Robustness of spike deconvolution for neuronal calcium imaging

Marius Pachitariu\textsuperscript{1,2,3}, Carsen Stringer\textsuperscript{1,4} and Kenneth D. Harris\textsuperscript{2,3}

\textsuperscript{1}HHMI Janelia, Ashburn, 20147, USA.
\textsuperscript{2}UCL Institute of Neurology, London WC1N 3BG, United Kingdom.
\textsuperscript{3}UCL Department of Neuroscience, Physiology, and Pharmacology, London WC1E 6BT, United Kingdom.
\textsuperscript{4}Gatsby Computational Neuroscience Unit, London W1T 4JG, United Kingdom.

DOI: 10.1523/JNEUROSCI.3339-17.2018

Received: 25 November 2017
Revised: 23 June 2018
Accepted: 27 June 2018
Published: 6 August 2018

Author contributions: M.P. and K.D.H. designed research; M.P. and C.S. performed research; M.P. contributed unpublished reagents/analytic tools; M.P. and C.S. analyzed data; M.P., C.S., and K.D.H. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

Correspondence to marius10p@gmail.com

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.3339-17.2018

Alerts: Sign up at www.jneurosci.org/cgi/alerts to receive customized email alerts when the fully formatted version of this article is published.
Robustness of spike deconvolution for neuronal calcium imaging

Marius Pachitariu1,2,3*, Carsen Stringer1,4, and Kenneth D. Harris 2,3

1 --- HHMI Janelia, Ashburn, 20147, USA.
2 --- UCL Institute of Neurology, London WC1N 3BG, United Kingdom.
3 --- UCL Department of Neuroscience, Physiology, and Pharmacology, London WC1E 6BT, United Kingdom.
4 --- Gatsby Computational Neuroscience Unit, London W1T 4JG, United Kingdom.
* --- Correspondence to marius10p@gmail.com

Abstract
Calcium imaging is a powerful method to record the activity of neural populations in many species, but inferring spike times from calcium signals is a challenging problem. We compared multiple approaches using multiple datasets with ground truth electrophysiology, and found that simple non-negative deconvolution (NND) outperformed all other algorithms on out-of-sample test data. We introduce a novel benchmark applicable to recordings without electrophysiological ground truth, based on the correlation of responses to two stimulus repeats, and used this to show that unconstrained NND also outperformed the other algorithms when run on “zoomed out” datasets of ~10,000 cell recordings from the visual cortex of mice of either sex. Finally, we show that NND-based methods match the performance of a supervised method based on convolutional neural networks, while avoiding some of the biases of such methods, and at much faster running times. We therefore recommend that spikes be inferred from calcium traces using simple NND, due to its simplicity, efficiency and accuracy.

Significance statement
The experimental method that currently allows for recordings of the largest numbers of cells simultaneously is two-photon calcium imaging. However, use of this powerful method requires that neuronal firing times be inferred correctly from the large resulting datasets. Previous studies have claimed that complex supervised learning algorithms outperform simple deconvolution methods at this task. Unfortunately, these studies suffered from several problems and biases. When we repeated the analysis, using the same data and correcting these problems, we found that simpler spike inference methods perform better. Even more importantly, we found that supervised learning methods can introduce artifactual structure into spike trains, that can in turn lead to erroneous scientific conclusions. Of the algorithms we evaluated, we found that an extremely simple method performed best in all circumstances tested, was much faster to run, and was insensitive to parameter choices, making incorrect scientific conclusions much less likely.

Introduction
Two-photon calcium imaging can be used to monitor the activity of populations of up to 10,000 neurons (Pachitariu et al., 2016; Stringer et al., 2018). Nevertheless, calcium-sensitive
fluorescence signals are an indirect readout of cellular activity. Therefore, accurate and well-calibrated data processing methods will be required to make optimal use of this activity (Andilla and Hamprecht, 2014; Deneux et al., 2016; Friedrich et al., 2017; Jewell and Witten, 2017; Kazemipour et al., 2016; Reynolds et al., 2015; Sebastian et al., 2017; Theis et al., 2016; Vogelstein et al., 2010). One important problem is developing methods for spike detection: inferring the times of action potentials from the fluorescence traces. The earliest such methods rely on spike deconvolution algorithms, which infer a spike train under the assumption that the fluorescence trace represents an approximate convolution of the underlying spike train with the cell’s calcium response (Vogelstein et al., 2010). This is often a good approximation (Chen et al., 2013), though situations exist when it breaks down. More complex spike deconvolution algorithms take these extreme cases into account (Deneux et al., 2016).

Recently, a new approach to spike detection, based on supervised learning, has been claimed to outperform several existing deconvolution algorithms (Theis et al., 2016). Supervised algorithms learn to solve the spike detection problem by training on “ground truth” data where spike times are also measured electrophysiologically. In principle, such methods should give the most accurate results: however, they may generalize poorly to “out-of-sample” data, i.e. recordings made under different conditions to the available training data.

Since the supervised approach was first introduced, ground truth data was released in a public competition called “spikefinder” (Berens et al., 2017) which has allowed the comparison of several old and new algorithms. Three of the better performing algorithms (called “Elephant”, “Purgatorio”, “conv6”) employ supervised methods based on convolutional neural networks, and appear to slightly outperform unsupervised methods based on non-negative deconvolution (NND; “oopsi”, “Suite2p”). However, this performance may be due to specific design features of the spikefinder challenge, rather than true improvements in spike deconvolution quality. First, the spikefinder benchmarks are run on in-sample data, and may thus not reflect generalization performance to new recordings. Second, multiple metrics for the similarity of decoded and actual spike trains are possible; because supervised methods can be trained to optimize the particular metric used, they will have an advantage over unsupervised methods, unless the latter are also optimized for the particular quality metric used.

