On the use of calcium deconvolution algorithms in practical contexts

Mathew H. Evans, Rasmus S. Petersen & Mark D. Humphries [add affilations]

May 24, 2019

Abstract

Calcium imaging is a powerful tool for capturing the simultaneous activity of large populations of neurons. Studies using it to address scientific questions of population dynamics and coding often use the raw time-series of changes in calcium fluorescence at the soma. But somatic calcium traces are both contaminated with multiple noise sources and are non-linearly related to spiking. A suite of methods are available to recover spike-evoked events from the raw calcium, from simple deconvolution to inferring the spikes themselves. Here we explore the extent to which our choice of raw or deconvolved calcium time-series affects the scientific inferences we can draw. Our results show the choice qualitatively changes the potential scientific inferences we draw about neural activity, coding, and correlation structure. We show that a substantial fraction of the processing methods fail to recover simple features of population activity in barrel cortex already established by electrophysiological recordings. Raw calcium time-series contain an order of magnitude more cells tuned to task features; yet there is also qualitative disagreement between deconvolution methods on which neurons are tuned. Finally, we show that raw and processed calcium time-series qualitatively disagree on the structure of correlations within the population and the dimensionality of its joint activity. We suggest that quantitative results obtained from population calcium-imaging be verified across multiple forms of the calcium time-series.

1 Introduction

Calcium imaging is a wonderful tool for high yield recordings of large neural populations [Harris; Stringer; Ahrens; Orger; others]. Many pipelines are available for moving from pixel intensity across frames of video to a time-series of calcium fluorescence in the soma of identified neurons [cite loads; including van Rossum latest Sci Report paper].

But raw calcium fluorescence is nonlinearly related to spiking, and contains noise from a range of sources. These issues have inspired a wide range of deconvolution algorithms [cite Theis benchmarking; Stringer Curr Opin], which attempt to turn raw somatic calcium into something more closely approximating spikes. We address here the question facing any systems neuroscientist using calcium imaging: do we use the raw calcium, or attempt to clean it up? Thus our aim is to understand if our choice matters: how do our scientific inferences depend on our choice of raw or deconvolved calcium time-series.

Deconvolution algorithm themselves range in complexity from simple deconvolution with a fixed kernel of the calcium response [Yaksi], through detecting spike-evoked calcium events [LZero, Suite2p], to directly inferring spike times [MLspike, Peeling]. This continuum of options raise the further question of the extent to which we should process the raw calcium signals.

We proceed in two stages. In order to use deconvolution algorithms, we need to choose their parameters. Wed like to know whether it is worth taking this extra step: how good can these algorithms be in principle, and how sensitive their results are to the choice of parameter values. We thus first evaluate qualitatively different deconvolution algorithms, by optimising their parameters against ground truth data with known spikes. With our understanding of their parameters in hand, we then turn to our main question, by analysing a large-scale population recording from the barrel cortex of a mouse performing a whisker-based decision task. We compare the scientific inferences about population coding and correlations we obtain using either raw calcium signals, or a range of time-series derived from those calcium signals, covering simple deconvolution, event detection, and spikes.

We find contrasting answers. A substantial fraction of the methods used here fail to recover basic features of population activity in barrel cortex established from electrophysiology. The inferences we draw about coding qualitatively differ between raw and deconvolved calcium signals. In particular, coding analyses based on raw calcium signals detect an order of magnitude more cells tuned to task features. Yet there is also qualitative disagreement between deconvolution methods on which neurons are tuned. The inferences we draw about correlations between neurons do not distinguish between raw and deconvolved calcium signals, but can qualitatively differ between deconvolution methods. Our results thus suggest care is needed in drawing inferences from population recordings of somatic calcium, and that one solution is to replicate all results in both raw and deconvolved calcium signals.

2 Results

2.1 Performance of deconvolution algorithms on ground-truth data-sets

We select here three deconvolution algorithms that infer discrete spike-like events, each an example of the state of the art in qualitatively different approaches to the problem: Suite2p (Pachitariu et al.), a peeling algorithm that matches a scalable kernel to the calcium signal to detect spike-triggered calcium events; LZero (Jewell and Witten, 2017), a change-point detection algorithm, which finds as events the step-like changes in the calcium signal the imply spikes [Check]; and MLspike (Deneux et al., 2016), a forward model, which fits an explicit model of the spike-to-calcium dynamics in order to find spike-evoked changes in the calcium signal, and returns spike times. We emphasis that these methods were chosen as exemplars of their approaches, and are each innovative takes on the problem; we are not here critiquing individual methods, but using an array of methods to illustrate the problems and decisions facing the experimentalist when using calcium imaging data.

We first ask if these deconvolution methods work well in principle. We fit the parameters of each method to a data-set of 21 ground-truth recordings (Chen et al., 2013), where the spiking activity of a cell is recorded simultaneously with 60 Hz calcium imaging using high-signal-to-noise juxtacellular recording techniques (Figure 1a). To fit the parameters for each recording, we sweep each method's parameter space to find the parameter value(s) with the best match between the true and inferred spike train.

The best-fit parameters depend strongly on how we evaluate the match between true and inferred spikes. The Pearson correlation coefficient between the true and inferred spike train is a common choice (Brown et al., 2004; Paiva et al., 2010; Theis et al., 2016; Reynolds et al., 2017; Berens et al., 2018), typically with both trains convolved with a Gaussian kernel to allow for timing errors. However, we find that choosing parameters to maximise the correlation coefficient can create notable errors. The inferred spike trains

