Blind sparse deconvolution for inferring spike trains from fluorescence recordings

Jérôme Tubiana,[1,*] Sébastien Wolf,[2,3] and Georges Debregeas[2,3]
[1] *Laboratoire de Physique Théorique, Ecole Normale Supérieure and CNRS,*
PSL Research, Sorbonne Universités UPMC, 24 rue Lhomond,75005 Paris, France
[2] *Sorbonne Universités, UPMC Univ. Paris 06, UMR 8237, Laboratoire Jean Perrin, F-75005 Paris, France*
[3] *CNRS UMR 8237, Laboratoire Jean Perrin, F-75005 Paris, France*

(Dated: June 27, 2017)

The parallel developments of genetically-encoded calcium indicators and fast fluorescence imaging techniques makes it possible to simultaneously record neural activity of extended neuronal populations *in vivo*, opening a new arena for systems neuroscience. To fully harness the potential of functional imaging, one needs to infer the sequence of action potentials from fluorescence time traces. Here we build on recently proposed computational approaches to develop a blind sparse deconvolution algorithm (BSD), which we motivate by a theoretical analysis. We demonstrate that this method outperforms existing sparse deconvolution algorithms in terms of robustness, speed and/or accuracy on both synthetic and real fluorescence data. Furthermore, we provide solutions for the practical problems of thresholding and determination of the rise and decay time constants. We provide theoretical bounds on the performance of the algorithm in terms of precision-recall and temporal accuracy. Finally, we extend the computational framework to support temporal super-resolution whose performance is established on real data.

Introduction

In the last two decades, functional calcium imaging has emerged as a popular method for recording brain activity *in-vivo*. This technique relies on calcium sensors, either synthetic or genetically expressed, that are designed to optically report the transient rise in intra-cellular calcium concentration that accompany spiking events. Calcium imaging offers several assets in comparison with standard electrophysiology methods: it is non-invasive, it allows monitoring extended neuronal networks (up to a few tens of thousands of units), and it can be combined with genetic methods in order to target specific neuronal populations.

The main limitation of calcium imaging is that it only provides a proxy measure of the neuronal activity. The kinetics of the calcium/reporter complexation being relatively slow, the spike-evoked fluorescence transients last much longer (0.1-1s) than the action potential itself (<5ms). As the fluorescence signal is generally noisy and/or weakly sampled, its interpretation heavily relies on deconvolution methods to infer approximated spike trains. With the rapid increase in data-throughput offered by current fast imaging techniques, these methods need to be fast and unsupervised, as any manual check of the produced inference signals would prove impractical.

Standard inference methods are based on a generative model, which describes the relationship between a spike train and the resulting fluorescence time trace. It reads:

$$F_i \equiv F(t_i) = a \int K(\tau) N(t_i - \tau) d\tau + b + \epsilon_i \tag{1}$$

where $t_i = i\Delta t$ is the time of measurement, $N(t) = \sum_j \delta_{t,t_j}$ denotes the spike train, b is the baseline fluorescence (spikeless signal), and ϵ_i is a discrete gaussian white noise: $<\epsilon_i>= 0$, $<\epsilon_i \epsilon_j>= \sigma^2 \delta_{i,j}$. The convolution kernel $K(t)$, which reflects the complexation kinetics, is of the form:

$$K(t) \propto (e^{-\frac{t}{\tau_d}} - e^{-\frac{t}{\tau_r}}) \mathbb{1}_{t>=0} \tag{2}$$

where the rise and decay time constants τ_r, τ_d – typically in the range of 10-100ms and 50-1000ms, respectively – mostly depend on the calcium indicator but can also vary with the targeted neuron. In the following, we normalize K such that $\max(K) = 1$, hence each spike produces a transient of maximum height a. The signal to noise ratio (SNR) is thus defined as $\text{SNR} = \frac{a}{\sigma}$. The noise stems from fluctuations of intra-cellular chemical concentration, light source and