We show here that non-negative deconvolution (NND) – with very simple parameter settings, and using the fast OASIS implementation (Friedrich et al., 2017) – outperforms supervised algorithms, when it is: 1) evaluated on out-of-sample data, and 2) adapted to the performance metric of the spikefinder challenge. In addition, we find that NND is highly robust to assumptions on the assumed shape of the calcium response to single spikes (henceforth called a “kernel”), such that a simple decaying exponential kernel performs better than more biologically accurate kernels that include a rising time segment, and even performs better than kernels estimated directly from ground truth data. Moreover, large changes to the timescale of the exponential kernels did not affect performance significantly, and optimizing these timescales for each cell actually hurts performance. Finally, we propose a new benchmark that can be used without electrophysiological ground truth, and show that simple NND again outperforms other algorithms on this benchmark. The new benchmark can be used to compare algorithms on a wide range of realistic in vivo datasets, and we provide a code repository to facilitate these comparisons, accompanied by a release of the large-scale test datasets used here (data: Pachitariu et al., 2018, code: https://github.com/MouseLand/pachitariu-et-al-2018a).
Materials and methods

*Imaging in visual cortex*

All experimental procedures were conducted according to the UK Animals Scientific Procedures Act (1986). Experiments were performed at University College London under personal and project licenses released by the Home Office following appropriate ethics review.

The experimental methods were similar to those described elsewhere (Dipoppa et al., 2016). Briefly, surgeries were performed in adult mice of either sex (P35–P125) in a stereotaxic frame and under isoflurane anesthesia (5% for induction, 0.5-1% during the surgery). During the surgery we implanted a head-plate for later head-fixation, made a craniotomy with a cranial window implant for optical access, and, on relevant experiments, performed injections of the GCaMP6m virus with a beveled micropipette using a Nanoject II injector (Drummond Scientific Company, Broomall, PA 1) attached to a stereotaxic micromanipulator. Viruses were acquired from University of Pennsylvania Viral Vector Core. Injections of 50-200 nl virus (1-3 x10^{12} GC/ml) were targeted to monocular V1, 2.1-3.3 mm laterally and 3.5-4.0mm posteriorly from Bregma and at a depth of L2/3 (200-400 μm). Some mice were transgenic and expressed tdTomato in certain cell classes. However, we did not use that information here.

**NND model**

NND models infer the most likely spike train \( s(t) \), given the fluorescence timecourse \( F(t) \), and a response kernel \( k \). Models based on deconvolution define a cost function of the form

\[
C(s) = \|F - s \ast k\|^2 + \lambda \cdot L(s)
\]

such that \( s(t) \geq 0 \), for all \( t \).

Here, \( s \ast k \) describes a temporal convolution of a positive timecourse \( s \) and the kernel \( k \), and \( L(s) \) describes a penalty function on the inferred spike trains.

We tested a suite of three unsupervised spike detection methods. The first is an approximate optimization algorithm where \( L(s) = \|s\|_0 \) is the L0 norm, i.e. the number of non-zero entries in \( s \) (code available at github.com/cortex-lab/Suite2P). The L0 penalty enforces the constraint that the inferred spike trains should be very sparse, because neurons fire rarely. The second method has \( L(s) = \|s\|_1 \), the L1 norm, and chooses the kernel from a parametrized class of functions (Friedrich et al., 2017). The third unsupervised model is unconstrained non-negative deconvolution (NND), with \( L(s) = 0 \). For our initial analysis, we chose the sparsity penalty \( \lambda \) for both the L0 and L1 methods in such a way as to output spike trains with similar levels of sparsity: only ~5% of the deconvolved samples were non-zero, for data sampled at 100Hz. We then varied the sparsity penalties, as well as the parameters that define the kernel \( k \), to understand how they influence performance.

**Supervised learning models**

We compared performance of these NND algorithms against two supervised methods: the method of Theis et al. (2016), and also a publicly available convolutional neural network algorithm (code from https://github.com/PTRRupprecht/Spikefinder-Elephant/tree/master/elephant). In both cases we used default parameter settings.
We also tested an additional, novel L0 deconvolution algorithm, obtained by developing a novel optimization procedure for a standard spike generation model. This optimization is an extension of a well-known algorithm called “matching pursuit” (Mallat and Zhang, 1993; Smith and Lewicki, 2006) which can be fast enough to apply to the large datasets considered here (Krstulovic and Gribonval, 2006). Briefly, the matching pursuit algorithm identifies putative spikes by their similarity to the calcium kernel. It then subtracts the kernel scaled with an appropriate factor (the “spike amplitude”) from the location of the identified spikes. On successive iterations, more putative spike locations are identified greedily, and their activity subtracted off. This continues until no new spikes can be introduced, because they do not have large enough amplitudes to explain a significant portion of the variance.

This basic matching pursuit algorithm is limited by local minima, because the greedy procedure cannot always resolve nearby spikes that have overlapping calcium activity. There have been various approaches proposed to address this problem, for example orthogonal matching pursuit, which re-estimates the magnitudes of the atoms in the support set at each iteration (Pati et al., 1993). An “atom” is defined as any potential element that can be picked up by matching pursuit, in our case potential spike location, while the “support set” refers specifically to the atoms that have been picked up to some iteration. One disadvantage of OMP and similar approaches (Soussen et al., 2013) do not allow atoms in the support set to be removed during the optimization. Other recent generalizations do, for example OMP with replacement or CoSaMP (Jain et al., 2011; Needell and Tropp, 2009).

Nevertheless, these generalizations are often computationally intensive, and do not explicitly address the problem we encountered most often: early on during greedy extraction, an atom is introduced to account for an entire burst of spikes, and is thus introduced at the average time of the spikes in the burst. To avoid local optima while replacing this single atom with two smaller ones would require adjusting the relative timing of the original atom once others are introduced nearby. However, such a step requires first a large sacrifice in explained variance (dropping a valuable atom): only if an algorithm knows it can immediately compensate by reintroducing the atom at a nearby time, will this sacrifice be chosen by an optimization scheme.

