Deconvolution of Sustained Neural Activity From Large-Scale Calcium Imaging Data

Younes Farouj, *Member, IEEE*, Fikret Işık Karahanoğlu, and Dimitri Van De Ville, *Senior Member, IEEE*

***Abstract*— Recent technological advances in light-sheet microscopy make it possible to perform whole-brain functional imaging at the cellular level with the use of Ca^{2+} indicators. The outstanding spatial extent and resolution of this type of data open unique opportunities for understanding the complex organization of neuronal circuits across the brain. However, the analysis of this data remains challenging because the observed variations in fluorescence are, in fact, noisy indirect measures of the neuronal activity. Moreover, measuring over large field-of-view negatively impact temporal resolution and signal-to-noise ratio, which further impedes conventional spike inference. Here we argue that meaningful information can be extracted from large-scale functional imaging data by deconvolving with the calcium response and by modeling moments of sustained neuronal activity instead of individual spikes. Specifically, we characterize the calcium response by a linear system of which the inverse is a differential operator. This operator is then included in a regularization term promoting sparsity of activity transients through generalized total variation. Our results illustrate the numerical performance of the algorithm on simulated signals; i.e., we show the firing rate phase transition at which our model outperforms spike inference. Finally, we apply the proposed algorithm to experimental data from zebrafish larvæ. In particular, we show that, when applied to a specific group of neurons, the algorithm retrieves neural activation that matches the locomotor behavior unknown to the method.**

***Index Terms*— Temporal deconvolution, light-sheet microscopy, calcium imaging, generalized total variation, ℓ_1-minimization.**

I. Introduction

SIMULTANEOUS recordings of activity from large neural populations are within reach today thanks to technical developments of whole-brain functional imaging based on light-sheet microscopy [1], [2]. This technology enables *in vivo* volumetric measurement of Ca^{2+} concentrations through genetically encoded fluorescent markers. While developments of new calcium indicators are still a field of active research [3]–[6], the measured signals are intrinsically noisy and their dynamics are slower than the underlying action potential (AP). This issue has prompted a broad interest in methods for finding evidence when AP firing occurs. The classical and predominant approaches are derivative thresholding [7], [8] and matching procedures [9], [10]. The first one employs a simple threshold on the normalized fluorescence signal to obtain event onset times. The second one aims at finding a fluorescence trace that matches the waveform of AP firing. Both these techniques have two main drawbacks. First, they incorporate precise priors on the absolute signal amplitude of the fluorescence signal. This amplitude is, however, changing depending on the spatial location and also on the calcium indicator that is deployed and its photo-bleaching properties. Second, they do not allow the retrieval of rapid successive events. Linear deconvolution techniques were then used to tackle this issue [11]. By incorporating an assumption about the positivity of the spikes, Vogelstein *et al.* [12] developed an efficient fast algorithm for nonnegative deconvolution of the spike train. Machine learning triggered some works that are based on supervised and unsupervised learning [13]–[15]. However, supervised approaches require the availability of large datasets with electrophysiological ground truth for the training phase. Moreover, in spite of the sophistication of these methods, unsupervised *deconvolution* methods were recently shown to give the best results in practice [16]. More recently, techniques that are based on finite rate of innovation (FRI) [17] were introduced to retrieve spikes by fitting them to the AP waveform without priors on the amplitude of the fluorescence signal [18]. Finally, the increasing interest in causal investigation of neural circuits [19] motivated the development of state-of-the-art online deconvolution methods [20]. In most of this literature, spike-retrieval methods are evaluated using datasets that are acquired at very fast temporal sampling rates. However, this prerequisite is no longer fulfilled in large field-of-view setups for whole-brain or whole-animal imaging. In such a scenario, complex temporal patterns of firing cannot be resolved because the Ca^{2+} responses start to heavily overlap. This has motivated recent work to deviate from classical deconvolution and estimate the underlying average activity or firing rates accounting for unobserved spikes within predefined time intervals [21].

In this work, we approach this problem from a different viewpoint by performing regularized deconvolution; i.e., we assume that the deconvolved signal consists of moments of sustained activation. Mathematically, the piecewise-constant nature of a signal can be expressed by total variation, which can be combined with the inverse filter for deconvolution [22].

Manuscript received July 9, 2019; revised September 11, 2019; accepted September 15, 2019. Date of publication September 23, 2019; date of current version April 1, 2020. This work was supported by Carl ZEISS AG under the Research-IDEAS Initiative of ZEISS and EPFL. *(Corresponding author: Younes Farouj.)*

Y. Farouj and D. Van De Ville are with the Institute of Bioengineering, École Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland, and also with the Department of Radiology and Medical Informatics, University of Geneva, 1211 Geneva, Switzerland (e-mail: younes.farouj@epfl.ch).

F. I. Karahanoğlu is with MGH/HST Athinoula A. Center for Biomedical Imaging, Harvard Medical School, Boston, MA 02215 USA.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2942765

The inspiration of this model comes from the field of human neuroimaging, in particular, functional magnetic resonance imaging (fMRI). This modality observes a hemodynamic proxy of neural activity that is several orders of magnitude slower than the underlying spiking activity. It has been shown that spatial patterns of transitions between periods of sustained activation are meaningful building blocks of evoked and spontaneous activity [23], [24]. In the context of calcium imaging, the model of sustained activity is motivated by three main observations:

- Overlapping responses are mainly occurring during bursting phenomena. Bursts might encode important information in their duration or frequency [25], features that are easily accessible if burst periods are modeled as activity blocks instead of a sequence of individual spikes.
- Neurons in many brain areas act as integrators [26]; they show persistent and accumulated neural activity; e.g., neurons driving sustained oculomotor (eye fixation) function [27].
- Whole-brain imaging aims mainly at mapping neural correlates of function and behavior in terms of distributed patterns of activity. Such a systems-level perspective can highly benefit from a meaningful simplification of spiking activity.

The proposed framework does not incorporate any prior information about the timing or duration of periods of sustained activation. Given the large field-of-view, correlation analysis of temporal patterns that contain alternations between sustained activation and baseline can allow to retrieve distributed spatial patterns of coupled or anti-coupled activity.

We start off in Sect. II by introducing the signal model and the linear-system characterization of the calcium response. Then, in Sect. III, we highlight the regularized deconvolution approach which involves three signals: the innovation signal that encodes the transients, the deconvolved signal, and the denoised signal consist with the solution. We give insights about the numerical guarantees of the proposed algorithm by analyzing its performance on simulated signals. In Sect. IV, we demonstrate the relevance of sustained activity modeling on state-of-the-art light-sheet microscopy data of zebrafish larvæ.

II. Signal Model

We start with some definitions and notations that we are going to use throughout the rest of the paper. Then, we describe how the fluorescence trace is linked to moments of transient activity and we model this link by a differential operator. For clarity of exposition, we start with a description in continuous settings. Discretization issues will be addressed in a separated paragraph. Next, we will often consider real-valued one-dimensional signals. Continuous-time signals are denoted with parentheses; e.g., $f(t)$, $t \in \mathbb{R}$. Discrete-time sampled signals are denoted with brackets; e.g., $f_s[k]$, $k \in \mathbb{Z}$. Continuous-time operators are denoted in calligraphic letters (e.g., $\mathcal{H}$) and their discrete counterparts in normal font (e.g., H).