---

[*] Electronic address: `jerome.tubiana@ens.fr`

detector noise, incorrect baseline estimation, and other modeling errors. Because the SNR is often small in practice ($\in [2, 10]$) simple inference methods such as naive linear deconvolution and Wiener filtering are inadequate. In the last decade, numerous alternative deconvolution algorithms have been proposed [1–15]; among them, a powerful family of algorithms is based on non-negative sparse deconvolution [5, 13, 14]. In short, it consists in solving the inverse problem (equation (1)) using the *a priori* knowledge that the spikes are sparse and non-negative. This framework, introduced by Vogelstein *et al.* in [5], was shown to efficiently recover spike trains from fluorescence signals. However, its performance is strongly dependent on the algorithm's hyperparameters, namely the sparsity prior λ and the rise and decay time constants τ_r, τ_d. Despite extensive efforts for automatically adjusting these parameters [13, 14], progress are still needed to achieve the adequate robustness of the inference [16]. Another drawbacks of such algorithms is that the interpretation of the output can be challenging due to a paucity of theoretical understanding of its expected performance. First, the inferred signal is a continuous time series, which thus needs to be thresholded to yield a discrete set of spikes. Current methods remain elusive regarding how the threshold value controls the precision-recall of spike detection. Second, the temporal accuracy of the ouptut signal remains elusive: the probability that a given spike be inferred at the wrong time bin - in advance or delayed with respect to the true spike - is unkown. Conversely, one may expect that for high SNR configurations, the temporal resolution of the inferred signal may exceed the fluorescence sampling rate.

In the context of our laser-sheet calcium imaging experiments on zebrafish larvae, we built on the fast-oopsi algorithm to develop a so-called *Blind Sparse Deconvolution (BSD)* algorithm. This implementation features automatic estimation of the hyperparameters, enhanced speed and similar-to-better reconstruction performances and super-resolution capabilities. It is benchmarked on both synthetic and real data. We additionally provide thresholding guidelines and theoretical bounds on its performance, in terms of precision-recall and temporal accuracy, for large range of realistic experimental conditions. The program is available at `https://github.com/jertubiana/BSD`.

The article is organized as follows. In section 1, we present the principles of non-negative sparse deconvolution methods. In section 2, we focus on the determination of the sparsity prior λ; we compare our method to existing solutions and compare performances on synthetic data. In section 3, we compute the spike detection theoretical limits in terms of precision-recall and temporal resolution. In section 4, we focus on the automatic determination of the parameters of the generative model, in particular the time-constants of the convolution kernel K. In section 5, we extend BSD to support super-resolution, and characterize the resulting gain in temporal resolution on synthetic data. Finally, in section 6, we test our algorithm on real data, (i) fluorescence recordings in mice cortex with electrophysiology ground truth [17, 18], and (ii) whole-brain recordings of larval zebrafish [19, 20].

I. NON-NEGATIVE SPARSE DECONVOLUTION

In this first section, we review the existing deconvolution approaches for inferring $N(t)$. We rewrite Eqn. 1 as:

$$\begin{aligned} F_i &= a \sum_l K(t_i - t_l) + b + \sigma \epsilon_i \\ F_i &= a \sum_{j=1}^{T} K\left[\Delta t(i-j+1)\right] N_j + b + \sigma \epsilon_i \\ \Leftrightarrow \mathbf{F} &= a \mathcal{K} \mathbf{N} + \mathbf{b} + \sigma \epsilon \end{aligned} \tag{3}$$

where $i \in [1, T]$ is the time frame index, $\Delta t \equiv \frac{1}{f}$ is the sampling interval, $\mathcal{K}$ is the convolution matrix $\mathcal{K}_{ij} = K\left[\Delta t(i-j+1)\right]$ and $N_j = \int_{t'=(j-1)\Delta t}^{j\Delta t} N(t') \in \mathbb{N}$ is the number of spikes in the time interval $[(j-1)\Delta t, j\Delta t]$ [29]. Note that the first and second lines are not equivalent; in doing so, we assume that:

- The boundary condition $N(t) = 0, \forall t < 0$ holds. It is mostly true in typical recordings that start during inactive periods but this simplification can be easily relaxed. [30]
- We can approximate $K(t_i - t_l) = K(i\Delta t - t_l)$ as $K\left[i\Delta t - (j_l - 1)\Delta t\right]$ where $j_l - 1 = \lfloor \frac{t_l}{\Delta t} \rfloor$. This discretization error is negligible when Δt is small, yet it ensures that the matrix $\mathcal{K}$ is translation invariant, *i.e.* $\mathcal{K}_{ij} = \phi(i-j)$ [31].