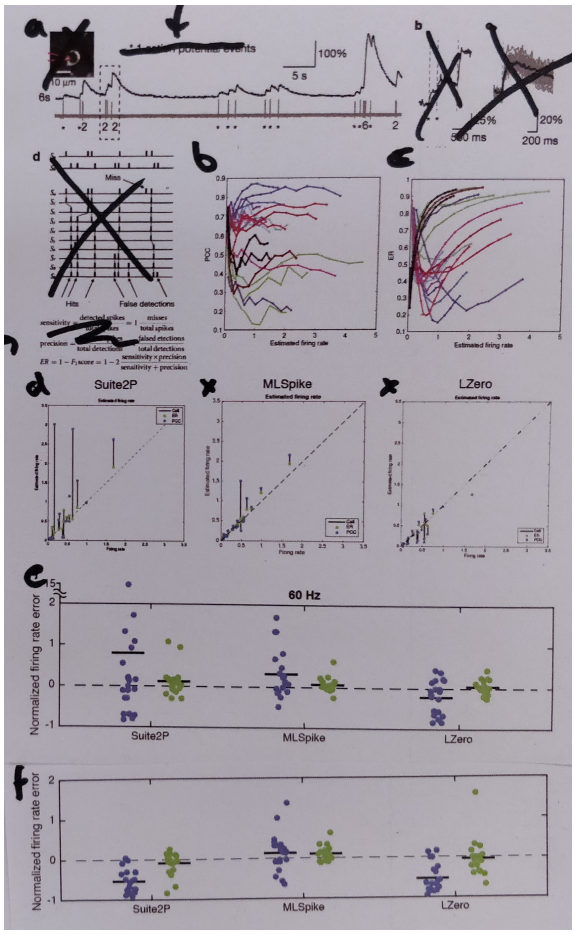

Figure 1: Ground truth data analysis.
(a) Example simultaneous recording of somatic voltage and calcium activity imaged at 60Hz from (Chen et al., 2013). Spikes are marked with asterisks.
(b) Error in estimating the true firing rate when using optimised parameters, across all three methods. One symbol per recording. We separately plot errors for parameters optimised to maximise the correlation coefficient (PCC), the errors for parameters optimised to minimise the error rate (ER). Horizontal black bars are means. Error is computed relative to the true firing rate: $(Rate_{true} - Rate_{estimated}/Rate_{true})$. For LZero and Suite2p, $Rate_{estimated}$ is computed from event times.
(c) As for (b), but with the somatic calcium down-sampled to 7Hz before optimising parameters for the deconvolution methods.
(d) Dependence of MLspike's deconvolution performance on the firing rate of the inferred spike train. For each of MLSpike's free parameters, we plot the correlation coefficient between true and inferred spikes as a function of the firing rate estimated from the inferred spikes. One line per recording. Parameters: A: [explain]; τ [explain]; σ: explain [Full range on y axis: (0,1)] [Here and throughout: lose the box axes (upper and right)]
(e) as in (d), but using Error Rate between the true and inferred spikes.
(f) Dependence of Suite2p's deconvolution performance on the firing rate of the inferred event train. Left: correlation coefficient; right: Error Rate.

from MLSpike have too many spikes on average (mean error: 31.72%), and the accuracy of recovered firing rates widely varies across recordings (Fig 1b, blue symbols). We attribute these errors to the noisy relationship between the correlation coefficient and the number of inferred spikes (Figure 1c): for many recordings, there is no well-defined maximum coefficient, especially for the amplitude parameter A [Correct description?], so that near-maximum correlation between true and inferred trains is consistent with a wide range of spike counts in the inferred trains. We see the same sensitivity for the event rates from recordings optimised using Suite2p and LZero (Figure 1f). If we compare their inferred event rates to true firing rates (Fig 1b), we see Suite2p estimates far more events than spikes (mean error 79.47%) and LZero fewer events than spikes (mean error: -21.14%). These further errors are problematic: there cannot be more spike-driven calcium events than spikes, and LZero's underestimate is considerably larger than the fraction of frames with two or more spikes [I would guess: WHICH IS?].

To address the weaknesses of the Pearson correlation coefficient, we instead optimise parameters using the Error Rate metric of Deneux et al. (2016). Error Rate returns a normalised score between 0 for a perfect match between two spike trains, and 1 when all the spikes are missed. This comparison between inferred and true spike trains is most straightforward for algorithms like MLSpike that directly return spike times; for the other algorithms, we use here their event times as inferred spikes, a reasonable choice given the low firing rate and well separated spikes in the ground truth data. Choosing parameters to minimise the Error Rate between the true and inferred spike-trains results in excellent recovery of the true number of spikes for all three deconvolution methods (Fig 1b, green symbols), with mean errors of 12% for Suite2P, 7.3% for MLSpike, and 5% for LZero. As we show in Figure 1e for MLSpike and Figure 1f for Suite2p, the Error Rate has a well-defined minima for almost every recording. Consequently, all deconvolution methods can, in principle, accurately recover the true spike-trains given an appropriate choice of parameters.

A potential caveat here is that the ground-truth data are single neurons imaged at a frame-rate of 60Hz, an order of magnitude greater than is typically achievable in population recordings (Peron et al., 2015a). Such a high frame-rate could allow for more accurate recovery of spikes than is possible in population recordings. To test this, we downsample the ground-truth data to a 7Hz frame-rate, and repeat the parameter sweeps for each deconvolution method applied to each recording. As we show in Figure 1c, optimising parameters using the minimum Error Rate still results in excellent recovery of the true spike rate (and interestingly for some recordings reduces the error when using the correlation coefficient). Lower frame-rates need not then be an impediment to using deconvolution methods.

2.2 Parameters optimised on ground-truth are widely distributed and sensitive

What might be an impediment to using deconvolution methods on population recordings is that the best parameter values vary widely between cells. Figure 2a-b plots the best-fit parameter values for each recording across deconvolution methods and sampling rates. Each method has at least one parameter with substantial variability across recordings, [SOMETHING HERE ON RANGE: are these ranges a high proportion of the total range of that parameter?]. This suggests that the best parameters for one cell may perform poorly for another cell.