TABLE I
DYNAMICS FOR THE FAST, MEDIUM AND SLOW VERSIONS OF GCaMP6 IN RESPONSE TO SINGLE AP FIRINGS

Ca^{2+} indicator	GCaMP6f	GCaMP6m	GCaMP6s
t_{peak} (seconds)	0.15	0.25	0.4
$t_{1/2}$ (seconds)	0.4	0.8	1

A. From AP Firing to the Fluorescence Trace

We assume that we are measuring calcium concentration on a single neuron. The observed signal is a rescaled and noisy version of the original fluorescence trace f. The signal rescaling is due to the imaging system and the received number of photons, while random noise and nuisance components can be caused by many factors such as low frequency fluctuations, residual errors from motion correction and normalization. In the sequel, we will consider that the observed signal is normalized with respect to a baseline fluorescence level in a way that only activity-related variations are observed. For the sake of simplicity, we also assume that the rescaling coefficient is included in the definition of f. As illustrated in Figure 1, this indirect measure is obtained from the actual signal describing AP firings, s, convolved with the calcium impulse response function (CIRF), h:

$$f(t) = (s * h)(t). \tag{1}$$

Here, s is a spike train consisting of J spikes occurring at times t_j

$$s(t) = \sum_{j=1}^{J} s_j \delta(t - t_j), \quad \text{where } s_j \in \mathbb{R}+. \tag{2}$$

Notice that no restriction is made on the inter-spike interval (ISI); i.e., two consecutive spikes can be arbitrary close. On the other hand, the function h models the calcium dynamics, which vary depending on the calcium indicator that is used. Different generations of the GCaMP calcium indicators [6] lead to signals with different activity transients; i.e., peak amplitude is reached after $t_{\text{peak}} \approx 150 - 400$ ms, before decaying back to the baseline with half decay time in the range $t_{1/2} \approx 400 - 1000$ ms. Table I reports some of the typical experimental parameter settings that we are going to consider.

B. Characterization of the CIRF Operator

The continuous-time CIRF is typically characterized by a double exponential that models the rise and decay of activity as follows:

$$h(t) = e^{-at} - e^{-bt}, \tag{3}$$

where a and b are two nonnegative scalars [18]. More precisely, a and b are given in Hz as the inverse of the rise time and the decay time; $a = \frac{1}{t_{\text{peak}}}$ and $b = \frac{t_{1/2}}{\log(2)}$. To describe the action of h as a differential operator, we first consider its Fourier domain description

$$\widehat{h}(\omega) = \frac{b - a}{(j\omega + a)(j\omega + b)}. \tag{4}$$

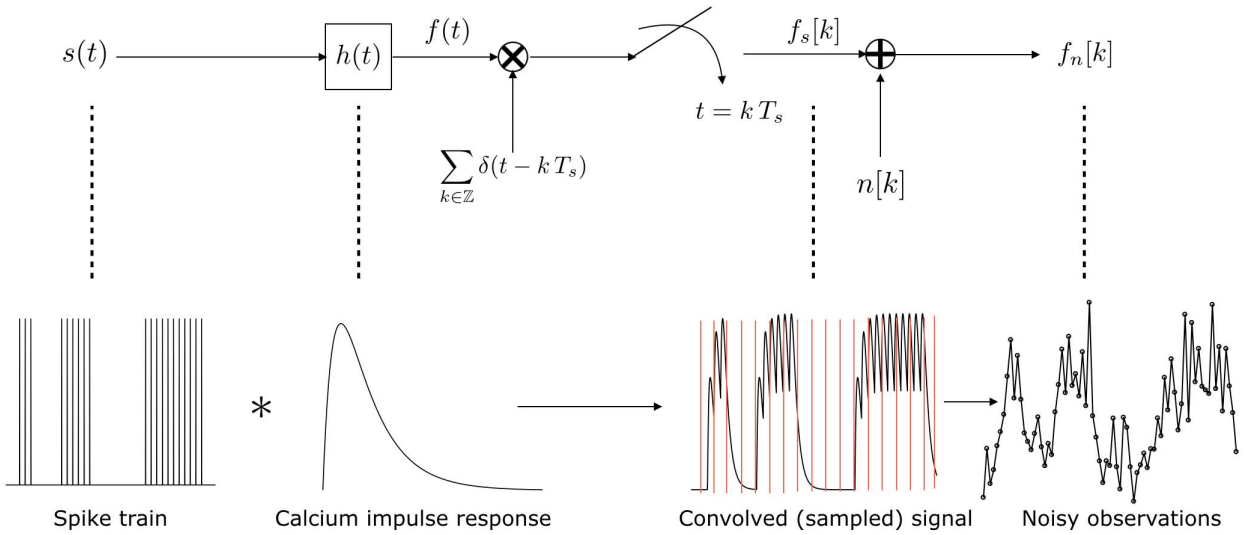

Fig. 1. Signal model: The observed recording f_n is a noisy version of the sampled fluorescence trace f_s; i,e, evaluated at the discrete imaging time grid (red lines).

Now, we denote by $\mathcal{D}$ and $\mathcal{I}$, the continuous derivative and identity operators, respectively. The characterization (4) happens to be the Fourier domain expression of the following operators:

$$\mathcal{H} = (b - a)\mathcal{D}_a^{-1}\mathcal{D}_b^{-1}, \tag{5}$$

where $\mathcal{D}_a = (\mathcal{D} + a\mathcal{I})^{-1}$ and $\mathcal{D}_b = (\mathcal{D} + b\mathcal{I})^{-1}$. The operator description in (5) is crucial for the present work as it will be used to construct a temporal parsimony-promoting regularization that enables recovering the fluorescence trace and an estimation of the underlying neural events.

C. Discrete Settings

We denote $T_s > 0$ as the sampling period, given in seconds. In practice, the fluorescence trace is observed for a finite duration $T > 0$. At each time step $k = 1, \ldots, K$, $K = T/T_s$, the observed noisy signal verifies $k = 1, \ldots, K$, $K > 0$:

$$f_n[k] = f_s[k] + n[k], \quad n[k] \sim \mathcal{N}(0, \sigma^2), \tag{6}$$

where f_s is a sampled version of f and the noise follows a Gaussian distribution with zero mean and standard deviation $\sigma > 0$.

D. Limits of Spike Inference

In any practical setting, spike inference from calcium recordings operates on sampled data. However, the underlying AP firing pattern and induced fluorescence signal are happening in continuous time, consistent with the previously introduced signal model. There are two ways to approach this issue. The first approach seeks at finding the spikes locations and amplitudes off-grid, that is, in continuous time domain. Parametric approaches such as the FRI framework developed in [18] are variants of Prony's method [28]. This requires a pre-estimation of the number of spikes J which is not always possible in practice. Additionally, these methods are sensitive to noise [29]. Another idea is to minimize the total variation over the space of Radon measures; i.e., the sum of Dirac measures. The obtained minimization problem is referred to as Beurling Lasso [30] and can be solved, for example, using continuous basis pursuit [31]. It is, however, known that this type of modeling requires a minimum separation distance between spikes that is above the theoretical limit[1] [32], while also still fails at low signal-to-noise ratio regimes [33]. Another minimum requirement for off-grid estimation is that, theoretically, at least two measurement are needed *per* spike in order to recover the two unknowns that are location and amplitude. Moreover, because of the decay of the convolution kernel, these samples should have a minimum proximity to the spike location [34]. The second approach is to give up on the exact locations and seek for the nearest spikes on the grid. This is the idea behind ℓ_1-*deconvolution*, in discrete domain, often used by practitioners [35]. An important point that we want to convey in this work is that when successive spikes occur in a short time window (i.e., small ISI with respect to the sampling period), inferring single spikes fails, both on the grid *and* off the grid. At some spiking rate, it becomes beneficial to consider a sustained activity model, as we propose here. The experimental results will show that there is a phase transition depending on the interplay between calcium dynamics and sampling.

III. Method

A. Sparsity-Promoting Temporal Deconvolution

From a temporal point-of-view, the main purpose is to invert the effects of the CIRF on the signal. Because of the presence of noise, this should be done using a regularity prior on the signal in order to make the problem well-posed. We exploit the expected sparsity of the activity transients to

[1]This limit is defined as $1/f_c$, where f_c is the cutoff frequency of the convolution kernel.