Following 3, a naive estimate for N is:

$$\begin{aligned}\hat{\mathbf{N}} &= \frac{1}{a}\mathcal{K}^{-1}(\mathbf{F}-\mathbf{b}) \\ \Leftrightarrow \hat{\mathbf{N}} &= \arg\min_{N}\left\{\frac{1}{2}\sum_{i=1}^{T}[F_i - a(\mathcal{K}N)_i - b]^2\right\}\end{aligned} \tag{4}$$

As shown in Fig. 1a, this approach fails to recover any spike at typical noise level $SNR = 2.5$. To understand this failure, one may reason in the continuous framework for which Eqn. 4 writes $\hat{N}(t) \propto \int K^{-1}(\tau)\left[F(t-\tau) - b(t-\tau)\right] d\tau$. Here, the inverse convolution kernel K^{-1} is proportional to $\delta''(t) - \left[\frac{1}{\tau_r} + \frac{1}{\tau_d}\right]\delta'(t) + \frac{1}{\tau_r\tau_d}\delta(t)$ thus the naive deconvolution reads:

$$\hat{N} \propto \partial_t^2 F(t) + \left[\frac{1}{\tau_r} + \frac{1}{\tau_d}\right]\partial_t F(t) + \frac{1}{\tau_r\tau_d}F(t) \tag{5}$$

A naive estimator of the signal involves computing the derivatives of the original signal, and is thus extremely sensitive to high frequency noise. The reasoning is similar for discrete signals, which involve discrete time derivatives. An intuitive solution to mitigate this issue consists in filtering out the high frequency component before carrying out the deconvolution, as is the basis of the Wiener deconvolution method. Vogelstein et al. showed that it also performs poorly because it smoothes out the fast rise of the fluorescence signal at spikes. In contrast, non-negative sparse deconvolution estimators achieve both filtering of the noise while preserving the high-frequency signal. They are given by the outcome of the following optimization problem:

$$\hat{\mathbf{N}} = \arg\min_{N\geq 0}\left\{\frac{1}{2}\sum_{i=1}^{T}[F_i - a(\mathcal{K}N)_i - b]^2 + \lambda N_i\right\} \tag{6}$$

or equivalently :

$$\hat{\mathbf{N}} = \frac{1}{a}\arg\min_{N'\geq 0}\left\{\frac{1}{2}\sum_{i=1}^{T}(F_i - (\mathcal{K}N')_i - b)^2 + \lambda' N_i'\right\} \tag{7}$$

where λ, $\lambda' = \frac{\lambda}{a}$ are L_1 penalty coefficients that control the sparsity of the optimum (the higher λ, the sparser the optimum). When $\lambda = 0$ and the $N \geq 0$ constraint is relaxed, the optimal value $\hat{N}$ is exactly given by Eqn. 4. As shown in the next section, the choice of λ is crucial for efficient denoising and proper spike inference. Notice that the optimization problem is convex and can be solved efficiently in $\mathcal{O}(T)$ for double-exponential kernels using the interior-point method, see [5]. This unusual linear scaling for a matrix inversion-like operation owes to the fact that $\mathcal{K}^{-1}$ is tridiagonal for double-exponential kernels: $\mathcal{K}^{-1}_{ij} \propto \delta_{ij} - \gamma_1\delta_{i,j+1} + \gamma_2\delta_{i,j+2}$, with $\gamma_1 = \exp\left(-\frac{\Delta t}{\tau_r}\right) + \exp\left(-\frac{\Delta t}{\tau_d}\right)$, $\gamma_2 = \exp\left(-\frac{\Delta t}{\tau_r} - \frac{\Delta t}{\tau_d}\right)$. In [14], the authors apply the Pool-Adjascent Violator Algorithm originally developed for isotonic regression problems to solve this optimization in an even faster yet greedy way.