The problem of between-cell variation in parameter values would be compensated somewhat if the quality of the inferred spike or event trains is robust to changes in those

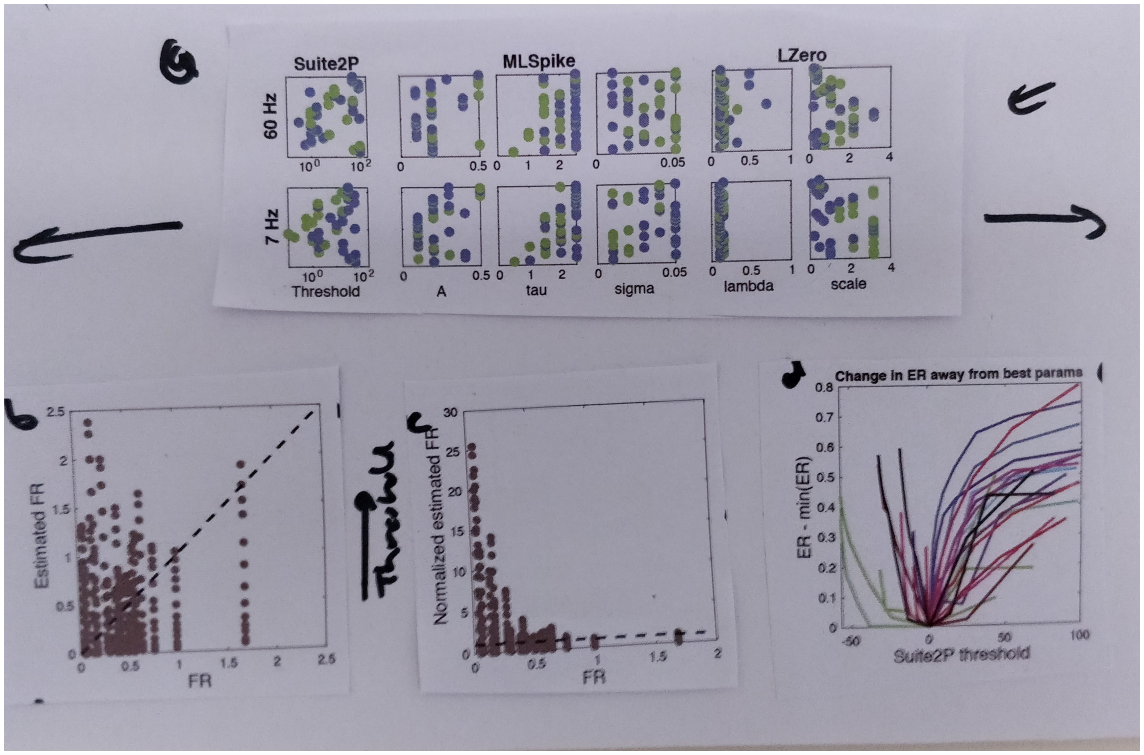

Figure 2: Variation in best-fit spike deconvolution parameters across ground-truth recordings.
(a) Distributions of optimised parameter values across recordings. In each panel, we plot parameter values on the x-axis against the recording ID on the y-axis (in an arbitrary but consistent order). Parameter values are plotted for those optimised using the Error Rate (green). Top row: fits to the original 60 Hz frame-rate data; bottom row: fits to data down-sampled to 7 Hz. [1: link recordings of the same cell] [2: add density histograms to top of each plot, to show distribution of parameters; one for PCC, one for ER]
(b) Change in error rate as a function of the change away from a parameter's optimum value, for each of MLSpike's free parameters. One line per recording. [3 panels here]
(c) Change in the error rate with change in Suite2p's threshold value away from its optimum for each recording [label x-axis correctly "Change in threshold"]. One line per recording.

values. However, we find performance is highly sensitive to changes in some parameters. Figure **??**b-c shows that for most recordings the quality of the inferred spike train abruptly worsens with small increases or decreases in the best parameter. Thus using deconvolution algorithms on population recordings comes with the potential issues that parameters can be both sensitive and vary considerably across cells.

2.3 Deconvolution of population imaging in barrel cortex during a decision task

We turn now to seeing if and how these issues play out when analysing a large-scale population recording with no ground-truth. The data we use are two-photon calcium imaging time-series from a head-fixed mouse performing a whisker-based two-alternative decision task (Fig. 3a-b), from the study of (Peron et al., 2015a). We analyse here a single session with 1552 simultaneously recorded pyramidal neurons in L2/3 of a single barrel in somatosensory cortex, imaged at 7 Hz for just over 56 minutes, giving 23559 frames in total across [X] trials of the task.

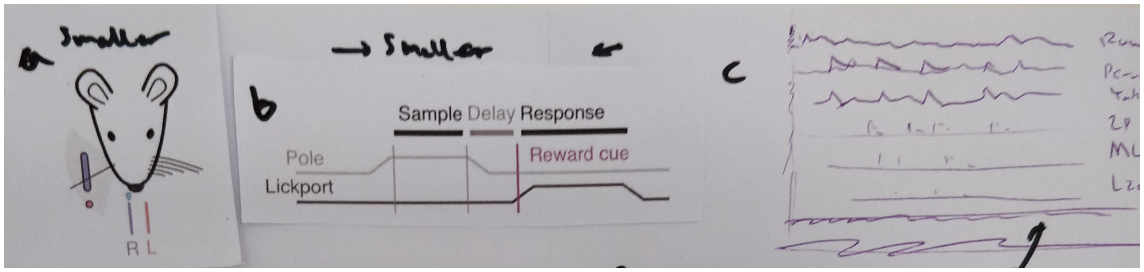

Figure 3: Experimental data from (Peron et al., 2015a).
(a) Schematic of task set-up. A pole was raised within range of the single left-hand whisker; its position, forward (red) or backward (purple) indicated whether reward would be available from the left or right lick-port. [Redraw]
(b) Schematic of trial events. The pole was raised and lowered during the sample period; a [light?] cue indicated the start of the response period. [Redraw]
(c) All deconvolution methods applied to one raw calcium signal from the same neuron.

Our primary goal is to understand how the choices of deconvolving these calcium-imaging data alter the scientific inferences we can draw. As our baseline, we use the "raw" $\Delta F/F$ time-series of changes in calcium indicator fluorescence. We use the above three discrete deconvolution methods to extract spike counts (MLSpike), event occurrence (LZero), or event magnitude (Suite2p) per frame. For comparison, we use (Peron et al., 2015a)'s own version of denoised calcium time-series, created by detecting calcium event onsets by a threshold and convolving events with a spike-response kernel. As an example of simpler methods, we use Yaksi and Friedrich (2006)'s simple deconvolution of the raw calcium with a fixed kernel of the calcium response to a single spike. And finally we create smoothed versions of the discrete-deconvolution methods, by convolving their recovered spikes/events with a spike-response kernel. Figure 3c show an example raw calcium time-series for one neuron, and the result of applying each of these 8 processing methods. We thus repeat all analyses on 9 different sets of time-series extracted from the same population recording.