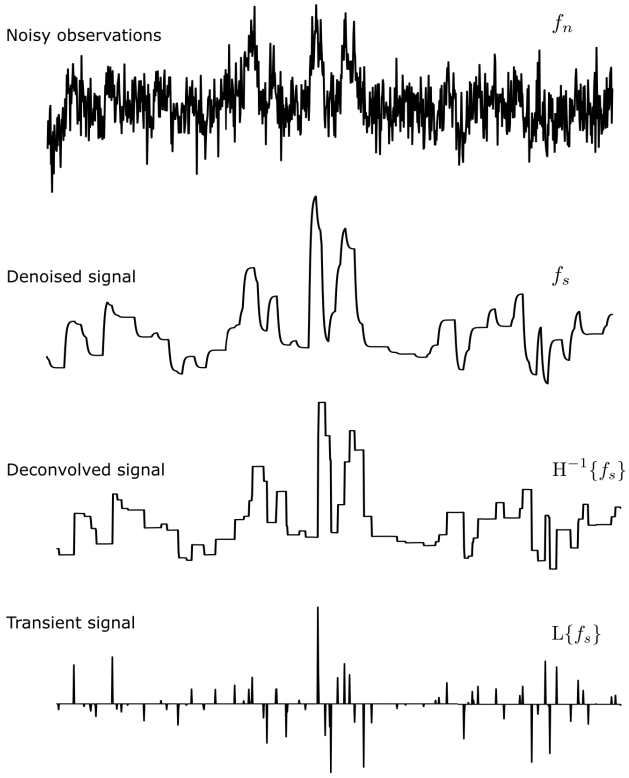

Fig. 2. In the *block model*, we are interested only in moments of transition to rapid firing activity that can be resolved at slow imaging rates. We exploit the sparsity of the transient signal; i.e., the derivative of a sustained version of s, to drive the deconvolution. The sparsifying operator is then considered to be the composition of the inverse of the CIRF operator and temporal derivation: $\mathrm{L} = \mathrm{D}\mathrm{H}^{-1}$.

drive the deconvolution within a variational framework. More precisely, we construct an operator that sparsifies the sampled fluorescence trace f_s on which we use the ℓ_1-*norm* to penalize a least squares optimization problem.

1) Construction of the Sparsifying Operator: We are interested in constructing an operator $\mathcal{L}$, whose discrete version L sparsifies f_s, that is $\mathrm{L}\, f_s$ is a train of Dirac pulses. First note that by the definition of $\mathcal{H}$, we have:

$$s = \mathcal{H}^{-1} f, \quad \text{with } \mathcal{H}^{-1} = \frac{1}{b-a}\mathcal{D}_a\mathcal{D}_b. \tag{7}$$

In a noiseless ideal sampling scenario, the relation (7) should hold in discrete domain. Therefore, the first option is to consider that $\mathcal{L}$ is simply given as $\mathcal{H}^{-1}$ since s is sparse *per se*. We refer to this model as the *spike model*. The second option is to consider the problem of detecting on/off moments of rapid firing and link the obtained signal to the firing rate (see Figure 2). Here we exploit the sparsity of the transient signal that is the derivative of s. This will be refereed to as the *block model*. Here, $\mathcal{L}$ consists of a combination of H^{-1} with a temporal derivative. To summarize, we have:

$$\mathcal{L} = \begin{cases} \mathcal{H}^{-1}, & \text{if } \textit{spike model}, \\ \mathcal{D}\mathcal{H}^{-1}, & \text{if } \textit{block model}. \end{cases} \tag{8}$$

In both cases we obtain a differential operator. In the *block model*, s is considered to be a block-wise signal with discontinuities. Therefore, its derivative should be understood in the distribution sense [36]. Notice that depending on the chosen model, the null-space of $\mathcal{L}$ consists of polynomials of degree 2 or 3, respectively. For the discrete operator L, whose construction is described in the appendix, the null-space becomes empty when using zero-boundary conditions.

Algorithm 1 FISTA Iterations for Temporal Deconvolution

Input: f_n, μ, $\tilde{\sigma}$, $f_s^{(0)} = 0$, $r_1 = 1$
Output: Estimate $\tilde{f_s}$
Initialize $z^{(0)} = \mathrm{L}\, f_s^{(0)}$, $v^{(1)} = z^{(0)}$, $\lambda^1 = \tilde{\sigma}$, $i \longleftarrow 1$
repeat
1: $z^{(i)} = \frac{1}{\lambda^i \mu}\mathrm{L}\, f_n + \left(I - \frac{1}{\mu}\mathrm{L}\,\mathrm{L}^{+}\right) v^{(i)}$
2: $z^{(i)} = \mathcal{P}(z^{(i)})$
3: $r_{i+1} = \frac{1+\sqrt{1+4r_i^2}}{2}$
4: $v^{(i+1)} = z^{(i)} + \frac{r_i - 1}{r_{i+1}}\left(z^{(i)} - z^{(i-1)}\right)$
5: $f_s^i = y - \mathrm{L}^{+} z^{(i+1)}$
6: $\lambda^{i+1} = \frac{K\tilde{\sigma}}{\frac{1}{2}\sum_{k=1}^{K}\left(f_n[k] - f_s^i[k]\right)^2}\lambda^i$
7: $i \longleftarrow i + 1$
until convergence or number of maximum iterations are reached.
$\tilde{f_s} \longleftarrow f_s^i$

2) The ℓ_1-Penalized Least Squares Optimization Problem: We want to construct a ℓ_1-penalized least squares optimization problem to recover the sampled fluorescence signal f_s. The sparsity-promoting norm involves L as explained above. The obtained minimization problem reads:

$$\tilde{f_s} = \underset{f \in \mathbb{R}^K}{\operatorname{argmin}} \frac{1}{2}\sum_{k=1}^{K}\left(f_n[k] - f[k]\right)^2 + \lambda ||\mathrm{L}\, f||_1, \tag{9}$$

where $||.||_1$ is the vector ℓ_1- *norm*; $||\mathrm{L}\, f||_1 = \sum_{k=1}^{K} |\mathrm{L}\{f\}[k]|$.

Once an estimate $\tilde{f_s}$ is found, the deconvolved signal s can be obtained simply by inverting the CIRF effect. The formulation in (9) is equivalent to the often used synthesis framework in which the inverse—or pseudo-inverse—of L is plugged in the data-fidelity term [37]. As we will see in the subsequent section, our choice for the analysis framework is motivated by the existence of an automatic choice of the regularization parameter and corresponding fast algorithms. A second advantage is that it gives direct access to both f_s and s. This will enable us to measure the performances of the model in terms of noise removal. It is noteworthy that the regularization term in (9) can be seen as a particular case of the generalized total variation paradigm [22]. Finally, solving (9) under the *spike model* is equivalent to the popular spike deconvolution algorithm [12], but without non-negativity constraints. As we mentioned in the Introduction, at slow acquisition rates we are more interested in unraveling