II. DETERMINATION OF THE SPARSITY PRIOR λ

The choice of the regularization parameter λ is crucial. If it is too large, the inferred spike train is $\hat{N} = 0$ and all spikes are missed, whereas if it is too small, noise-induced transients are interpreted as spikes $\hat{N} \neq 0$, yielding large false positive rates. Intuitively, we expect the optimal choice to depend on the parameters of the generative model (noise level, spike amplitude, etc.) Here we review the expressions of λ previously used and we then introduce our method. We adopt the convention from Eqn. 7 and drop the primes. We assume for now that all generative model parameters are known.

A. Review of existing methods: fast-oopsi and constrained-oopsi

In [5], the authors derive the non-negative sparse deconvolution from an approximate Maximum A Posteriori principle. They assume that the spike count N_i at time step i follows a Poisson distribution of mean $\nu\Delta t$, where ν

is the firing rate. After approximating the Poisson prior with an exponential distribution, they compute the negative log-likelihood $-\log P(F, N)$, which they find to be proportional to (7) with a sparsity prior λ given by:

$$\lambda_{oopsi} = \frac{\sigma^2}{a\nu\Delta t} \tag{8}$$

This approach thus provides an analytical expression for λ. However, due to the exponential approximation, this expression proves to be ineffective in several realistic experimental conditions as shown in Section 2C,D and illustrated in Figure 1.

To address this issue, a non-analytical method called constrained-oopsi (refered to as con-oopsi in the following) was recently introduced in [13]. The authors propose the following constrained deconvolution:

$$\begin{aligned} \hat{N} &= \arg\min_{N\geq 0} \sum_i N_i \\ \text{subject to } &\sum_i [F_i - (\mathcal{K}N)_i]^2 \leq \sigma^2 T \end{aligned} \tag{9}$$

Where T is the number of observations. The problem can be rewritten using the Karush-Kuhn Tucker conditions by introducing the Lagrangian $\mathcal{L} = \sum_i N_i + \rho \sum_i [F_i - (\mathcal{K}N)_i]^2$ where ρ is the Lagrange multiplier associated with the constraint. There exists ρ such that the critical point $N^\star$ of $\mathcal{L}$ is the solution of the constrained optimization problem. Clearly, $N^\star$ satisfies the constraint only if ρ is non-negative; in this case $\mathcal{L}$ is convex and the critical point is a minimum of $\mathcal{L}$. Overall, the optimization problem can be rewritten as:

$$\hat{N}(\rho) = \arg\min_{N\geq 0} \left\{ \sum_i N_i + \rho \sum_i [F_i - (\mathcal{K}N)_i]^2 \right\} \tag{10}$$

Identifying $\lambda = \frac{1}{2\rho}$, the constrained deconvolution is equivalent to a sparse deconvolution with an adaptive sparsity prior λ. Since $\sum_i \hat{N}_i(\rho)$ is a decreasing function of ρ, the expression for λ reads:

$$\lambda_{con-oopsi} = \max\{\lambda \in \mathbb{R}^+, \sum_i [F_i - (\mathcal{K}N)_i]^2 \leq \sigma^2 T\} \tag{11}$$

In practice, $\lambda_{con-oopsi}$ is found by alternatively solving Eqn. 7 and updating λ, by decreasing it if the reconstruction error is too large, or increasing it otherwise. This non-analytical approach performs better than fast-oopsi, see Section 2C,D and Figure 1. However, it comes at a cost of increased computational time, because the deconvolution problem must be solved many times and the number of iterations required for convergence is not known in advance.

B. Blind Sparse Deconvolution

We propose a different analytical expression for λ, inspired by [21]. It is deduced from the analysis of the optimization problem for two simple configurations, in which there is either zero or one spike in the original signal. We show that this solution combines the computational speed of fast-oopsi and the robustness of con-oopsi.