We developed such an optimization scheme, by adding a step at every iteration where we allow existing spikes to change their location and/or magnitude to better account for the calcium trace. The optimal changes can be calculated exactly and efficiently at all small temporal offsets from existing atoms, by using the pre-computed filtered trace. This step allows “old” spikes to adjust their locations and activity in the context of “new” spikes, thus reaching more accurate solutions.

We note that another algorithm for solving a similar problem has recently been proposed (Jewell and Witten, 2017). This algorithm does not impose positivity, is restricted to exponential kernels, and is slower than our approach, but it does obtain an exact solution for their respective problem, which our greedy approach does not.

**Kernel choices**

For the NND, L0- and L1- regularized models, we used kernels that were either exponential, or a difference of exponentials to model the fluorescence rise time. The timescale of the exponential was chosen as one of three possibilities: fast (0.7s), medium (1.25s) or slow (2s), according to the calcium sensor used. Timescales for the GCaMP6 sensors were assigned.
according to their version. For other sensors, timescales were assigned based on the literature (OGB: medium, GCaMP 5k: fast, jRCaMP1a: slow, jRGECO1a: fast). To model fluorescence rise time, we used difference of exponentials, with the decaying exponential having the same timescale as before. The exponential accounting for the rising phase had a timescale that was varied as a fraction (0 to 0.5) of the timescale of the decaying exponential (Figure 2e,f, Figure 3h).

Simulations

To simulate fluorescence traces, we first generated the underlying spike trains. The number of spikes \( s(t) \) in bin \( t \) of 10ms was simulated from a Poisson process, \( s(t) \sim \text{Poisson}(r(t)) \), where the firing rate \( r(t) \) is constructed from

\[
r(t) = r_0 + \alpha \cdot b(t)
\]

Here, \( r_0 \) is the baseline firing rate, \( \alpha \) is a scaling factor that determines the ratio of bursting to tonic firing. \( b(t) \) is 1 during bursting and 0 otherwise. Bursting periods occurred randomly at least twice per minute and lasted for 250 ms.

We then simulated the calcium fluorescence by first filtering the spike train \( s(t) \) with an exponential kernel (timescale 1 s) and adding independent Gaussian noise at each timepoint to simulate shot noise (see below). Finally, to simulate temporally-correlated noise such as movement and neuropil contamination we added Gaussian-filtered white noise (half-width of 500 ms), to simulate other sources of noise, like movement and neuropil contamination.

\[
F(t) = \langle s * K \rangle(t) + \beta n_{\text{independent}}(t) + \gamma n_{\text{correlated}}(t)
\]

The kernel was an exponential filter with a timescale of 1 second. To produce multiple datasets (Figure 2g,h), we randomly varied simulation parameters with a uniform distribution: mean firing rate (0.05 - 1 Hz), burstiness \( \alpha \) (0 – 0.5), shot noise amplitude \( \beta \) (0 – 0.02) and correlated noise amplitude \( \gamma \) (0 – 0.015). The parameters were fixed within each simulated dataset of 200 cells.

In figure 3b, we also use these simulations. For three fixed firing rates and with \( \alpha = 0 \) we generated three sets of 200 spike trains. From each of these sets, we generated 20 datasets with different combinations of shot noise and correlated noise, but keeping the underlying spike trains fixed.

Performance metrics

The main performance metric we used was \( s_{\text{corr}} \): the Pearson correlation of a ground truth spike train \( s_{\text{GT}} \) and an estimated spike train \( s \), both binned with 40ms resolution (Figures 1,2,3). This was the same metric as used by the spikefinder challenge (Berens et al., 2017; Theis et al., 2016). We also used a variant of the van Rossum spike train distance, where both \( s \) and \( s_{\text{GT}} \) are smoothed with a Gaussian before taking their correlation (Rossum, 2001; Schreiber et al., 2003). The width of the Gaussian was varied systematically (Figure 1i).

We also used the area under the ROC curve (AUC) as a metric (Vogelstein et al., 2010; Theis et al., 2016) (Figure 2cd). The AUC can be computed as the fraction of times that the deconvolution result for a bin containing a spike is larger than the deconvolution result for a bin containing no spike. When both results are exactly equal to each other, we count this as a "half correct" value, or 0.5.
Finally, we introduce a new metric $\sigma_{\text{stim}}$ based on the stimulus response reliability. Specifically, $\sigma_{\text{stim}}$ is the Spearman correlation of trial-averaged tuning curves, obtained from two independent halves of a recording, where the stimuli were presented in random order on each half. Defining the trial-averaged responses of neuron $n$ to stimulus $k$ as $r_1(k, n)$ on the first repeat and $r_2(k, n)$ on the second repeat, the correlation $\sigma_{\text{stim}}$ of stimulus responses is simply:

$$\sigma_{\text{stim}} = \text{Spearman} (r_1(k, n), r_2(k, n))_k$$

where the $k$ subscript indicates that the correlation is taken over stimuli $k$.

The correlation $\sigma_{\text{stim}}$ is not trivially 1 because the calcium signals are not identical across repeats. This trial-to-trial variability will reflect both genuine differences in neural spike trains, and the noisy transformation from electrical spiking to fluorescence. The task of a spike deconvolution algorithm is to invert this transformation, thereby denoising the recovered signals. The more successful an algorithm is, the more noise it will have removed, and the higher its $\sigma_{\text{stim}}$ value should be, even though these values cannot reach 1 due to neuronal variability.

We chose the rank correlation (Spearman) rather than the Pearson correlation to avoid potential biases introduced by nonlinearities in the spike deconvolution process. For example, if the result of deconvolution to both repeats were transformed through the same nonlinearity, the Pearson correlation could be artificially increased, while the Spearman correlation will remain the same.

Convolutional neural networks

Of the SPIKEFINDER challenge winners, one team has so far made their code publicly available (https://github.com/PTRRupprecht/Spikefinder-Elephant/tree/master/elephant), so we used their CNN configuration, called “Elephant”. This method fits >100,000 free variables, which were learned using default parameter settings.