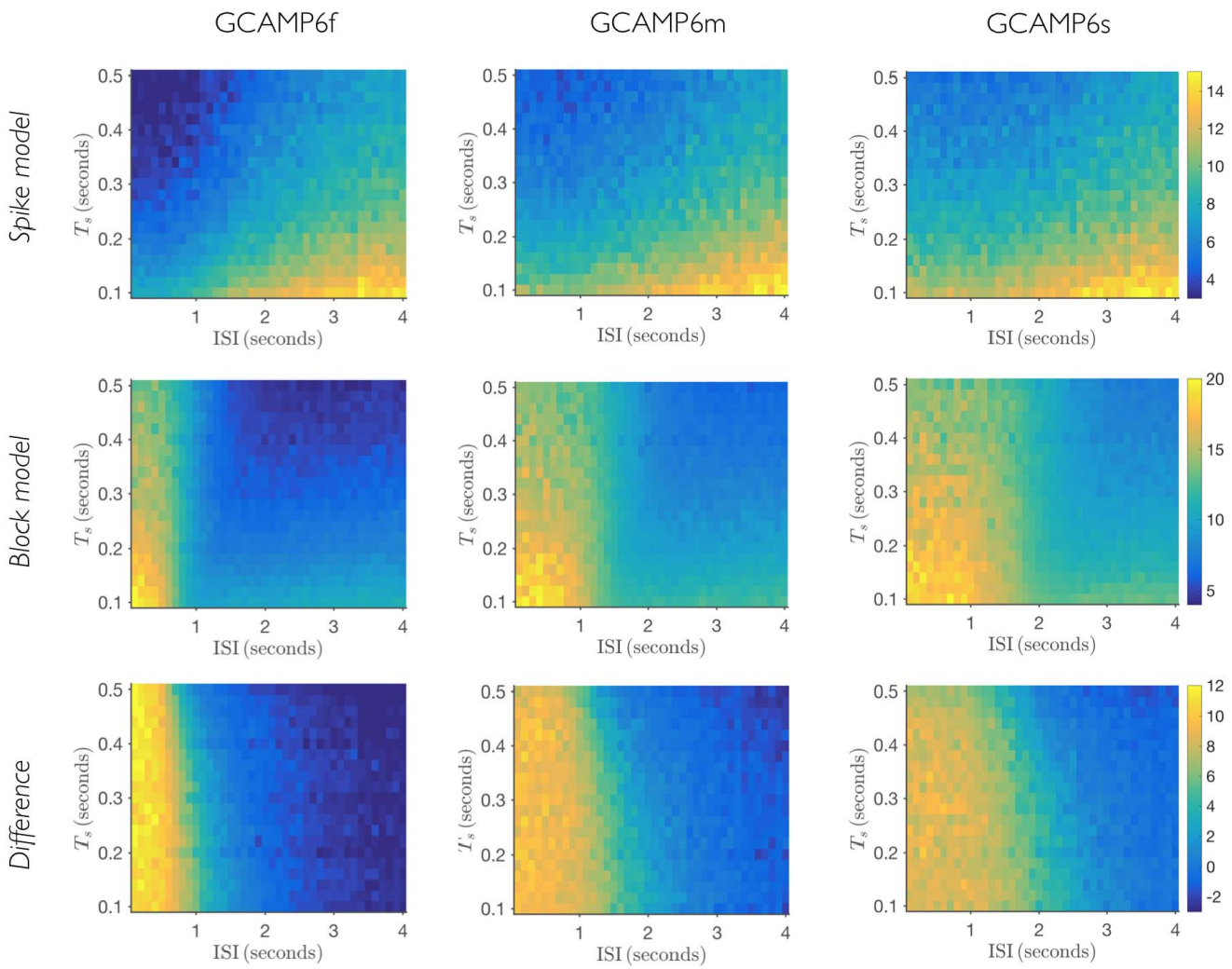

Fig. 3. SNR difference between recovered fluorescence from the two models for different ISI and T_S values. The plots show that, depending on the interactions between ISI and T_S, there are phase transitions at which one of the models outperforms the other. In particular, the slower the dynamics of the Ca^{2+} response are, the largest the region in which the *block model* is advantageous.

alternations between increase and decrease of neural activity compared to baseline firing.

B. Solution to the Minimization Problem via Fast Iterative Soft Thresholding

In order to solve (9), we use an accelerated version of the forward-backward splitting [38]. This splitting technique relies on two steps which are a gradient descent for data-fidelity and a proximal operator for ℓ_1-*norm* correction. Its accelerated version consists in updating the relaxation parameter for a faster convergence. Here, we customize the fast iterative soft thresholding algorithm (FISTA) [38], to deal with L. It requires to set a gradient descent step $\mu = \sigma_{\max}(\mathrm{L}^{+})$, the maximum eigenvalue of the adjoint operator L^{+}. It also requires to fix the regularization parameter λ. The entire routine is given in Algorithm 1 where $\mathcal{P}$ is the point-wise projection on the unit ball $\mathcal{P}(z) = \mathrm{sign}(z)\max(1, |z|)$. Step 6 in the loop is an updating technique that enables an automatic tuning of λ. It is due to Chambolle [39] and it is a distinctive feature of the analysis framework. However, λ needs to be initialized. Here, we used a pre-estimation of the noise variance $\tilde{\sigma}$ form a wavelet decomposition (*cf.* Supplementary Material). An open repository containing the full code is available at https://github.com/ufaro/CalciumDec.

IV. Experiments & Results

A. Simulated Signals

To assess the performance of the algorithm, we tested it on synthetic signals of varying spiking rate. Through these experiments we aim at exploring how the interplay between the temporal resolution and the calcium dynamics affects the fluorescence signal recovery.

The synthetic signal consists of a spike train of 2.5 min composed of three uniform bursting periods of 10 s, 20 s and 40 s, as illustrated in Figure 1, was first generated at a very high sampling rate to model the continuous domain (1000 Hz). We considered firing rates corresponding to ISI varying between 0.1 s and 4 s. The obtained signals where convolved with CIRFs corresponding to GCaMP6s, GCaMP6m and GCaMP6f. These responses where, then, sampled to a time resolution that is comparable to conditions in experimental data. Here, we varied the sampling period, T_s, between 0.1 s

and 0.5 s. Finally, we added a noise component such as the observed signal has a signal-to-noise ratio (SNR) of 5 dB. We solved the optimization problem (9) for both *spike model* and *block model* and we measured the performance in term of the SNR of the respective estimates $\tilde{f}_s$:

$$\mathrm{SNR}(\tilde{f}_s) = 10\log_{10}\left(\frac{\sum_{k=1}^{K}\tilde{f}_s[k]^2}{\sum_{k=1}^{K}\left(f_s[k]-\tilde{f}_s[k]\right)^2}\right). \quad (10)$$

In Figure 3, we report the performances of the two models. The results show that the *block model* gives an estimate that fits better the observed data when the ISI is small and the T_s is large. On the opposite, the *spike model* is beneficial when the spikes are separated enough or the resolution is high enough to capture single firings. Moreover, depending on the dynamics of the calcium indicator, the regions of accurate estimation are more or less large. In particular, the faster the dynamics (i.e., CIRF filter has a narrower support), the more the *spike model* is accurate and vice-versa for the *block model*. Notice that such regimes where sparse estimation fails are known to emergence in convex minimization approaches [40]. The difference between the two models in terms of SNR enables to observe phase transitions at which one of the models surpasses the other.

We have also computed the mean amplitude of the recovered activity signals in function of ISI for the three indicators. The results are reported in Figure 4 and demonstrate that this amplitude is proportional to the firing rate, as expected. Moreover, the slower the indicator dynamics are, the higher the amplitude because of activity accumulation. This means that, when undone from the effect of the CIRF, the changes in amplitude of the deconvolved signals are a sign of variations in the underlying firing rate.

Finally, in order to apprehend the outputs of the algorithm, Figure 5 displays the recovered signals at $T_s = 2$ s with ISI= 0.5 s for an initial noise level of 5 dB. The plots illustrate the behavior of the algorithm in situations of sustained activity with small ISI values. The *block model* recovers an on/off activity signal showing moments of rapid firing, whereas the *spike model* cannot resolve single spikes in this situation. When the activity is rather sparse than sustained; i.e, the ISI is large, detecting single spikes becomes feasible. In such case, the *block model* favors short blocks with small amplitudes on the actual spikes (*cf.* Supplementary Material - Section III). It is also noteworthy that the algorithm enables not only the detection of the beginning and the end of activation moments, but also changes in firing rates (*cf.* Supplementary Material - SectionV).