1. *Spikeless Signal*

In the following, we use matrix notations and rewrite the cost function as :

$$\mathcal{L}(\mathbf{N}) = \frac{1}{2} [\mathbf{F} - \mathcal{K}\mathbf{N}]^T [\mathbf{F} - \mathcal{K}\mathbf{N}] + \lambda \mathbf{1}^T \mathbf{N} \tag{12}$$

The gradient writes:

$$\begin{aligned} -\nabla_{\mathbf{N}} \mathcal{L} &= \mathcal{K}^T (\mathbf{F} - \mathcal{K}\mathbf{N}) - \lambda \mathbf{1} \\ -\nabla_{\mathbf{N}} \mathcal{L} &= -(\mathcal{K}^T \mathcal{K})\mathbf{N} + \mathcal{K}^T \mathbf{F} - \lambda \mathbf{1} \end{aligned} \tag{13}$$

Let's first assume that the signal is spikeless, such that $F_i = \sigma\epsilon_i$, where ϵ_i is a gaussian white noise. Since $(\mathcal{K}^T\mathcal{K})N > 0$, we have:

$$-\frac{\partial\mathcal{L}}{\partial N_i} < \sigma(\mathcal{K}^T\epsilon)_i - \lambda \sim \mathcal{N}\left(-\lambda, \sigma^2\|K\|^2\right)$$
$$\Rightarrow P\left[-\frac{\partial\mathcal{L}}{\partial N_i} > 0\right] < \Phi\left[\frac{\lambda}{\sigma\|K\|}\right] \tag{14}$$

where $\Phi(x) = \int_x^{+\infty}\frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$, and $\|K\| \equiv \sqrt{\sum_{i=-\infty}^{\infty} K^2\left[i\Delta t\right]}$

Therefore, if $\lambda = \lambda_1 \equiv z_1\sigma\|K\|$ with z_1 large enough, the gradients are almost always negative, and the global optimum of $\mathcal{L}$ is $\hat{\mathbf{N}} = \mathbf{0}$. Hence for instance, setting $z_1 = 2.326$, yields a probability of false positive event per time bin $P_{FP} < 0.01$.

2. *Single spike signal*

We now examine a configuration in which a single spike is present in the data :

$$N_i^0 = \delta_{i,i_0}$$
$$F_i = aK\left[\Delta t(i - i_0 + 1)\right] + \sigma\epsilon_i \tag{15}$$

The gradient writes:

$$-\nabla_{\mathbf{N}}\mathcal{L} = -(K^TK)(\mathbf{N} - a\mathbf{N^0}) + \sigma K^T\epsilon - \lambda\mathbf{1} \tag{16}$$

We look for an optimum of the form $\hat{N}_i = an\delta_{i,i_0}$. The optimization with respect to n gives:

$$n = \max\left\{1 - \frac{\lambda}{a\sum_{i=1}^T K^2\left[(i-i_0+1)\Delta t\right]} + \frac{\sigma}{a}\frac{\sum_{i=1}^T K\left[(i-i_0+1)\Delta t\right]\epsilon_i}{\sum_{i=1}^T K^2\left[(i-i_0+1)\Delta t\right]}, 0\right\}$$
$$n \sim \max\left[\mathcal{N}\left(1 - \frac{\lambda}{a\|K\|^2}, \frac{\sigma^2}{a^2\|K\|^2}\right), 0\right] \tag{17}$$

where the last line assumes that $\sum_{i=1}^T K^2\left[(i-i_0)\Delta t\right] \approx \sum_{i=-\infty}^{\infty} K^2\left[i\Delta t\right] \equiv \|K\|^2$, which is true provided that i_0 is far from the boundaries. Thus, if the spike position is known in advance, the inferred spike is a thresholded gaussian variable.