We choose the algorithm parameters as follows. Simple deconvolution Yaksi and Friedrich (2006) is just kernel of the calcium response to a single spike; here is GCAMP6s[?], so parameterise kernel to that [Same as Perons kernel?]. For the three discrete deconvolution methods, we choose the modal values of the best-fit parameters that optimised the Error Rate over the ground-truth recordings. This seems a reasonably consistent choice, of using the most consistently performing values obtained from comparable data: neurons in the same layer (L2/3) in the same species (mouse), in another primary sensory area (V1). Most importantly for our purposes, choosing the modal values means we avoid pathological regions of the parameter space. [FILL IN DETAILS AND TIDY]

2.4 Deconvolution methods disagree on estimates of simple neural statistics

We first check how well each approach recovers the basic statistics of neural activity event rates in L2/3 of barrel cortex. Electrophysiology has shown that the distribution of firing rates across neurons in a population is consistently long-tailed, and often log-normal, all across rodent cortex (Wohrer et al., 2013); and L2/3 neurons in barrel cortex are no different (O'Connor et al., 2010), with median firing rates less than 1 Hz, and a long right-hand tail of rarer high-firing neurons. We thus expect the calcium event rates or spike rates from our time-series would follow such a distribution. (Event rates for raw calcium,

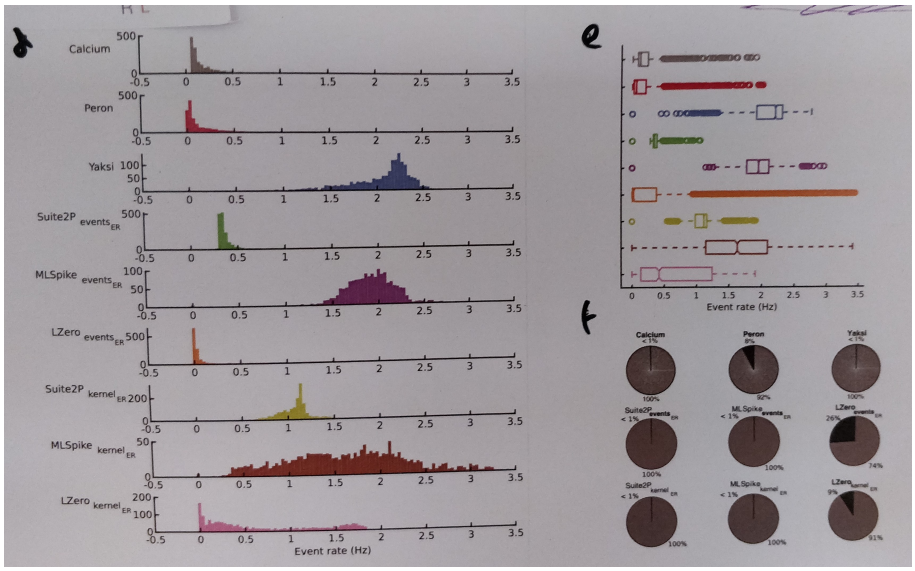

Figure 4: Estimates of population-wide event rates vary qualitatively across deconvolution methods.
(a) The distribution of event rate per neuron across the recorded population, according to each deconvolution method. For raw calcium and the five denoising methods (upper 6 panels), events are detected as fluorescence transients greater in magnitude than three standard deviations of background noise. The discrete deconvolution methods (lower 3 panels) return per frame: a spike count (MLSpike), a binary event detection (LZero), or an event magnitude (Suite2p); these time-series were thus sparse, with most frames empty.
(b) Proportion of active (gray) and silent (black) cells for each method. Silent cells are defined following (Peron et al., 2015b) as those with an event rate less than 0.0083Hz.

Peron, Yaksi and the convolved obtained by thresholding the calcium time-series)

Figure 4a shows that the raw calcium and two of the discrete deconvolution methods (Suite2p, LZero) have qualitatively correct distributions of event rates (median near zero, long right-hand tails). The Peron time-series also have the correct distribution of event rates, which is unsurprising as it was tuned to do so. All other methods give qualitatively wrong distributions of spike rates (MLSpike) or event rates (all other methods). There is also little overlap in the distributions of spike rates between the three discrete deconvolution methods. Applying a kernel to their inferred spikes/events shifts rather than smooths the firing rate distributions (Suite2P$_{kernel}$, MLSpike$_{kernel}$, LZero$_{kernel}$), suggesting noise in the deconvolution process is amplified through the additional steps of convolving with a kernel and thresholding.

Cell-attached recordings in barrel cortex have shown that ~26% of L2/3 pyramidal cells are silent during a similar pole localisation task, with silence defined as emitting fewer than one spike every two minutes O'Connor et al. 2010. For the nine approaches we test here, six estimated the proportion of silent cells to be less than 1%, including two of the discrete deconvolution methods (Figure 4c). For raw calcium and methods returning continuous time-series, raising the threshold for defining events will lead to more silent cells, but at the cost of further shifting the event rate distributions towards zero. Even for simple firing statistics of neural activity, the choice of time-series gives widely differing, and sometimes wrong, results.

2.5 Inferences of single cell tuning differ widely between raw calcium and deconvolved methods

We turn now to what we can infer about simple properties of neural coding, and how our choice of deconvolution method can alter those inferences. The decision task facing the mouse (Fig. 3a) requires that it moves its whisker back-and-forth to detect the position of the pole, delay for a second after the pole is withdrawn, and then make a choice of the left or right lick-port based on the pole's position (Fig. 3b). As the imaged barrel corresponds to the single spared whisker (on the contralateral side of the face), so the captured population activity during each trial likely contains neurons tuned to different aspects of the task. We show here that the number and identity of such task-tuned neurons in the population differ widely between deconvolution methods.

Following Peron et al. 2015a, we define a task-tuned cell as one for which the peak in its trial-averaged histogram of activity exceeds the predicted upper limit from shuffled data (Fig.5a0; see Methods). When applied to the raw calcium time-series, close to half the neurons are tuned (Fig.5a). This is more than double the proportion found for the next nearest method (Yaksi's simple deconvolution), and at least a factor of 5 greater than the proportion of tuned neurons resulting from any discrete deconvolution method, which each report less than 10% of the neurons are tuned.