B. Experimental Data

We tested the method on neurons belonging to the anterior rhombencephalic turning region (ARTR) in maturing brain of zebrafish larvæ. ARTR was identified as the population of neurons responsible for setting the direction of spontaneous swimming during zebrafish exploratory locomotion [41]. This

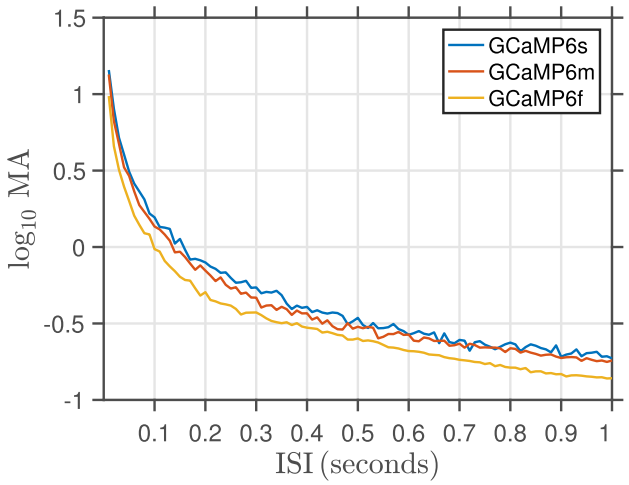

Fig. 4. Mean Amplitude (MA) of the recovered activity blocks in terms of ISI ($T_s = 0.2$ s). Since the signals are deconvolved, only variations due activity accumulation are visible. The longer the life-time of the calcium indicator, the larger the amplitude of the retrieved blocks.

allows us to evaluate the estimated activity moments by comparing them to the turning vector.

The data we used was obtained using transgenic zebrafish expressing the genetically encoded calcium indicator GCaMP6f [6]. Whole-brain neural activity was captured by a dual-laser light-sheet microscope capable of scanning the entire brain while avoiding direct exposure of the zebrafish retina to the laser beam [42]. Roughly, 1.87 brain volumes were recorded per second (approximately, every 530 ms) during fictive spontaneous locomotion. Heading directions over time were recorded with a high-speed camera and summarized in a vector of the same size of the time bins. At each time-point a value is assigned to describe the motion; the magnitude represents the strength of the motor event while the sign gives the direction (positive for right turns and negative for left turns). The data consists of approximately 27 minutes of simultaneous recordings from 94214 neurons within 31 z-plane slices. The signals were detrended to remove non-activity related baseline drifts. Similarly to the case of simulated data, the regularization parameter was estimated from a fine scale wavelet expansion.

We have, first, run our algorithm on the normalized fluorescence signal from a single neuron belonging to the right ARTR. The results are shown in Figure 6, in which the lower plots in the **B** and **C** panels clearly reveal estimated moments of activation along with the turning vector (positive for right turns and negative for left turns) in panel **A**. This result demonstrates how the temporal deconvolution approach succeeds at retrieving a neural activity that matches the locomotor behavior of the fish. Sustained activation moments in the left ARTR indicate periods of anti-clockwise turning. Another advantage of the block-wise model is its capacity to also retrieve sustained baseline moments. The result shows that these moments fits with the clockwise turning which confirms a negative correlation between turning and opposite ARTR regions.

Figure 6-C shows the estimated activity from a neuron outside the ARTR region. Here, baseline and activation moments

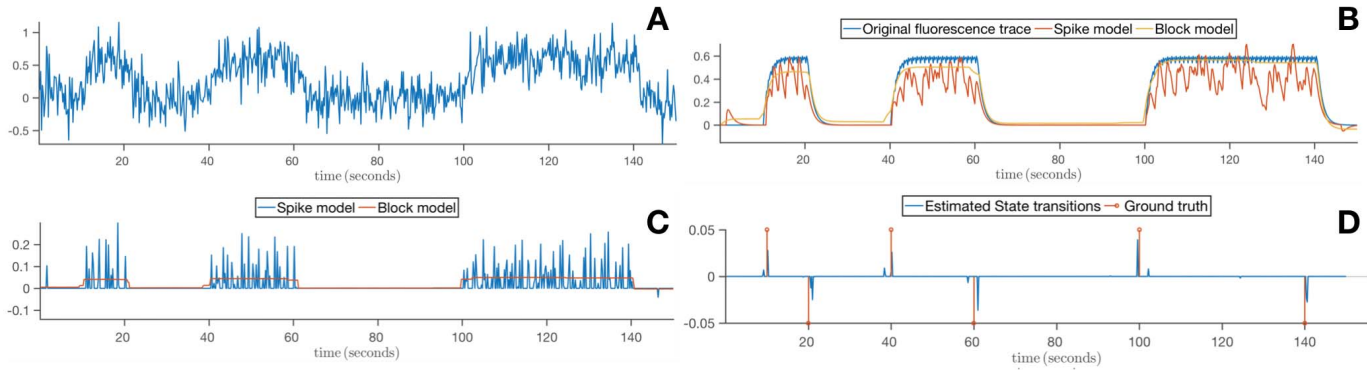

Fig. 5. Example of recovered activity (normalized) using GCAMP6m dynamics with a T_s of 0.2 s with ISI $= 0.5$ s for an initial noise level of 5 dB. **A**: noisy signal, **B**: denoised signal, **C**: deconvolved signal and **D**: Transient signal together with original ground truth transitions.

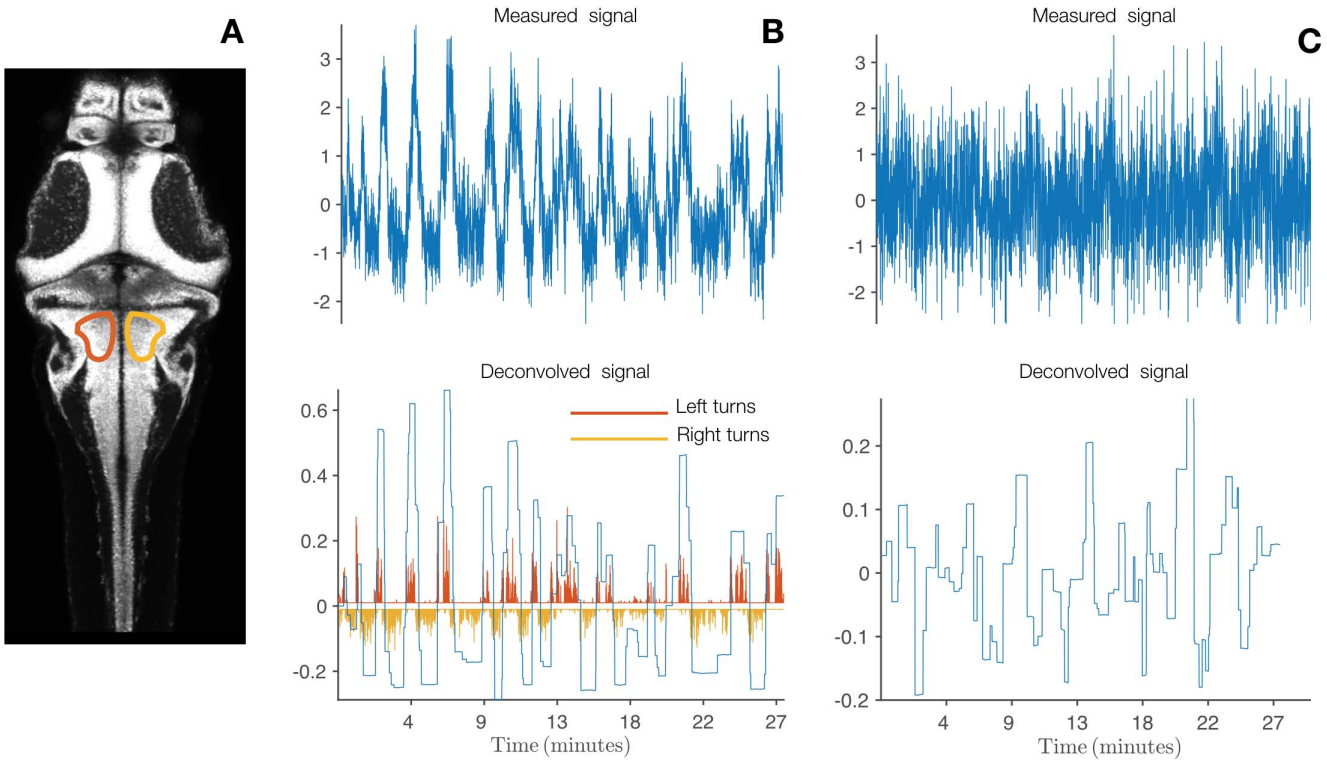

Fig. 6. Temporal deconvolution applied to experimental recordings. **A**: An anatomical scan of zebrafish larvæ [43] including neurons of the left and right ARTR region (highlighted in colors). In the bottom, the fish locomotion vector shows moments of right and left turns. **B**: Recovered activity from a neuron inside the left ARTR region: this region shows moments of persistent activity in which neurons continue firing when the associated turning direction is chosen. When the opposite direction is chosen, a drop of Ca^{2+} concentration is observed, which explains negatively valued blocks. **C**: Recovered activity from a neuron outside the ARTR region: here, activity moments are rather sparse without apparent oscillatory behavior.