One may observe that the noise-to-signal level $\sigma' = \frac{\sigma}{a\|K\|}$ that appears in this expression is smaller than $\frac{\sigma}{a}$ by a factor $\frac{1}{\|K\|}$. This has an important consequence: since $\max(K) = 1$, the norm $\|K\| = \sqrt{\sum_i K(i\Delta t)^2}$ of the discretized kernel is proportional to $\sqrt{M}$ where M is the typical number of time frames over which K is non-zero. Thus, $\sigma' \sim \frac{\sigma}{a\sqrt{M}}$ as if the noise had been averaged over the duration of the transient. This suggests that low SNR signals can be efficiently inferred provided that the spike-induced fluorescence transient is sufficiently well sampled.

Eqn. 17 shows that when λ is too large, the probability that a given spike is undetected reads:

$$P_{FN} = P\left[n = 0\right] = \Phi\left[\frac{\|K\|\left(a - \frac{\lambda}{\|K\|^2}\right)}{\sigma}\right] \tag{18}$$

Therefore, setting $\lambda \leq \lambda_2 \equiv \|K\|^2a - z_2\sigma\|K\|$, with, say, $z_2 = 2.326$, guarantees a low false negative rate (FNR) as the probability that a spike is detected is then larger than 0.99.

The sparsity prior λ_{BSD} is chosen to minimize both the FPR and FNR. Hence, for $a = 1$, $\sigma = 0.1$, $\tau_r = 0.1$, $\tau_d = 0.5$, $f = 10Hz$, $\lambda_2 = 4.1379$ is much higher than λ_1. In this case, setting $\lambda_{BSD} = \lambda_1$ is the best solution, as smaller values of λ_{BSD} lead to less signal deformation. In contrast, for configurations such that $\lambda_1 > \lambda_2$, *i.e.* when $\sigma > \sigma^{max} = \frac{a\|K\|}{z_1+z_2}$, it is impossible to satisfy both constraints (low FPR and low FNR); in this case we use the crossover value $\lambda = \frac{z_1a\|K\|}{z_1+z_2} = \lambda_1(\sigma_{max})$.

3. *Sparsity prior for BSD*

To summarize, in our Blind Sparse Deconvolution (BSD) algorithm, the sparsity prior is set analytically as:

$$\lambda_{BSD} = z_1 \|K\| \min\left(\sigma, \frac{a\|K\|}{z_1 + z_2}\right) \tag{19}$$

where $\|K\| = \sqrt{\sum_{i=-\infty}^{+\infty} K(i\Delta t)^2}$ is the L_2 norm of the discretized convolution kernel K, and z_1, z_2 are two numbers ~ 2 that control the precision and recall, respectively.

4. *Thresholding the BSD inferred signal*

Some applications, such as network connectivity inference, may require to threshold the signal in order to get a binary spike train. Unlike previous methods, BSD provides rationale for choosing a threshold. Indeed, the computations performed in Section 2B1,2B2 shows that the (unnormalized) inferred spikes in the absence (resp. presence) of spikes are thresholded gaussians variables, with means $-\frac{\lambda}{\|K\|^2}$ and $a - \frac{\lambda}{\|K\|^2}$, respectively, and identical variance $\frac{\sigma^2}{\|K\|^2}$. Picking a threshold that separates the two distributions yields:

$$\theta = \min\left[z_3 \frac{\sigma}{\|K\|}, u\left(a - \frac{\lambda_{BSD}}{\|K\|^2}\right)\right] \tag{20}$$

where z_3 is a quantile of the normal distribution, and u a number between 0 and 1, say, 0.5. When θ equals the left term, the vast majority of the noise is efficiently filtered out such that any non-zero value in the output signal can be safely assigned to a spike; the right-hand term in turn prevents the threshold from becoming larger than the signal itself.