Results

For the first set of benchmarks, we considered two main classes of datasets with simultaneously-recorded ground truth electrophysiology, which we will refer to in short as “GENIE” and “SPIKEFINDER”. The GENIE collection consists of five datasets recorded by the GENIE project (Akerboom et al., 2012; Chen et al., 2013; Dana et al., 2016; GENIE project, 2015), that have been used in the original descriptions of the GCaMP calcium sensors and their red variants and are available on CRCNS.org (GENIE project, 2015). The SPIKEFINDER collection consists of the five datasets analyzed in Theis et al, 2016, made publically available as part of the “spikefinder” challenge (spikefinder.codeneuro.org). Surprisingly, the two state-of-the-art algorithms, “oopsi” and “stm” (Theis et al., 2016; Vogelstein et al., 2010), have different performance on the two GT datasets, with oopsi winning on GENIE and stm winning on SPIKEFINDER.

To measure deconvolution performance, requires metrics to compare the algorithms’ output to ground truth spike trains. A standard metric, termed $\sigma_{\text{st}}$ throughout this paper, consists of the Pearson correlation coefficient of the true and inferred spike trains, after binning both in a preset window size (i.e. 40ms in the spikefinder challenge, also the standard value used here). However, this metric may not be well suited for spike trains, which are very sparse quantities.
For example, if an inferred spike is offset by just one bin from a GT spike, it will be counted as a complete miss by the correlation metric, similarly to a temporal mismatch of several bins. Supervised algorithms, if assessed on in-sample test data, will have automatic protection from this effect as they are directly trained to minimize the correlation metric; however, this behavior can lead to worse out-of-sample performance as shown below. Unsupervised algorithms can be adapted to perform well by the correlation metric by smoothing their output temporally, which reduces the effect of temporal mismatches between true spikes and deconvolved spikes. In addition, a temporal offset might be required for some datasets (for example due to hardware synchronization issues), which can be learned automatically by the supervised methods, but has to be inferred post-hoc in the unsupervised algorithms.

A short segment of fluorescence from a cell recorded with ground truth is shown in Figure 1ab, together with the reconstructions obtained with three unsupervised models (the quantity $s \ast k$ in equation 1). All three models track the large fluorescence changes, but some of the smallest changes are only tracked by the unconstrained model (NND), which is relatively less constrained than the L0- or L1-penalized methods. For example, the L1-penalized model overly penalizes single spike events in some cases, reducing their amplitude relative to other single or multi spike events (Figure 1c). Nonetheless, for this cell, all three deconvolution methods returned roughly similar spike trains, which correlated well with the known ground truth electrophysiology (Figure 1bc).

Simple non-negative deconvolution outperforms the state-of-the-art results

Many calcium deconvolution algorithms have recently been described (Andilla and Hamprecht, 2014; Deneux et al., 2016; Friedrich et al., 2017; Jewell and Witten, 2017; Kazemipour et al., 2016; Reynolds et al., 2017; Theis et al., 2016; Vogelstein et al., 2010) some of which have provided their code publicly. However, little effort has been made to compare the performance of these algorithms to each other, with the notable exception of (Theis et al., 2016), who concluded that supervised algorithms, trained on available ground truth data, perform better than the more routinely used unsupervised algorithms.

We investigated this claim on the same datasets used by (Theis et al., 2016), which were since made available publicly in the “spikefinder” challenge. We indeed found that the supervised algorithms performed better than the L0 and L1-penalized algorithms, when evaluated by the correlation metric (Figure 1de, “original”). However, we found that unsupervised algorithms became superior after very simple modifications.

First, we chose an appropriate timelag parameter for a subset of datasets where the timing of the ground truth spikes was not perfectly synchronized with the fluorescence (Figure 1de, “+time lag”). This parameter was chosen to maximize the correlation with the ground truth spikes, for each of the ten datasets separately. The inferred timelag was 0 for all GENIE datasets, and ranged between -2 and 3 for SPIKEFINDER datasets.

Second, we applied smoothing to the deconvolved traces (but not to the ground truth spike trains), which reduced the effect of the correlation metric (Figure 1de, “+smoothing”). The smoothing was performed with a Gaussian-shaped kernel with a standard deviation of two samples for all GENIE datasets, and 8 samples for all SPIKEFINDER datasets. (These values were empirically found to perform well.) Note that all datasets have been upsampled at 100Hz, and are benchmarked at 25Hz, following the standards of the original spikefinder benchmarks.
Finally, we did not allow the algorithms to estimate the best fit calcium kernels, or the kernel’s decay timescale, as we found that all methods failed to recover appropriate parameters. Instead, we fixed the timescales of the calcium kernel to be approximately the measured values from the literature (Akerboom et al., 2012; Chen et al., 2013; Dana et al., 2016) (Figure 1de, “+fixed taus”). For simplicity, we divided all sensors into a fast, a medium and a slow category, assigning them corresponding timescales of 0.75, 1.25 and 2 seconds. As we show below, the precise values for these timescales were not critical.

These improvements, together, increased the benchmark performance for nearly all cells, and surpassed both the supervised and unsupervised “state-of-the-art” approaches submitted on the website spikefinder by their developers (stm and oopsi). Furthermore, the best performing model in the benchmark was unconstrained NND, with the L0- and L1-based methods slightly lagging behind (Figure 1de). Across datasets, unconstrained NND reliably performed as well as or better than L0- and L1-based methods (Figure 1f), with only small differences between models. Similarly, the differences were small at the level of single cells (Figure 1g), suggesting that it was not the case that some cells were better deconvolved by some models.

To test the performance of the algorithms at different timescales, we varied the binning size from 10ms to 1.28s. Again, we found the three top algorithms to be virtually indistinguishable (Figure 1h). To test whether the distance metric used to evaluate spike train accuracy might influence the result, we turned to a variant of the van Rossum distance (Rossum, 2001; Schreiber et al., 2003) which smoothed the spike trains before computing correlations. Over a wide range of smoothing widths, the correlation between ground truth and estimated spike trains followed a similar pattern, with all three unsupervised algorithms performing very similarly (Figure 1i).