Worse, few neurons are detected as tuned in time-series resulting from multiple methods (Fig.5b). Only XXX neurons are labelled as tuned in at least two sets of time-series, and just 21 (X %) are labelled as tuned in all nine. Even separately considering the continuous and discrete time-series, we find only 38 cells are tuned across all six continuous methods, and 25 neurons for all three discrete deconvolution methods (Fig.5c). Figure 5d illustrates the diversity of detected tuning even amongst the neurons with the greatest agreement between methods.

These results suggest that raw calcium alone over-estimates tuning in the population, but also that there can be substantial disagreement between deconvolution methods. One solution for robust detection of tuned neurons is to find those agreed between the raw

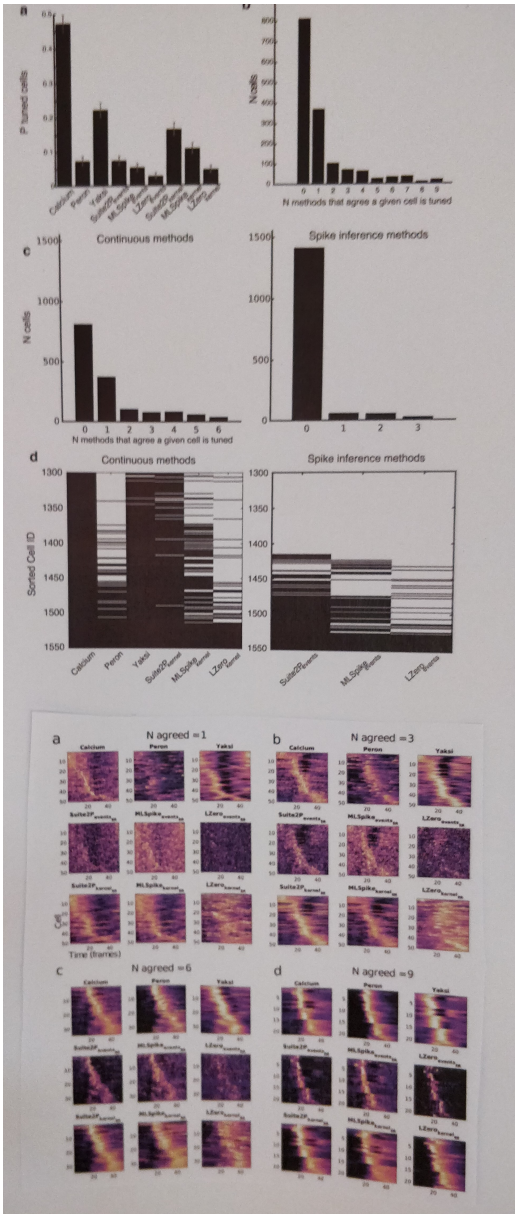

Figure 5: Inferences of single cell tuning show poor agreement between raw calcium and deconvolution methods, and between methods.
(a0) Examples of a tuned (left) and non-tuned (right) cell from the raw calcium time-series. X: Data; Y: upper 95% interval from shuffled data.
(a) Number of tuned cells per deconvolution method. Error bars are 95% binomial confidence intervals.
(b) Agreement between methods. For each neuron, we count the number of methods (including raw calcium) for which it is labelled as tuned. Bars show the number of cells classified as tuned by exactly N methods.
(c) Similar to (b), but breaking down the cells into: agreement between methods (raw or denoising) resulting in continuous signals (left panel); and agreement between discrete deconvolution methods (right panel).
(d) Comparison of cell tuning across methods. Each row shows whether that cell is tuned (black) or not (white) under that deconvolution method. Cells are ordered from bottom to top by the number of methods that classify that cell as tuned.
(e-h) Identifying robust cell tuning. Panel groups (e) to (h) show cells classed as tuned by increasing numbers of deconvolution methods. Each panel within a group plots one cell's normalised (z-scored) trial-average histogram per row, ordered by the time of peak activity. The first panel in a group of 9 shows histograms from raw calcium signals; each of the 8 subsequent panel shows trial-average histograms for the same cells, but following processing by each of the eight deconvolution methods.
[Q: How were the neurons chosen?]

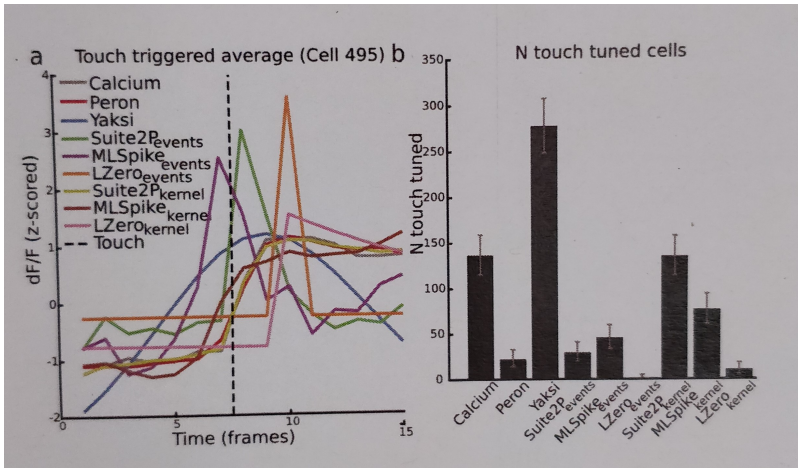

Figure 6: Touch-triggered neuron responses.
(a) Touch-triggered average activity from one neuron, across all deconvolution methods. The dotted line is the imaging frame in which the whisker touched the pole.
(b) Number of touch-tuned cells across deconvolution methods. A cell is classed as touch-tuned if its peak touch-triggered activity is significantly greater than shuffled data. Error bars are Jeffreys intervals

calcium time-series and more than one deconvolution method. In Figure 5e-h, we show how increasing the number of methods required to agree on a neuron's tuned status creates clear agreement between time-series processed with all methods, even if a particular method did not reach significance for that cell. Even requiring agreement between the raw calcium and just two other methods is enough to see tuning of many cells. The identification of unambiguously task-tuned cells could thus be achieved by triangulating the raw calcium with the output of multiple deconvolution methods.