C. Qualitative Comparison

As we have just shown, the sparsity prior expression λ_{BSD} allows one to simultaneously minimize both the FPR and FNR. In contrast, the expression of λ_{oopsi} offers no guarantee that either is small in all situations. For instance, if $a = 1, \sigma = 0.1, \tau_r = 0.1, \tau_d = 0.5, \Delta t = 0.1s, \nu = 1Hz$, we find $\lambda_{oopsi} = 0.1$ and $\lambda_1 = 0.51$; hence λ_{oopsi} is too small which results in multiple noise-induced false spikes. On the other hand, for $\sigma = 0.25$ and $\nu = 0.1Hz$, we have $\lambda_1 = 1.25,\ \lambda_2 = 3.39\ \lambda_{oopsi} = 6.25$; in this case, λ_{oopsi} is too large, such that most of the spikes are missed.

We expect good performance as well for $\lambda_{con-oopsi}$, although differences with λ_{BSD} may exist. In the absence of spikes, the noise is expected to be completely filtered out, because the optimum of Eqn. 9 is $\hat{N} = 0$ in the large T limit. In the presence of spikes, we expect the spikes train to be recovered to some extent, since $\hat{N} = 0$ violates the constraint of Eqn. 9 but in a slightly noisier fashion. Indeed, observe that as soon as $\lambda > 0$, the spikes are on average underestimated: $\hat{N}_i = n\delta_{i,i_0}$ with $n < 1$, see Eqn 17. Thus, if there is a 'good' value λ such that we retrieve exactly the spikes at their positions, we typically find at the optimum $\sum_{i=1}^{T} \left(F_i - anK\left[(i - i_0)\Delta t\right]\right)^2 > \sigma^2 T$. Hence the constraint is violated, and λ is decreased, until false (noise-induced) spikes appear and reduce the reconstruction error below $\sigma^2 T$.

As an illustration, we compare the results of the three inference algorithms for a signal with $\sigma = 0.4$, $f = 10Hz$ in Figure 1a. For fast-oopsi, the sparsity prior is too large, and no spikes are inferred, whereas for both con-oopsi and BSD, the signal is correctly recovered. We notice however that BSD infers slightly less false spikes than con-oopsi.

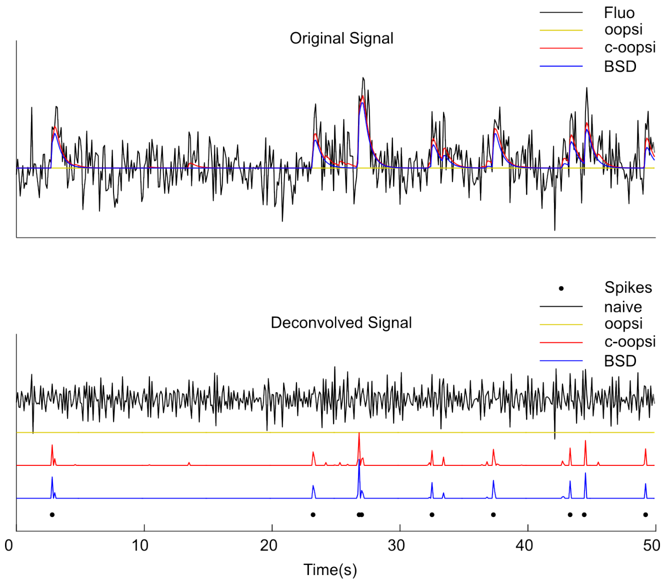

FIG. 1: Example of inference results.