**Robustness of non-negative deconvolution**

These results suggest that the choice of regularization method (if any) does not have major impact on the performance of unsupervised deconvolution. We tested this more systematically, by varying the regularization parameter $\lambda$. When $\lambda = 0$, both the L0 and L1-based algorithms are equivalent to the unconstrained approach. We found that for both algorithms, performance was best when $\lambda = 0$, corresponding to unconstrained NND (Figure 2ab). We confirmed this result with another metric: the area under the ROC curve (AUC) (Figure 2cd), which can be computed as the fraction of times that the deconvolution result for a bin containing a spike is larger than the deconvolution result for a bin containing no spike.

These results suggest that the simple NND method consistently matches or outperforms L0- and L1- methods. However, the GENIE and SPIKEFINDER datasets contain several tens of cells each, which may not be a large enough sample to test how generally this result holds. To extend the range of our analysis, we generated simulated datasets where we varied for each neuron its firing rate, burstiness, shot noise amplitude and correlated noise amplitude. Each simulated dataset was generated under a random combination of these four parameters. For this analysis, we chose the value of the regularization parameter for L0- and L1- methods that gave optimal test set performance: these methods can therefore choose a value of 0 if that performs best, and thus by definition cannot perform worse than simple NND. Nevertheless, we observed only minimal improvements due to regularization, and only in a small number of
scenarios (Figure 2g). The regularizer was most often set to near zero, but sometimes high values were chosen by the L1-based method (Figure 2f). We conclude that regularization provides little if any benefit in both simulated data and real data with available ground truth.

The unsupervised algorithms were robust not just to the value of the regularizer, but also to large changes in the shapes of the calcium kernels. Lengthening or shortening the assumed timescale of the calcium indicators by a factor of 2 did not significantly affect performance (Figure 3ab). Furthermore, adding another component to the kernel (a “rising” timescale) did not improve performance (Figure 3cd). Finally, performance was not improved by using the “ground truth” kernel, obtained directly by regressing the fluorescence onto the ground truth spikes (Figure 3cd, dotted lines).

*Ground truth*-free benchmarks using stimulus responses

NND therefore exceeds the performance of supervised methods, and at least matches the performance of L0- and L1-regularized based methods in both the SPIKEFINDER and GENIE datasets, as well as a range of simulations. Still, however, these conditions might be different from the conditions in many experiments. For example, the SPIKEFINDER and GENIE datasets were recorded under anaesthesia, with an invasive electrode attached to the cell. To assess the performance of spike deconvolution methods in realistic recordings, we developed a novel benchmark that does not require ground truth electrophysiology (Figure 4a).

Our new benchmark is based on the intuition that a more accurate spike detection method will yield more similar responses to repeated stimuli. Specifically, this benchmark is the Spearman correlation $\sigma_{\text{stim}}$ of deconvolved responses, trial-averaged on two separate halves of the data:

$$\sigma_{\text{stim}} = \text{corr}(\hat{s}_{\text{repeat}1}, \hat{s}_{\text{repeat}2}),$$

where $\hat{s}_{\text{repeat}1}$ and $\hat{s}_{\text{repeat}2}$ are the binned, trial-averaged responses to $N$ different stimuli. The stimuli must be presented in randomized order on each of the two repeats, so that the only common information contained in the neural responses is related to the spiking activity. $\sigma_{\text{stim}}$ computes a trade-off between signal and noise variance, with $1 - \sigma_{\text{stim}}$ representing the proportion of noise variance (reflecting a sum of biological and measurement noise; see Methods). Since the signal can only originate in the true spiking $s$, $\sigma_{\text{stim}}$ captures the ability of the deconvolution to reconstruct the true spiking. Deconvolution can fail to capture the stimulus variance in $s$ in one of two ways: 1) failure to distinguish spikes from noise, or 2) failure to correctly invert the forward calcium model, for example assigning spikes to incorrect stimulus bins. Even with perfect spike detection, the actual spike train would also differ between repeats; thus, $\sigma_{\text{stim}}$ cannot reach 1, but will have a maximal value, given by the Spearman correlation of the underlying spike trains. This maximum will be achieved when the deconvolved spike trains exactly match the original spike trains at the time resolution used to quantify stimulus responses.

This new metric gave nearly identical results to the standard metric on simulated data. To show this, we first generated three sets of spike trains of different mean firing rates, in response to multiple repeats of simulated stimuli. From each set of spike trains, we generated multiple fluorescence traces by varying the noise parameters of each simulation (shot noise and correlated noise), and then ran multiple deconvolution algorithms on each trace, using multiple parameter settings. For each spike train, the deconvolution performance as assessed by $\sigma_{\text{stim}}$ was nearly perfectly correlated with performance assessed by the standard measure $\sigma_{\text{GT}}$ (the correlation between deconvolved spike trains and the true underlying spike trains) (Figure 4b).

Nevertheless, this relation could not be extrapolated between spike trains: for the same value of...
\( \sigma_{GT} \), different spike trains could have different values of \( \sigma_{stim} \). Thus, \( \sigma_{stim} \) can be used to compare the relative deconvolution performance of multiple algorithms on the same spike train, but not to estimate on an absolute scale how similar the deconvolved spikes are to the true, unknown spikes.

