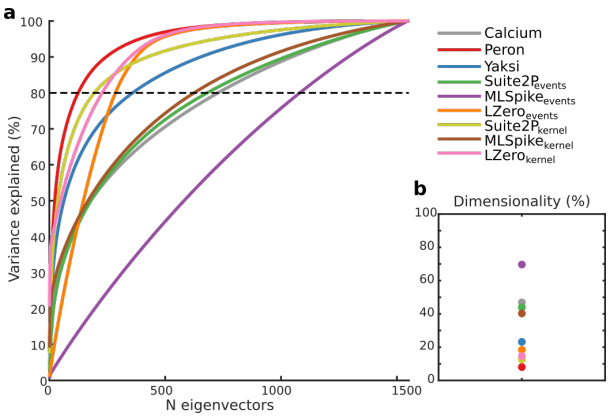

Figure 12: Dimensionality. (a) Cumulative variance explained by N eigenvectors following Eigendecomposition of the data covariance matrix for each method. (b) Proportion (percentage) of eigenvectors required to explain 80% of the variance in the data. Different pre-processing methods result in widely different estimates of the dimensionality of the data.

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 - the number of orthogonally separable sources of variance in the data. We applied eigendecomposition to the example data from Peron et al. 2015b and that processed by the eight different de-noising, deconvolution and spike-inference methods. Fig 12 (a) 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 (Fig 12 (b)) from 125 (Peron) to 1081 (MLSpike$_{events}$) *(i.e. 8% and 70%)*. This result indicates that the same dataset can appear low dimensional ($<$10% dimensions required to explain 80% of the variance) or high-dimensional (>50% of dimensions required) depending on analysis method. Critically, there is no consistent pattern within each method class - spike inference methods result in the highest (MLSpike$_{events}$) and amongst the lowest (LZero$_{events}$) dimensional datasets. Peron analysis resulted in the lowest dimensional dataset. As in the pairwise correlation analysis, the parameter choices of Peron (aiming to eliminate false positives) seem to have resulted in de-correlated neural activity and a compact representation, likely missing some sources of variance in the data.

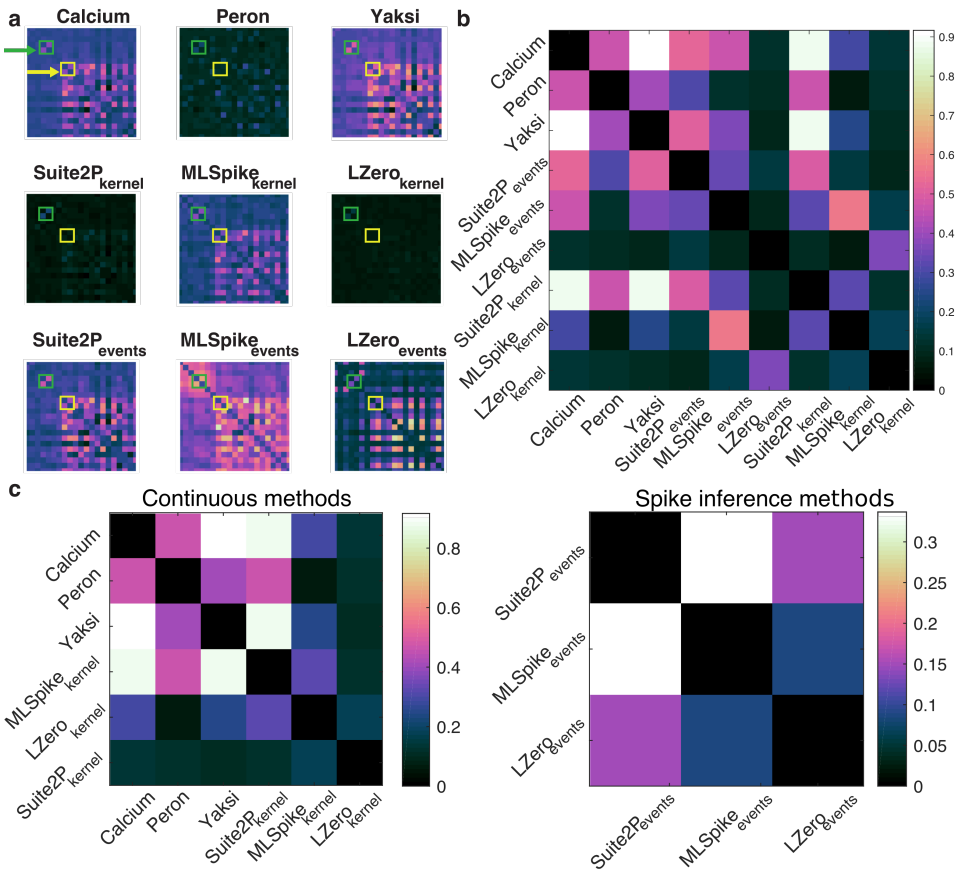

Figure 11: Pairwise correlations. (a) Example pairwise correlations for 50 cells. Some pairs of cells are consistently correlated across different methods (green arrow and boxes). Other pairs appear correlated when processed with one method but not with others (yellow arrow and boxes). (b) Correlation 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).