are rather sparse and show less structure than locomotion driving neurons. This fact is highlighted even more in Figure 7. Therein, the estimated deconvolved traces from groups of neurons belonging to two different z-plane slices are shown. The first slice includes neurons from the ARTR region. The left and right activity components of the ARTR have the particularity of being asymmetric; i.e., they show anti-correlated patterns. Moreover, during exploratory swimming, this activity oscillates between the two sides permanently. If we look at the rest of neurons in the same slice, but outside the ARTR region, we can observe that the activity blocks have shorter durations. This suggests that neurons that are controlling swimming direction have perhaps a persistent activity similarly to oculomotor neurons that hold eye position [27]. Moreover, if we consider a slice not containing the ARTR regions as shown in Figure 7-(b), we can observe that activity is much more hemispherically symmetric in opposition to the ARTR area.

To appreciate more the results, we provide movies of spatial activity maps in the Supplementary Material. An example of an intensity map is shown in Figure 8-(a) where we can see how the activity patterns are spatially clustered and organized. The results also suggests that recordings from the ARTR region have strong intensity when involved; i.e., high increase of calcium concentration compared to the rest of the brain. Figure 8-(b) shows a map of estimated noise variance from a slice involving these region. The figure shows the variability of noise levels depending on the spatial position. In particular, signals that are recorded from the ARTR regions are of better quality than the ones recorded from the rest of the brain.

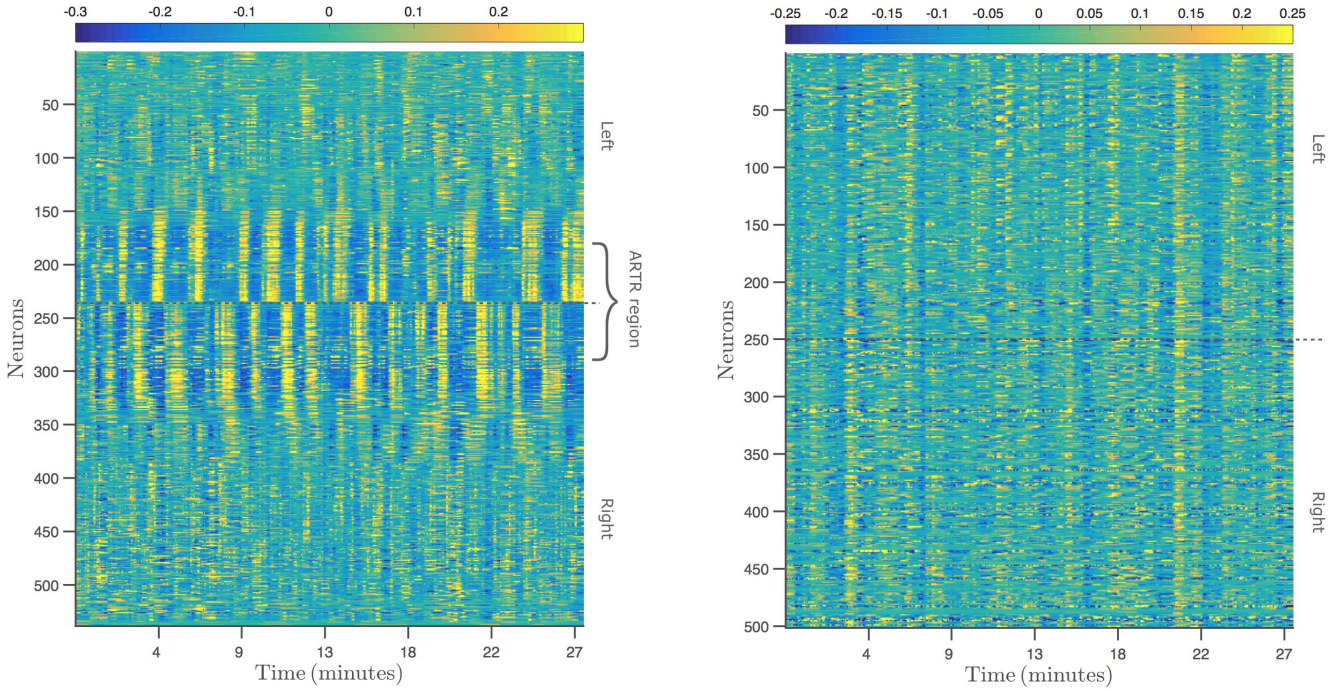

Fig. 7. Recovered fluorescence from two different z-plane slices. (a) A slice including a group of neurons from the ARTR region. These neurons are continuously oscillating between activation and de-activation moments. Moreover, these oscillations alternate between neurons in the left and the right hemispheres of the brain. (b) Another slice shows a sparse activity that is much more symmetric.

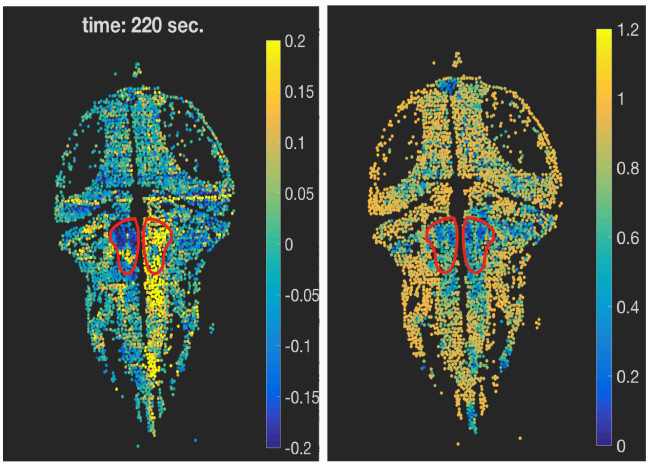

Fig. 8. A slice containing the ARTR regions (highlighted in red): (a) Intensity of the deconvolved signal at a fixed time point. (b) Estimated noise standard deviation.

On the computational side, our experiments were conducted on a laptop with Intel Core i7 CPU, 2.6 GHZ processor and 16 GB of RAM under MacOS 10.13.4, using MATLAB v.9.1, 64-bit. Because of the regularization parameter estimation, the speed of the algorithm varies depending on the noise variance and also on the number and duration (regularity) of the underlying signal. It took the algorithm (*block model*) between 0.3 and 1.2 seconds to process a single signal. The algorithm is of course parallelizable. For example, it took the algorithm 570 seconds to process a single z-slice (3402 single neurons recordings of 3050 time-points) on 4 parallel pools.