Having established \( \sigma_{stim} \) as a useful metric for comparing algorithms, we applied it to some 2-photon recordings of \( \sim 10,000 \) cells from primary visual cortex of awake mice (Stringer et al., 2018), with full field drifting grating stimuli of varying orientation, spatial and temporal frequency. A raster plot of these responses (deconvolved by NND) is shown in Figure 4c. After averaging over 8 repeats, the responses of many single cells were generally reliable, but still contained some noise (Figure 4d). Nearly all cells had positive Spearman correlations \( \sigma_{stim} \) of their tuning curves between the two stimulus halves (Figure 4e). Taking \( \sigma_{stim} \) as a measure of deconvolution performance (higher is better), we repeated the types of analyses from Figures 2 and 3, for 6 datasets recorded from 4 mice. Again we found that simple NND outperformed the regularized versions, as well as the supervised STM algorithm, which in turn only slightly outperformed the raw fluorescence (Figure 4f). We also again found that the NND methods (with or without regularization) are robust to the kernel timescale up to a factor of 2 (Figure 4g), and that the AR(2) kernel does not help performance (Figure 4h). This new metric therefore reinforces the results obtained on a more limited dataset with simultaneous electrophysiology. In addition, we note that this method can be easily applied to other recordings, with as few as two repeats of the same stimulus, allowing users to benchmark multiple approaches on their own data. We provide a code repository and example data to facilitate these comparisons (code: https://github.com/MouseLand/pachitariu-et-al-2018a, data: Pachitariu et al., 2018).

**Comparisons with convolutional neural networks**

We have shown that unsupervised methods, when adapted to the correlation-based benchmark, outperform the supervised approach described in Theis et al., 2016, on the same ground truth data used there. However, it remains possible that other supervised methods might perform better still. In the recent “spikefinder” challenge, convolutional neural networks (CNNs) appear to outperform all other methods. However, to successfully apply CNNs on standard recordings without ground truth, it must be shown that CNNs generalize to datasets not included in the training set. We suspected this generalization might be imperfect, because CNNs, like all supervised methods, can overfit to the specific statistics of the ground truth datasets.

To evaluate the generalization performance of CNNs, we trained the “Elephant” method on nine of the ten available datasets, testing it on the tenth. We found that under this testing protocol, the performance of supervised CNNs and unsupervised NND was nearly identical, both on the GENIE datasets (CNN: 0.44, NND: 0.45) and on the SPIKEFINDER datasets (CNN: 0.59, NND: 0.60), although there were slight variations in performance from neuron to neuron (Figure 5a).

However, the performance of the CNNs was worse in the ground-truth free benchmarks we just introduced above (CNN: 0.234 vs NND: 0.262), with some variation from neuron to neuron (Figure 5b).

Another disadvantage of complex CNNs is speed: even using a high-performance GPU (GTX 1080), the method is two orders of magnitude slower than all unsupervised methods we tested, which can run efficiently on standard CPUs (Figure 5c).

*Supervised methods impose biases on out-of-sample data*
The similarity in performance we found might appear at odds with the results of the spikefinder challenge, where the CNN methods outperformed unsupervised approaches by ~10% (CNN: ~0.46 vs NND+L0: ~0.43). However, for that challenge the CNN was tested on within-class data, thus being able to take advantage of the particular statistics of spiking and fluorescence for each recording. One such statistic is the autocorrelogram structure of the spike trains, which was far from Poisson, reflecting either that stimuli were presented during some of the recordings, or the structure of spontaneous activity in the recorded neurons. The autocorrelogram structure can be used by supervised approaches to perform better spike prediction. However, this strategy is undesirable, because it will enforce the properties of the training data on new data, potentially leading to an erroneous scientific conclusion that all recorded neurons share the same temporal dynamics as the neurons used to train the algorithm.

To demonstrate this transfer of constraints between training and test data, we simulated spike trains with Poisson statistics (flat auto-correlograms), and generated fluorescence traces from them with a calcium decay timescale of 1 second. The deconvolved spike trains using CNNs had a large, spurious auto-correlation at short timelags, which was much less pronounced in methods based on OASIS (NND and NND+L1) and absent using the L0-based method (Figure 5d). The “black-box” nature of CNN algorithms raises a further concern that other features of the training data may be erroneously imposed on new data, in ways that are unknown to the user.

Discussion

We conclude that the performance of simple NND-based deconvolution algorithms matches or exceeds all tested alternatives. Adding L0/L1 regularization to the NND model did not improve its performance, perhaps because non-negativity is already a strong regularizer by itself. NND was robust to changes in kernel timescale and shape, with values taken from the literature providing close to optimal performance. Automated identification of kernel parameters appeared to be counterproductive, resulting in mismatched parameters that impaired performance. While supervised methods gave apparently superior performance in previously reported benchmarks, this reflected their ability to optimize particular evaluation metrics, and compensate for phenomena such as synchronization lags within single example datasets. When tested out-of-sample, against unsupervised methods with appropriate compensatory mechanisms, we found their performance to be inferior.

The potential pitfalls of using more complex methods go beyond poor out-of-sample generalization. These methods may also introduce biases, usually due to their implicit or explicit priors. For example, L0- and L1- based deconvolution may introduce too much sparsity into the spike trains. Supervised models do not typically impose explicit priors, but learn implicit priors from the statistics of the training data. They then impose these priors on new test data, as we have shown here (Figure 5d), even if the new data has different statistics, potentially leading to erroneous scientific conclusions.

We therefore recommend simple, unconstrained NND, with fixed calcium decay timescale. OASIS (Friedrich et al., 2017) provides a very efficient algorithm for performing this deconvolution. In Suite2p, the calcium processing pipeline that we maintain, we provide wrappers for the OASIS toolbox, and additionally include the L0-based deconvolution code, which may provide advantages for some cases, such as avoidance of auto-correlation bias.
It remains possible that future methods will be able to significantly outperform simple NND; however, any such improvements need to be balanced against the simplicity and interpretability of the NND approach.

Acknowledgements

We thank Charu Reddy for assistance with surgeries and Michael Krumin for assistance with the two-photon microscopes. This work was supported by the Wellcome Trust (95668, 95669, 108726), and the Simons Foundation (325512). CS was funded by a four-year Gatsby Foundation PhD studentship.

References

Akerboom, J., Chen, T.-W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderón, N.C., Esposti, F., Borghuis, B.G., Sun, X.R., et al. (2012). Optimization of a GCaMP calcium indicator for neural activity imaging. J. Neurosci. 32, 13819–13840.

Andilla, F.D., and Hamprecht, F.A. (2014). Sparse space-time deconvolution for calcium image analysis. In Advances in Neural Information Processing Systems, pp. 64–72.