3 Discussion

MDH NOTES IN ITALICS: 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., 2016)). Specifically, PCC will not change between two cells if the firing rate is doubled or halved. Therefore neither false positives (FPs) nor false negatives (FNs) 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 8

*(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 this short discussion from Harris et al. (2016)

> *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 - i.e. it's not a mysterious black box, you just need to be careful. For example, you get more/less FPs/FNs depending on whether you are trying to explain every wrinkle in the Calcium trace vs match empirical firing rate distributions - as Peron have done. 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. many cortical studies (Goris, Movshon, Simoncelli on variability), motor cortex studies.

4 Methods

Ground truth data

Ground truth data was accessed from crcns.org (Svoboda, 2015), and the experiments have been described previously Chen et al. (2013). Briefly, mouse visual cortical neurons expressing the fluorescent Calcium reporter protein GCamP6s were imaged with two-photon microscopy at 60Hz. Loose-seal cell-attached recordings were performed simultaneously at 10kHz for twenty one recordings from nine cells.

Spike train metrics

Pearson correlation coefficient was computed between the ground truth and inferred spike trains following convolution with a gaussian kernel (61 samples wide, 1.02 seconds). *NB this is different to the Spikefinder method of taking the PCC from a downsampled (25Hz) signal, by convolving the spikes with a 40ms square kernel (equivalent to 15 samples here), though the area under both kernels is similar (sum of the gaussian kernel = 18).*
Error Rate was computed between the ground truth and inferred spike trains using the Deneux et al. (2016), implementation of normalised error rate, derived from Victor and Purpura (1996) Error Rate (code available https://github.com/MLspike). Briefly, the error rate is 1 - F1-score, where the F1-score is the harmonic mean of sensitivity (number of missed spikes divided by total spikes) and precision (number of falsely detected spikes divided by total detected spikes),

$$ErrorRate = 1 - 2\frac{sensitivity \times precision}{sensitivity + precision}.$$

Hits, misses and false detections were counted with a temporal precision of 0.5 seconds.

Parameter fitting

For each method the best parameters for each cell were determined by brute force search over an appropriate range. The modal best parameters, as determined using Error Rate on downsampled data, were then fixed for the population imaging data analysis.

In additional tests (*TO DO Supplemental figs, subscript 'PCC'*), the modal best parameters when assessed using correlation coefficent were used, and in the case of MLSpike parameters were additionally hand-tuned using a built-in GUI (*TO DO Supplemental figs, subscript 'hand'*).

Downsampling

Ground truth calcium data was downsampled from 60Hz to 7Hz in Matlab by up-sampling by 7 - `interp(ca,7)` and downsampling the resultant signal by 60, as Matlab's downsampling must be done in integer steps.

Population imaging data description

Population imaging data was accessed from crcns.org and have been described previously (Peron et al., 2015b). Briefly, volumetric two photon calcium imaging of primary somatosensory cortex (S1) was performed in awake head-fixed mice performing a whisker-based object localisation task. In the task a metal pole was presented on one of two locations and mice were motivated with fluid reward to lick at one of two lick ports depending on the location of the pole following a brief delay. Two photon imaging of GCaMP6s expressing neurons in superficial S1 was performed at 7Hz. Images were motion corrected and aligned, before regions of interest were manually set and neuropil-subtracted. A single recording from this dataset was used for population analysis. The example session had 1552 neurons recorded for a total of 23559 frames (56 minutes).

4.1 Event rate estimation

Spike inference methods (Suite2P$_{events}$,MLSpike$_{events}$, LZero$_{events}$) return estimated spikes which were converted into mean event rates (Hz) per cell. Event rate for continuous methods (Calcium, Peron, Yaksi, Suite2P$_{kernel}$,MLSpike$_{kernel}$, LZero$_{kernel}$) for each cell was determined by counting activity/fluorescence events greater than three standard deviations of the background noise. Background noise was calculated by taking a four-point moving average and subtracting this from the activity/fluorescence trace. Event rate was then computed in Hz.

Silent cells were defined as cells with event rates below 0.0083Hz (or fewer than one spike per two minutes of recording) as in (O'Connor et al., 2010).

4.2 Task-tuned cells

Task-tuning was determined for each neuron using the model-free approach of Peron et al. (2015b). Neurons were classed as task-tuned if their peak trial-average activity exceeded the 95th percentile of a distribution of trial-average peaks from shuffled data (10000 shuffles). The shuffle test was done separately for correct lick-left and lick-right trials and cells satisfying the tuning criteria in either case were counted at task-tuned.