V. Discussion & Perspectives

We presented a method for deconvolution of sustained neural activity from calcium imaging data. Our procedure exploits the sparsity of activity transients to construct a ℓ_1-penalized least squares optimization. For this purpose, we used an analysis sparsity prior: the Ca^{2+} response operator is incorporated in the regularization and not in the data-fidelity term as it is often the case in sparse deconvolution algorithms. A distinctive advantage of this approach is that it gives direct access to a clean version of the fluorescence trace. This allowed us to study the performances of the estimation in terms of SNR. We used this measure as a criteria for comparing two sparsity prior models; *spike model* and *block model*. We found that depending on the relation between temporal resolution and minimum separation between the spikes, it can become advantageous to use a block prior. It is the case for slow imaging rates ($\sim 2\,$Hz) such as the ones used in large spatial field of view setups.

In moderate imaging rates, the choice for an appropriate model depends more on the underlying firing rate. Some neurons show fast sustained bursting and consequently their activity is better modeled using a *block model*. It can be also argued if some acquisition regimes do not require deconvolution; i.e, if simple denoising is enough. We investigated this question and we found that, in some sampling scenarios, Total Variation (TV) denoising [44] leads to similar quantitative results but the estimated state transitions are less accurate (*cf.* Supplementary Material). We also showed that the amplitude of the estimated activity blocks is linked to the underlying firing rate and depends on the dynamics of the Ca^{2+} indicator.

We also showed that the slower the dynamics of the Ca^{2+} indicator are, the more advantageous the *block model* is. In fact, slow indicators correspond to impulse responses with a large spread, which makes inference between successive spikes difficult.

The problem of estimating the actual firing rates is more complex and requires calibrating experimentally the mean amplitude of a single spike response. This will be subject of future work. On experimental data, we showed an example of activity estimation from whole brain zebrafish imaging at relatively slow imaging rates. The results confirm that the observed increase in Ca^{2+} concentration is often the results of sustained activity.

Moreover, the algorithm allowed to retrieve moments of decrease in activity compared to baseline firing. Negative fluctuations are indicative of neuronal inhibition. For example, in the ARTR case, when the zebrafish choses a turning direction, the contralateral ARTR is suppressed by contralateral inhibition [41]. This results in blocks with negative amplitudes in the deconvolved signal.

The outcome of this work offers many possible extensions. From an experimental angle, the proposed algorithm could be applied to *in vivo* recordings from other small animals for which the nervous system is well-characterized such as *Caenorhabditis elegans* [45]. The retrieved activity characteristics could be then linked to the function of the neurons.

On the methodological side, more robustness could be reached by adding a spatial regularization to the temporal deconvolution process as it is done in the fMRI methods [23]. It was also observed that some neurons show more sustained activity than others. It could be interesting to classify the neuronal population into neurons that are showing bursts frequently and those that are better modeled with a spike model.

Some neurons may exhibit a hybrid behavior; i.e., they generate individual spikes with a considerable large inter-spike interval, but also bursting periods. If the sampling rate enables retrieving isolated spikes, a possible approach to retrieve both spikes and bursting periods efficiently, would be to incorporate two regularization terms instead of one; i.e one term with the derivative as in the *block model* and one term without derivative as in the *spike model*. Such an approach will impose both sparsity and piece-wise constancy as in *fused lasso* [46].

It is also often observed that the calcium dynamics are diverse among neurons with the same expression. A semi-blind extension of the proposed algorithm could be used to estimate both the activity and the calcium response. Another idea would be to use known stimuli moments to estimate the shape of the calcium response for specific groups of neurons as it is done for the hemodynamic response [47]. Finally, the recovered activity could be used to study the connectivity of brain regions via correlation analysis or more sophisticated classification techniques [2]. Moreover, the state transitions in activity that are revealed by the *block model* could be used to reveal dynamic interactions between brain networks.

From a theoretical point-of-view, the proposed estimation procedure is an extension of adaptive piecewise polynomial estimation [48]. It would be interesting to derive fast rate error bounds and study the effect of perturbing pure differentiation operators such is the case for the Ca^{2+} response.

APPENDIX
DISCRETE OPERATOR CONSTRUCTION

In order to perform the deconvolution algorithm presented in this paper, an important step consists in constructing a discrete version of the operator $\mathcal{L}$ defined in (8) at a given sampling period T_s. This operator is of the form:

$$\mathcal{L} = \prod_{i=1}^{N} (\mathcal{D} - \alpha_i \mathcal{I}), \tag{11}$$

where,

$$\begin{cases} N = 2, \quad \alpha_1 = a, \text{ and } \alpha_2 = b, & \text{if } \textit{spike model}, \\ N = 3, \quad \alpha_1 = a, \quad \alpha_2 = b \text{ and } \alpha_3 = 0, & \text{if } \textit{block model}. \end{cases}$$

with a and b defined as in (3). Hereafter, we conventionally note $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)$. Operators of the form (11) correspond to finite impulse response (FIR) filters of order N ($N+1$ filter taps). Therefore, the discrete operator L has an exact construction [22] similar to the one used to construct exponential splines [49]:

$$\{\mathrm{L}\, f_s\}[k] = \sum_{\ell=1}^{N} f_s[k-\ell]\Delta_L[\ell], \quad k \geq N+1 \tag{12}$$

where the discrete filter is given in the time domain by:

$$\Delta_L[\ell] = (-1)^{\ell} \sum_{|\boldsymbol{m}|=\ell} (e^{\boldsymbol{\alpha}}), \quad \boldsymbol{m} \in [0,1]^N, \ 0 \leq \ell \leq N. \tag{13}$$

with $\boldsymbol{m} = (m_1, \cdots, m_N)$, $|\boldsymbol{m}| = \sum_{\ell=1}^{N} m_\ell$ and the conventions $c^{\boldsymbol{m}} = (c^{m_1}, \cdots, c^{m_N})$ and $\boldsymbol{c}^{\boldsymbol{m}} = \Pi_{\ell=1}^{N} c_\ell^{m_\ell}$, In particular, for the *spike model*, we have the following filter:

$$\Delta_L = [1, \quad -(e^{a\,T_s} + e^{b\,T_s}), \quad e^{a\,T_s + b\,T_s}], \tag{14}$$

which is similar to common filters that are used for spike detection [50], while for the *block model* the filter writes:

$$\Delta_L = [1, \quad -(1 + e^{a\,T_s} + e^{b\,T_s}), \quad e^{a\,T_s} + e^{b\,T_s} + e^{a\,T_s + b\,T_s}, \quad -e^{a\,T_s + b\,T_s}]. \tag{15}$$

Note that from a (discrete) linear algebra point of view, the taps in (15) can be obtained by combining a finite difference filter $[1 \ -1]$ with (14). The filtering operation in (12) is defined up to the first N point. Here, we used a zero boundary condition to define the filter on these points. As a result of this construction, the discrete operator L is full-rank although its continuous counterpart has a null-space spanning the vector space of second-order polynomials. Finally, the deconvolution algorithm we use here requires computing the adjoint operator L^+. This operator is constructed similarly to L using the time-reversed filter $\Delta_L^+[l] = \Delta_L[-l]$.

ACKNOWLEDGMENT

The authors would like to thank Misha Ahrens, Ph.D., and Yu Mu at Janelia Research Campus for providing whole-brain imaging data of fictively behaving larval zebrafish and sharing their code for pre-processing.

REFERENCES

[1] M. B. Ahrens *et al.*, "Brain-wide neuronal dynamics during motor adaptation in zebrafish," *Nature*, vol. 485, no. 7399, pp. 471–477, 2012.

[2] M. B. Ahrens, M. B. Orger, D. N. Robson, J. M. Li, and P. J. Keller, "Whole-brain functional imaging at cellular resolution using light-sheet microscopy," *Nature Methods*, vol. 10, no. 5, pp. 413–420, 2013.

[3] L. Tian *et al.*, "Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators," *Nature Methods*, vol. 6, no. 12, pp. 875–881, Dec. 2009.

[4] H. A. Zariwala *et al.*, "A Cre-dependent GCaMP3 reporter mouse for neuronal imaging *in vivo*," *J. Neurosci.*, vol. 32, no. 9, pp. 3131–3141, 2012.