In the pole detection task considered here, neurons tuned to pole contact are potentially crucial to understanding the sensory information used to make a decision. Touch onset is known to drive a subset of neurons to spike with short latency and low jitter (O'Connor et al., 2010; Hires et al., 2015). Detecting such rapid, precise responses in the slow kinetics of calcium imaging is challenging, suggesting discrete-deconvolution methods might be necessary to detect touch-tuned neurons. To test this, in each of the 9 sets of time-series we identify touch-tuned neurons by a significant peak in their tough-triggered activity (Fig **??**a). Figure 6b shows that, while all data-sets have touch-tuned neurons, the number of such neurons differs substantially between them. And rather than being essential, discrete deconvolution methods disagree strongly on touch-tuning, with MLSpike (events) finding 45 touch-tuned neurons and LZero (events) finding one. Thus our inferences of the coding of task-wide or specific sensory events crucially depends on our choice of calcium imaging time-series.

2.6 Inconsistent recovery of population correlation structure across deconvolution approaches

The high yield of neurons from calcium imaging is ideal for studying the dynamics and coding of neural populations [cite some examples here]. Many analyses of populations start from pairwise correlations between cells, whether as measures of a population's synchrony or joint activity, or as a basis for further analyses like clustering and dimension reduction [Cunningham]. We now show how our inferences of population correlation structure also

depend strongly on the choice of deconvolution method.

Figure 7a shows that the distributions of pairwise correlations qualitatively differ between the sets of time-series we derived from the same calcium imaging data. The considerably narrower distributions from the discrete deconvolution time-series compared to the others is expected, as these time-series are sparse. Nonetheless, there are qualitative differences within the sets of discrete and continuous time-series. Some distributions are approximately symmetric, with broad tails; some asymmetric with narrow tails; the correlation distribution from the Peron method time-series is the only one with a median below zero. These qualitative differences are not due to noisy estimates of the pairwise correlations: for all our sets of time-series the correlations computed on a sub-set of time-points in the session agree well with the correlations computed on the whole session (Figure 7b). Thus pairwise correlation estimates for each method are stable, but their distributions differ between methods.

Looking in detail at the full correlation matrix shows that even for methods with similar distributions, their agreement on correlation structure is poor. Some neuron pairs that appear correlated from time-series processed by one deconvolution method are uncorrelated when processed with another method (Figure 7c). Over the whole population, the correlation structure obtained from the raw calcium, Yaksi and Suite2p (kernel) time-series all closely agree, but nothing else does (Figure 7d): the correlation structure obtained from LZero agrees with nothing else; and the discrete deconvolution methods all generate dissimiliar correlation structures (Figure 7e). Our inferences about the extent and identity of correlations within the population will differ qualitatively depending on our choice of imaging time-series.

2.7 Deconvolution methods show the same population activity is both low and high dimensional

Dimensionality reduction techniques, like principal components analysis (PCA), allow researchers to make sense of large scale neuroscience data (Cunningham and Yu, 2014) [more here], by reducing the data from N neurons to $d < N$ dimensions. Key to such analyses is the choice of d, a choice guided by how much of the original data we can capture. To assess such inferences of population dimensionality, we apply PCA to our 9 sets of imaging time-series to estimate the dimensionality of the imaging data (which for PCA is the variance explained by each eigenvector of the data's covariance matrix).

Figure 8a plots for each deconvolution method the cumulative variance explained when increasing the number of retained dimensions. Most deconvolution methods qualitatively disagree with the raw calcium data-set on the relationship between dimensions and variance. This relationship is also inconsistent across deconvolution methods; indeed the discrete deconvolution methods result in the shallowest (MLSpike_{events}) and amongst the steepest (LZero_{events}) relationships between increasing dimensions and variance explained. The number of dimensions required to explain 80% of the variance in the data ranges from $d = 125$ (Peron) to $d = 1081$ (MLSpike_{events}), a jump from 8% to 70% of all possible dimensions (Fig 8b). Thus we could equally infer that the same L2/3 population activity is low dimensional (<10% dimensions required to explain 80% of the variance) or high-dimensional (¿50% of dimensions required) depending on our choice of imaging time-series.