Berens, P., Freeman, J., Deneux, T., Chenkov, N., McColgan, T., Speiser, A., Macke, J.H., Turaga, S., Mineault, P., Rupprecht, P., et al. (2017). Community-based benchmarking improves spike inference from two-photon calcium imaging data. BioRxiv 177956.

Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al. (2013). Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499, 295–300.

Dana, H., Mohar, B., Sun, Y., Narayan, S., Gordus, A., Hasseman, J.P., Tsegaye, G., Holt, G.T., Hu, A., Walpita, D., et al. (2016). Sensitive red protein calcium indicators for imaging neural activity. Elife 5, e12727.

Deneux, T., Kaszas, A., Szalay, G., Katona, G., Lakner, T., Grinvald, A., Rózsa, B., and Vanzetta, I. (2016). Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. Nat. Commun. 7.

Dipoppa, M., Ranson, A., Krumin, M., Pachitariu, M., Carandini, M., and Harris, K.D. (2016). Vision and locomotion shape the interactions between neuron types in mouse visual cortex. BioRxiv 058396.

Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. PLOS Comput. Biol. 13, e1005423.

GENIE project (2015). Simultaneous imaging and loose-seal cell-attached electrical recordings from neurons expressing a variety of genetically encoded calcium indicators. Janelia Farm Campus HHMI Svob. Karel Contact.

Jain, P., Tewari, A., and Dhillon, I.S. (2011). Orthogonal matching pursuit with replacement. In Advances in Neural Information Processing Systems, pp. 1215–1223.
Jewell, S., and Witten, D. (2017). Exact Spike Train Inference Via $\ell_0$ Optimization. ArXiv Prepr. ArXiv170308644.

Kazempour, A., Liu, J., Solarana, K., Nagode, D.A., Kanold, P.O., Wu, M., and Babadi, B. (2016). Fast and stable signal deconvolution via compressible state-space models. BioRxiv 092643.

Krstulovic, S., and Gribonval, R. (2006). MPTK: Matching pursuit made tractable. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference On, (IEEE), pp. III–III.

Mallat, S.G., and Zhang, Z. (1993). Matching Pursuits With Time-Frequency Dictionaries. IEEE Trans. Signal Process. 41, 3397–3415.

Needell, D., and Tropp, J.A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. 26, 301–321.

Pachitariu, M., Stringer, C., Schröder, S., Dipoppa, M., Rossi, L.F., Carandini, M., and Harris, K.D. (2016). Suite2p: beyond 10,000 neurons with standard two-photon microscopy. BioRxiv 061507.

Pachitariu, M., Stringer, C., and Harris, K.D. (2018). Recordings of 10k neurons in V1 during drifting gratings (figshare).

Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference On, (IEEE), pp. 40–44.

Reynolds, S., Onativia, J., Copeland, C.S., Schultz, S.R., and Dragotti, P.L. (2015). Spike detection using FRI methods and protein calcium sensors: performance analysis and comparisons. In Sampling Theory and Applications (SampTA), 2015 International Conference On, (IEEE), pp. 533–537.

Rossum, M. van (2001). A novel spike distance. Neural Comput. 13, 751–763.

Schreiber, S., Fellous, J.-M., Whitmer, D., Tiesinga, P., and Sejnowski, T.J. (2003). A new correlation-based measure of spike timing reliability. Neurocomputing 52, 925–931.

Sebastian, J., Kumar, M.G., Sreekar, Y., Rikhye, R.V., Sur, M., and Murthy, H.A. (2017). GDspike: An Accurate Spike Estimation Algorithm from Noisy Calcium Fluorescence Signals.

Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. Nature 439, 978.

Soussen, C., Gribonval, R., Idier, J., and Herzet, C. (2013). Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares. IEEE Trans. Inf. Theory 59, 3158–3174.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C.B., Carandini, M., and Harris, K.D. (2018). Spontaneous behaviors drive multidimensional, brain-wide population activity. BioRxiv 306019.
Theis, L., Berens, P., Froudarakis, E., Reimer, J., Roman-Roson, M., Baden, T., Euler, T., Tolias, A.S., and Bethge, M. (2016). Benchmarking spike rate inference in population calcium imaging. Neuron 90, 471–482.

Vogelstein, J.T., Packer, A.M., Machado, T.A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. J. Neurophysiol. 104, 3691–3704.

Figures

(a) Example fluorescence recording of a neuron recorded by the GENIE project. The model trace reconstructions are shown in color (blue is NN, red is NN + L0, yellow is NN + L1).

(b) Simultaneous ground truth electrophysiology for the neuron shown in a.

(c) Deconvolved traces using three NND models.

(d) Correlation $\sigma_{GT}$ between deconvolved and ground truth spike trains, with various processing stages included, averaged over cells separately for datasets from the GENIE project (GENIE project, 2015) and for SPIKEFINDER datasets (Theis et al., 2016). The average values for stm and oopsi are taken from the spikefinder challenge (http://spikefinder.codeneuro.rg).

(f) Per dataset average improvement of L0 and L1-based methods over NND.

(g) Distribution across cells of the improvement of L0 and L1-based methods over NND.

(h) Mean $\sigma_{GT}$ across all datasets as a function of bin size.

(i) Same as (h) but instead of binning, spike trains were smoothed with a Gaussian of standard deviation $\sigma$. 

Figure 1. Deconvolution performance with ground-truth electrophysiology. 

- For dataset average improvement of L0 and L1-based methods over NND.
- For distribution across cells of the improvement of L0 and L1-based methods over NND.
- For mean $\sigma_{GT}$ across all datasets as a function of bin size. 
- For same as (h) but instead of binning, spike trains were smoothed with a Gaussian of standard deviation $\sigma$. 

14
Figure 2. No performance advantage from L0/L1 regularization.  

Effect of regularization parameter $\lambda$ on mean correlation $\sigma_{GT}$ or the area under the curve (AUC) between deconvolved spike trains and ground truth electrophysiology.