[5] J. Akerboom *et al.*, "Optimization of a gcamp calcium indicator for neural activity imaging," *J. Neurosci.*, vol. 32, no. 40, pp. 13819–13840, 2012.

[6] T.-W. Chen *et al.*, "Ultrasensitive fluorescent proteins for imaging neuronal activity," *Nature*, vol. 499, no. 7458, pp. 295–300, 2013.

[7] D. Smetters, A. Majewska, and R. Yuste, "Detecting action potentials in neuronal populations with calcium imaging," *Methods*, vol. 18, no. 2, pp. 215–221, 1999.

[8] B.-Q. Mao, F. Hamzei-Sichani, D. Aronov, R. C. Froemke, and R. Yuste, "Dynamics of spontaneous activity in neocortical slices," *Neuron*, vol. 32, no. 5, pp. 883–898, 2001.

[9] J. D. Clements and J. M. Bekkers, "Detection of spontaneous synaptic events with an optimally scaled template," *Biophys. J.*, vol. 73, no. 1, pp. 220–229, 1997.

[10] B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, and F. Helmchen, "High-speed *in vivo* calcium imaging reveals neuronal network activity with near-millisecond precision," *Nature Methods*, vol. 7, no. 5, pp. 399–405, 2010.

[11] T. F. Holekamp, D. Turaga, and T. E. Holy, "Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy," *Neuron*, vol. 57, no. 5, pp. 661–672, 2008.

[12] J. T. Vogelstein *et al.*, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *J. Neurophysiol.*, vol. 104, no. 6, pp. 3691–3704, 2010.

[13] T. Sasaki, N. Takahashi, N. Matsuki, and Y. Ikegaya, "Fast and accurate detection of action potentials from somatic calcium fluctuations," *J. Neurophysiol.*, vol. 100, no. 3, pp. 1668–1676, 2008.

[14] L. Theis *et al.*, "Benchmarking spike rate inference in population calcium imaging," *Neuron*, vol. 90, no. 3, pp. 471–482, 2016.

[15] P. Berens *et al.*, "Community-based benchmarking improves spike rate inference from two-photon calcium imaging data," *PLoS Comput. Biol.*, vol. 14, no. 5, 2018, Art. no. e1006157.

[16] M. Pachitariu, C. Stringer, and K. D. Harris, "Robustness of spike deconvolution for neuronal calcium imaging," *J. Neurosci.*, vol. 38, no. 37, pp. 7976–7985, 2018.

[17] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, Jun. 2002.

[18] J. Oñativia, S. R. Schultz, and P. L. Dragotti, "A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging," *J. Neural Eng.*, vol. 10, no. 4, 2013, Art. no. 046017.

[19] L. Grosenick, J. H. Marshel, and K. Deisseroth, "Closed-loop and activity-guided optogenetic control," *Neuron*, vol. 86, no. 1, pp. 106–139, 2015.

[20] J. Friedrich, P. Zhou, and L. Paninski, "Fast online deconvolution of calcium imaging data," *PLoS Comput. Biol.*, vol. 13, no. 3, 2017, Art. no. e1005423.

[21] E. Ganmor, M. Krumin, L. F. Rossi, M. Carandini, and E. P. Simoncelli, "Direct estimation of firing rates from calcium imaging data," Jan. 2016, *arXiv:1601.00364*. [Online]. Available: https://arxiv.org/abs/1601.00364

[22] F. I. Karahanoğlu, I. Bayram, and D. Van De Ville, "A signal processing approach to generalized 1-D total variation," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5265–5274, Nov. 2011.

[23] F. I. Karahanoğlu, C. Caballero-Gaudes, F. Lazeyras, and D. Van De Ville, "Total activation: FMRI deconvolution through spatio-temporal regularization," *NeuroImage*, vol. 73, pp. 121–134, Jun. 2013.

[24] F. I. Karahanoğlu and D. Van De Ville, "Transient brain activity disentangles fMRI resting-state dynamics in terms of spatially and temporally overlapping networks," *Nature Commun.*, vol. 6, Jul. 2015, Art. no. 7751.

[25] P. Reinagel, D. Godwin, S. M. Sherman, and C. Koch, "Encoding of visual information by LGN bursts," *J. Neurophysiol.*, vol. 81, no. 5, pp. 2558–2569, 1999.

[26] E. M. Izhikevich, "Neural excitability, spiking and bursting," *J. Bifurcation Chaos*, vol. 10, no. 6, pp. 1171–1266, 2000.

[27] A. A. Koulakov, S. Raghavachari, A. Kepecs, and J. E. Lisman, "Model for a robust neural integrator," *Nature Neurosci.*, vol. 5, no. 8, pp. 775–782, 2002.

[28] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, vol. 1. Upper Saddle River, NJ, USA: Prentice-Hall, 1997.

[29] Z. Doğan, C. Gilliam, T. Blu, and D. Van De Ville, "Reconstruction of finite rate of innovation signals with model-fitting approach," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6024–6036, Nov. 2015.

[30] J.-M. Azaïs, Y. De Castro, and F. Gamboa, "Spike detection from inaccurate samplings," *Appl. Comput. Harmon. Anal.*, vol. 38, no. 2, pp. 177–195, 2015.

[31] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4735–4744, Oct. 2011.

[32] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Commun. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, Jun. 2014.

[33] V. Duval and G. Peyré, "Sparse spikes super-resolution on thin grids II: The continuous basis pursuit," *Inverse Problems*, vol. 33, no. 9, 2017, Art. no. 095008.

[34] B. Bernstein and C. Fernandez-Granda, "Deconvolution of point sources: A sampling theorem and robustness guarantees," *Commun. Pure Appl. Math.*, vol. 72, no. 6, pp. 1152–1230, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21805

[35] H. L. Taylor, S. C. Banks, and J. F. McCoy, "Deconvolution with the ℓ_1 norm," *Geophysics*, vol. 44, no. 1, pp. 39–52, 1979.

[36] T. Debarre, J. Fageot, H. Gupta, and M. Unser, "B-spline-based exact discretization of continuous-domain inverse problems with generalized TV regularization," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4457–4470, Jul. 2019.

[37] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus synthesis in signal priors," *Inverse Problems*, vol. 23, no. 3, p. 947, 2007.

[38] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[39] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1, pp. 89–97, 2004.

[40] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Inf. Inference*, vol. 3, pp. 224–294, Jun. 2014.

[41] T. W. Dunn *et al.*, "Brain-wide mapping of neural activity controlling zebrafish exploratory locomotion," *Elife*, vol. 5, Mar. 2016, Art. no. e12741.

[42] N. Vladimirov *et al.*, "Light-sheet functional imaging in fictively behaving zebrafish," *Nature Methods*, vol. 11, no. 9, pp. 883–884, 2014.

[43] X. Chen *et al.*, "Brain-wide organization of neuronal activity and convergent sensorimotor transformations in larval zebrafish," *Neuron*, vol. 100, no. 4, pp. 876–890, 2018.

[44] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, 1992.

[45] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode Caenorhabditis elegans," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 314, no. 1165, pp. 1–340, 1986.

[46] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 67, no. 1, pp. 91–108, 2005.

[47] L. Chaari, T. Vincent, F. Forbes, M. Dojat, and P. Ciuciu, "Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach," *IEEE Trans. Med. Imag.*, vol. 32, no. 5, pp. 821–837, May 2013.

[48] R. J. Tibshirani, "Adaptive piecewise polynomial estimation via trend filtering," *Ann. Statist.*, vol. 42, no. 1, pp. 285–323, 2014.

[49] M. Unser and T. Blu, "Cardinal exponential splines: Part I—Theory and filtering algorithms," *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1425–1438, Apr. 2005.

[50] S. Reynolds, C. S. Copeland, S. R. Schultz, and P. L. Dragotti, "An extension of the FRI framework for calcium transient detection," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 676–679.