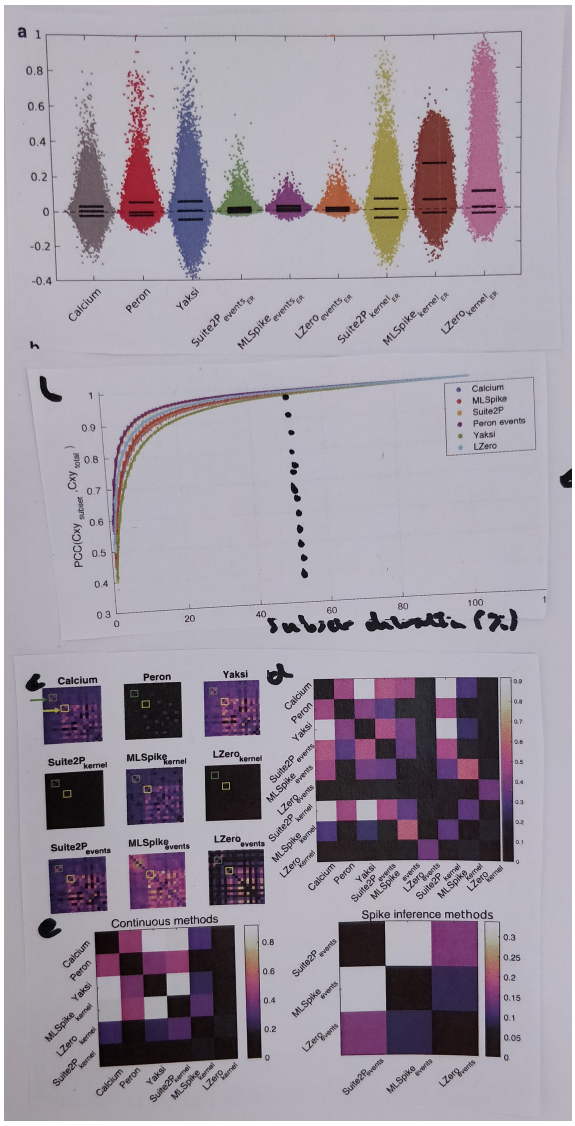

Figure 7: Effects of deconvolution on pairwise correlations between neurons.
(a) Distributions of pairwise correlations between all cells, for each deconvolution method (one dot per cell pair, x-axis jitter added for clarity). Solid black lines are 5th, 50th and 95th percentiles.
(b) Stability of correlation structure in the population. We quantify here the stability of the pairwise correlation estimates, by comparing the correlation matrix constructed on the full data (Cxy_{total}) to the same matrix constructed on a subset of the data (Cxy_{subset}). Each data-point is the correlation between Cxy_{total} and Cxy_{subset}; one line per deconvolution method.
(c) Examples of qualitatively differing correlation structure across methods. Each panel plots the pairwise correlations for the same 50 neurons on the same colour scale. As examples, we highlight two pairs of cells: one consistently correlated across different methods (green arrow and boxes); the other not (yellow arrow and boxes).
(d) Comparison of pairwise correlation matrices between deconvolution methods. Each square is the correlation between the full-data correlation matrix for that pair of methods. [REDO: compare correlation matrices using Spearmans to test rank ordering of pairs]
(e) as in (d), but split to show continuous methods (left) or discrete deconvolution methods (right). [Check all labels: clearly something inconsistent between (d) and (e): see Suite2pkernel results in both]

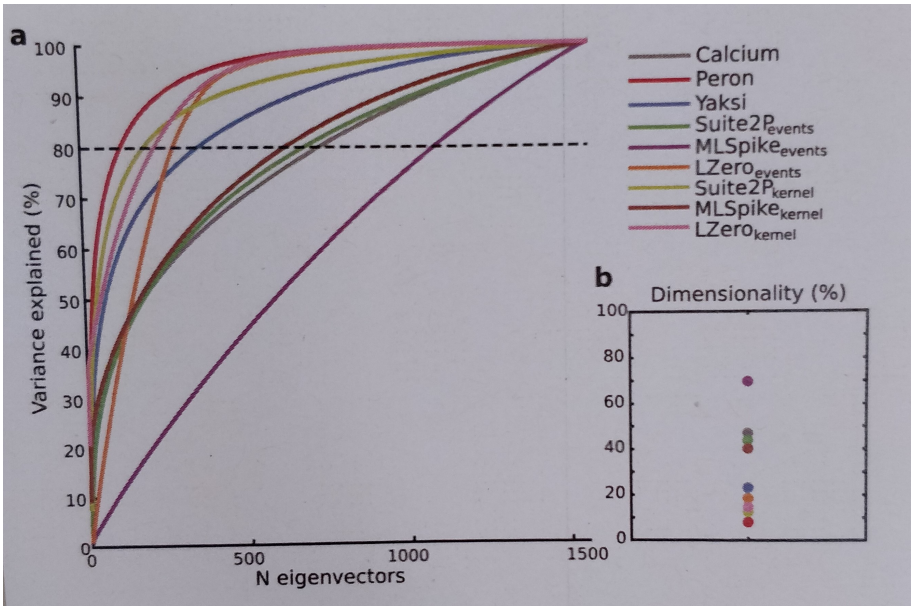

Figure 8: Dimensionality of population activity.
(a) Cumulative variance explained by each dimension of the datas covariance matrix, one line per deconvolution method. Dimensions are obtained from principal components analysis, and are ordered by decreasing contribution to the total variance explained. Dashed line is the 80% threshold used in panel (b).
(b) Proportion of dimensions required to explain 80% of the variance in the data.

3 Discussion

Imaging of somatic calcium is a remarkable tool for capturing the simultaneous activity of hundreds to thousands of neurons. But the time-series of each neuron's calcium fluorescence is inherently noisy and non-linearly related to its spiking. We sought here to address how our choice of corrections to these time-series – to use them raw, deconvolve them into continuous time-series, or deconvolve them into discrete events – affect the quality and reliability of the scientific inferences drawn.

Our results show the choice qualitatively changes the potential scientific inferences we draw about neural activity, coding, and correlation structure. We consistently observe that the analysis results differ sharply between the raw calcium and most, if not all, of the processed time-series. However, the deconvolved time-series also consistently disagreed with each other, even between methods of the same broad class (continuous or discrete time-series).

3.1 Accurate discrete deconvolution is possible, but sensitive

We find much that is encouraging. In fitting discrete deconvolution methods to ground-truth data, we found they can in principle accurately recover neural activity. A caveat here is that the choice of metric for evaluation and fitting of parameters is of critical importance. The widely-used 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. By contrast, the Error Rate metric (Deneux et al., 2016; Victor and Purpura, 1996) resulted in excellent recovery of ground-truth spike trains. Other recently developed methods for comparing spike-trains based on information theory (Theis et al.,

2016) or fuzzy set theory (Reynolds et al., 2017), may also be appropriate.

However, while good estimates of neural activity can be achieved with modern discrete deconvolution methods [cf Spikefinder, Pitcharriu 2018 JNS], the best parameters vary substantially between cells, and small changes in analysis parameters result in poor performance. This variation and sensitivity of parameters played out as widely-differing results between the three discrete deconvolution methods in analyses of neural activity, coding, and correlation structure.

3.2 Choosing parameters for deconvolution methods

A potential limitation of our study is that we use a single set of parameter values for each discrete deconvolution method applied to the population imaging data from barrel cortex. But then our situation is the same as that facing any experimentalist: in the absence of ground-truth, how do we set the parameters? Our solution here was to use the modal parameter values from ground-truth fitting, as these values are candidates for the most general solutions. We also felt these were a reasonable choice for the population imaging data from barrel cortex, given that the ground-truth recordings came from the same species (mouse) in the same layer (2/3) of a different bit of primary sensory cortex (V1).

Rather than use the most general parameters values, another solution would be to tune the parameters to obtain known gross statistics of the neural activity. This was the approach used by Peron and colleagues [cite] to obtain the denoised Peron time-series we included here. But as weve seen, this approach can lead to its own problems: for example, in the Peron time-series, it created a distributions of correlations that differed from any other set of time-series. Indeed, finding good parameter values may be an intractable problem, as it is possible each neuron requires individual fitting, to reflect the combination of its expression of fluorescent protein, and its particular non-linearity between voltage and calcium.