Tuned cell agreement was calculated as the number of methods that agreed to the tuning status of a given cell, for all methods and separately for continuous and spike inference methods.

4.3 Touch-related responses

Touch-tuned cells were determined by computing touch-triggered average activity for each cell, before calculating whether the data distribution of peak touch-induced activity exceeds the expected activity of shuffled data. In more detail, the time of first touch - between the mouse's whisker and the metal pole - on each trial was recorded. For each touch time one second (seven data samples)

was extracted before and after the frame closest to touch (15 samples total), and the mean activity was calculated. The time of peak touch-triggered average activity was calculated, and a ranksum test (bonferroni corrected) between the true data distribution at peak time and a matched random sample of data from the same cell. This test determined whether peak touch-triggered activity was significantly different from chance.

4.4 Pairwise correlations

Pairwise correlations were calculated for all pairs of neurons in Matlab (corrcoef) at the data sampling rate (7Hz). Correlations between correlation matrices (Fig. 11) were computed between the unique pairwise correlations from each method (i.e. `CXY(find(triu(CXY)))`).

Random 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.

4.5 Dimensionality

To determine the dimensionality of each dataset we performed eigendecomposition of the covariance matrix of each dataset. The resultant eigenvalues were sorted into descending order, and the variance explained (`cumsum(egs)/sum(egs)`) plotted.

List of deconvolution methods

Suite2P

`Suite2P` (https://github.com/cortex-lab/Suite2P) is actively developed by Marius Pachitariu (HHMI Janelia) 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 (Pachitariu et al.)

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-x8R0e84w3TqkUR1PNMXiem6ZIjJta-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 et al., 2016).

MLSpike can return a maximum a posteriori spike train, or a spike probability per time step. (*TO DO: We show results for both denoted MLSpike$_{events}$ and MLSpike$_{pspike}$ in Supplement*)

LZero

The method we refer to as `LZero` was developed by Sean Jewell and Daniela Witten from U.Washington, Seattle, 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 and Witten, 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. The method is detailed in the paper: (Yaksi and Friedrich, 2006).

Peron events

`Peron events` refer to the extracted events detailed in the original Peron et al. (2015b) paper. It is a version of the 'peeling' algorithm (Lütcke et al., 2013) 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%.

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 estimates, these events are convolved with a calcium kernel and plotted as Method$_{kernel}$.

5 Supplemental

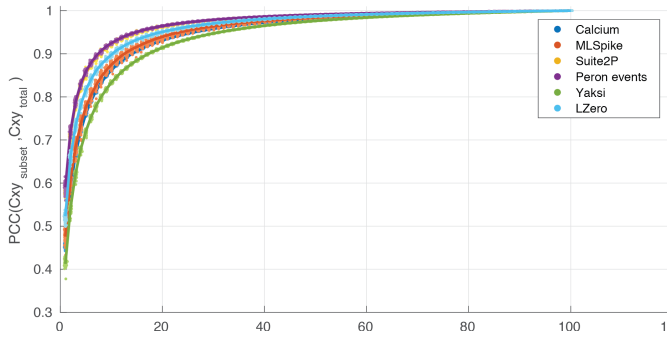

Figure S1: Example datasets are long enough to generate stable correlation estimates. Correlation between the pairwise correlation matrix for a given method, and an equivalent correlation matrix for subsets of the data. For each datapoint in the figure a subset (1%-100%) of the full dataset is extracted 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.

Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nat. Neurosci.*, 7(5):456–461, May 2004.

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.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.*, 17(11):1500–1509, November 2014.

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.

Kenneth D Harris, Rodrigo Quian Quiroga, Jeremy Freeman, and Spencer L Smith. Improving data quality in neuronal population recordings. *Nat. Neurosci.*, 19(9):1165–1174, August 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.

Henry Lütcke, Felipe Gerhard, Friedemann Zenke, Wulfram Gerstner, and Fritjof Helmchen. Inference of neuronal network spike dynamics and topology from calcium imaging data. *Front. Neural Circuits*, 7:201, December 2013.

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.

Michael Okun, Nicholas A Steinmetz, Lee Cossell, M Florencia Iacaruso, Ho Ko, Péter Barthó, Tirin Moore, Sonja B Hofer, Thomas D Mrsic-Flogel, Matteo Carandini, and Kenneth D Harris. Diverse coupling of neurons to populations in sensory cortex. *Nature*, 521(7553):511–515, May 2015.

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.

António R C Paiva, Il Park, and José C Príncipe. A comparison of binless spike train measures. *Neural Comput. Appl.*, 19(3):405–419, April 2010.

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.

K Svoboda. Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators. *GENIE project, Janelia Farm Campus, HHMI; CRCNS.org*, 2015.

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.