(ab) The penalty on sparsity was varied, for both L0 and L1-penalized models. The best value of 0 corresponds to unconstrained NND.

(cd) Same as (ab) for a different performance metric (area under the ROC curve, AUC).

(ef) Performance of L0 and L1 methods vs NND, in simulations obtained by randomly varying 4 parameters: firing rate, burstiness, shot noise and correlated noise. The optimal regularizer for L0 and L1 methods was chosen by cross-validation, and is shown in (f).
Figure 3. Robustness to kernel parameters.

(ab) The kernel timescales were varied; this had little effect and did not significantly improve performance over values taken from the literature.

(cd) A second kernel timescale was introduced and varied, to model the rise time of the fluorescence following a spike. The kernels were defined as a difference of exponentials, which defines a subclass of AR(2) kernels. The AR(2) version of OASIS was used for the L1-penalized method. Performance with ground truth-derived kernels is also shown as dotted lines (see Methods). The optimal rising timescale was 0, corresponding to a simple exponential kernel.
Figure 4. Benchmarking without ground truth electrophysiology.

(a) Schematic of the processes leading up to the deconvolved spike trains.

(b) The new performance metric $\sigma_{\text{sim}}$ plotted against $\sigma_{\text{GT}}$, for three example simulated spike trains. Dots of a single color represent different instantiations of the same spike train, with different added noise and different deconvolution algorithms. The horizontal lines indicate the maximum possible.

(c) Example deconvolved neural responses (NND algorithm) from ~10,000 simultaneously recorded neurons, sorted by their preferred stimulus. The stimuli shown were drifting gratings with one of 8 directions, one of 3 spatial frequencies and one of four temporal frequencies.

(d) Correlation between trial-averaged neural responses to half of the stimulus presentations, versus the other half of the stimulus presentations.

(e) Distribution of correlation coefficients between repeat halves (Spearman). The mean of these coefficients is used in f-h as a benchmark of deconvolution performance (higher is better).
(f) Deconvolution performance as the penalty on sparseness was increased, for L0 and L1-penalized models, as well as for raw non-deconvolved data, and for the stm model from Theis et al., 2016.

(g) Deconvolution performance as the timescales are increased or decreased by a fractional amount.

(h) Performance with AR(2) kernels that include a non-instantaneous rise with varying timescales.
Figure 5. Non-negative deconvolution matches the performance of supervised, convolutional neural networks.

(a) Spike detection performance ($\sigma_{GT}$) for all cells with ground truth electrophysiology. The means for all GENIE and all Theis et al. 2015 datasets are shown as circles. Each CNN was trained on all but one of the 10 datasets, and tested on the remaining dataset.

(b) Stimulus response correlation ($\sigma_{stim}$) of CNN trained on all 10 datasets, and tested on our data using the repeat similarity benchmark (Figure 3).

(c) Runtimes of CNN on a high-end GPU (GTX 1080), compared with runtimes for NND implemented by OASIS (with or without L1 regularization) and the L0 method on a standard CPU (Core i7).

(d) Autocorrelograms of deconvolved spike trains from simulations with Poisson ground truth statistics. The CNN approach heavily biases the statistics of the inferred spike trains. The L0 and L1 method were run with same parameters as in Figure 2 ($\lambda = 10$ and 100 respectively).
Calcium fluorescence

Ground truth (GT) electrophysiology.

Non-negative deconvolution (NND) \( \sigma_{GT} = 0.71 \)

NND + L0 constraints \( \sigma_{GT} = 0.71 \)

NND + L1 constraints \( \sigma_{GT} = 0.70 \)

Rescuing GT performance

No improvement over NND

Varying bin size

Smoother metric
No advantage of L0/L1 regularizer ($\sigma_{GT}$)

(a) GENIE
- NND
- NND + L0
- NND + L1

(b) SPIKEFINDER
- NND
- NND + L0
- NND + L1

No advantage of L0/L1 regularizer (AUC)

(c) GENIE
- Raw dF
- mean AUC

(d) SPIKEFINDER
- Raw dF
- mean AUC

Small advantage of L0/L1 regularizer in simulations

(e) mean $\sigma_{GT}$ vs mean $\sigma_{GT}$ (NND) for NND, L0, and L1.

(f) Scatter plot of regularizer $\lambda$ (cross-validated) vs mean $\sigma_{GT}$ (NND) for NND, L0, and L1.
Robustness to kernel timescale

No advantage of AR(2) kernel
**a** Benchmarking without GT electrophysiology

- Stimulus repeats 1-8
- Stimulus repeats 9-16
- Spikes
- Fluorescence
- Deconvolved spikes
- High correlation = good deconvolution

**b** Simulations
- Firing rate (Hz)
- $\sigma_{\text{stim}}$
- $\sigma_{\text{GT}}$

**c** ~10,000 cells imaged in visual cortex

- Neurons sorted by stimulus preference
- Orientation (deg) 0 360
- Spatial freq (cpd) 4 8 16 $\times 10^{-2}$
- Temporal freq (Hz) .5 1 2 4
- 1 minute

**d** Responses of one cell
- $\sigma_{\text{stim}} = 0.68$
- Repeats 9-16 (avg)
- Repeats 1-8 (avg)

**e** Distribution (1 recording)
- Number of cells
- $\sigma_{\text{stim}}$

**f** No advantage of L0/L1 penalty

- Mean $\sigma_{\text{stim}}$
- Penalty $\lambda$
- NND
- NND + L0
- STM
- Raw dF
- NND + L1

**g** Robustness to kernel timescale

- Mean $\sigma_{\text{stim}}$
- Timescale multiplier

**h** No advantage of AR(2) kernel

- Mean $\sigma_{\text{stim}}$
- Rising timescale (fraction)
Comparisons with the state-of-the-art CNN "Elephant"

(a) 

(b) 

(c) Runtime for 1000 cells

(d) Autocorrelograms (Poisson simulation)