3.3 Ways forward

The simplest solution to the inconsistencies between different forms of time-series is to triangulate them, and take the consensus across their results. For example, our finding of a set of tuned neurons across multiple methods is strong evidence that neurons in L2/3 of barrel cortex are responsive across the stages of the decision task. Further examples of such triangulation in the literature are rare; Klaus and colleagues (Klaus et al., 2017) used two different pipelines to derive raw $\Delta F/F$ of individual neurons from one-photon fibre-optic recordings in the striatum, and replicated all analyses using the output of both pipelines. Our results encourage the further use of triangulation to create robust inference: obtaining the same result in the face of wide variation increases our belief in its reliability (Munaf and Davey Smith, 2018).

There are caveats to using triangulation. For single neuron analyses, triangulation inevitably comes at the price of reducing the yield of neurons to which we can confidently assign roles. A further problem for triangulation is how to combine more complex analyses, such a pairwise correlations; the alternative is to rely on qualitative comparisons.

Many studies use the raw calcium signal as the basis for all their analyses [cite], perhaps assuming this is the least biased approach. Our result show this is not so: the discrepancy between raw and deconvolved calcium on single neuron coding suggests an extraordinary range of possible results, from about half of all neurons tuned to the task down to less 5 percent. The qualitative conclusion – there is coding – is not satisfactory. Thus our

results should not be interpreted as a call to abandon deconvolution methods; rather they serve to delimit how we can interpret their outputs.

Instead, we need deconvolution solved: as sensors with faster kinetics (though fundamentally limited by kinetics of calcium release itself) and higher signal-to-noise ratios are developed [cite GCAMP7; red-shifted], so the accuracy and robustness of de-noising and deconvolution should improve; and as the neuron yield continues to increase [Ahrens], so the potential for insights from inferred spikes or spike-driven events grows. Developing further advanced deconvolution algorithms will harness these advances, but are potentially always limited by the lack of ground-truth to fit their parameters. Our results may provide impetus for a different direction of research, focussing on how we can get consensus among the output of different algorithms, and thus provide robust scientific inferences about neural populations.

4 Methods

[FINISH all methods: add all details from deconvolution debugging doc; fix all issues flagged below; turn notes into prose.]

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. The data-set contains twenty one recordings from nine cells.

Spike train metrics

Pearson correlation coefficient was computed between the ground truth and inferred spike trains following convolution of both with a gaussian kernel (61 samples wide, 1.02 seconds). [WHAT were used as inferred spike trains]

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),

[write this in full, defining sensitivity and precision]

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

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

[Firing rate: how computed for true and inferred?]

For normalised estimation of errors in firing rate, we computed (estimated FR - true FR / true FR)

Parameter fitting

For each method the best parameters for each cell were determined by brute force search over an appropriate range. [Give the ranges. How was "appropriate" defined?]

The modal best parameters, as determined using Error Rate on downsampled data, were then fixed for the population imaging data analysis. These were: [Complete with parameter values]

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 [??? how was the downsampling done then? By taking every Nth frame from the interpolated data? Or averaging over every approx 8.5 frames of the 60Hz signal?].

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 per frame which were converted into mean event rates (Hz) per cell [FIX THESE descriptions]. The 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 [missing object in sentence: average of what?] and subtracting this [what is "this"?] from the activity/fluorescence trace. Event rate was then computed in Hz.

[what does this all mean? That each data-point for the background noise was an average over 4 adjacent frames, and the standard deviation of the noise was computed from those data-points? Why smooth the noise? And what segment of the data was treated as "background"? (i.e. how many data-points)? Or does this mean that the whole Ca2+ trace for each cell was smoothed using a 4-frame average (shifting 1 frame?), and the SD of that smoothed signal was used as an estimate of background noise?]

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 as 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 of activity (seven data samples) was extracted before and after the frame closest to touch (15 samples total); taking the mean of these gave the average touch

response for the cell. 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.

[Figure legend said: (Mann-Whitney U test, Benjamini Hochberg corrected); which is it? And what is α here?]

[unclear what the pairwise tests were between; the mean activity at the peak time, and a N-length vector of mean activity as the same time obtained from N shuffled datasets?]

[Shuffled how many times? Shuffled how?].

4.4 Pairwise correlations

Pairwise correlations were calculated for all pairs of neurons in Matlab (corrcoef) at the data sampling rate (7Hz). [State what forms of time-series were correlated - from debugging doc]

Correlations between correlation matrices (Fig. **??**) were computed between the unique pairwise correlations from each method (i.e. `CXY(find(triu(CXY)))`). [Add Spearman's rank here]

Stability of correlation. For each deconvolution method, we computed the pairwise correlation matrix using the entire sessions data, as above. We then sampled a subset of time-points (1%-100%) of the full dataset at random without replacement and compute its matrix of pairwise correlations. We compute the similarity between the [Total] and [Subset] matrices using Pearsons correlation coefficient. [Finish this])

4.5 Dimensionality

To determine the dimensionality of each dataset we performed eigendecomposition of the covariance matrix of each dataset. [Computed as per the pairwise correlations above?] 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.)

[Explain: the basic method, and its free parameters]
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 re-

cover 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).

[Explain: the basic method, and its free parameters]

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).

[Explain: the basic method, and its free parameters]

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).

[Explain: the basic method]

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%.

[Explain: the basic method, and its free parameters]

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$_{kernel}$.

[Explain: what are the parameters of the kernel]

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.

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.

Andreas Klaus, Gabriela J Martins, Vitor B Paixao, Pengcheng Zhou, Liam Paninski, and Rui M Costa. The spatiotemporal organization of the striatum encodes action space. *Neuron*, 95:1171–1180.e7, Aug 2017. ISSN 1097-4199. doi: 10.1016/j.neuron.2017.08.015.

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.

Marcus R Munaf and George Davey Smith. Robust research needs many lines of evidence. *Nature*, 553:399–401, January 2018. ISSN 1476-4687. doi: 10.1038/d41586-018-01023-3.

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.

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.