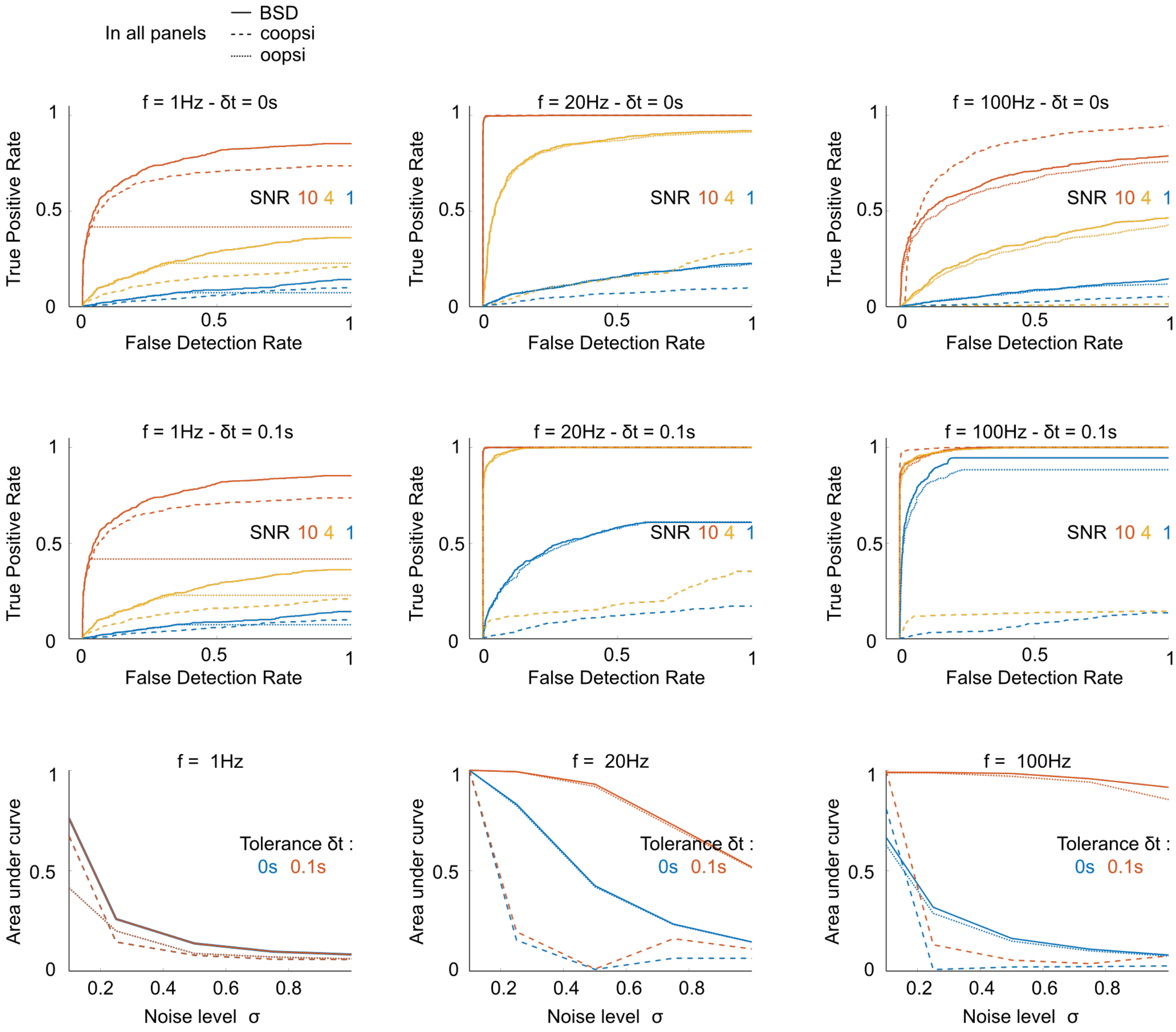

FIG. 2: $f = 10Hz$, $\sigma = 0.4$, $\tau_r = 0.1$, $\tau_d = 0.5$. Precision-recall curves and area under the curve. Top row: PR curves for $f = 1, 20, 100Hz$, and SNR, with a time tolerance for spike detection of $\delta t = 0$. Middle row: same, with time tolerance $\delta t = 0.1s$. Bottom row: Corresponding area under the curve as function of the noise level.

D. Quantitative Comparison

The performance of the three algorithms is now compared on a systematic benchmark. A random spike train is drawn from a Poisson distribution of mean firing rate $\nu = 0.1Hz$ over a duration $t = 10000s$. The signal is generated with the discrete model Eqn. 3 and a noise σ of varying amplitude. We use a double exponential kernel K with $\tau_r = 0.1$, $\tau_d = 0.5$, akin to a GCaMP6 reporter. Spike trains $\hat{\mathbf{N}}$ are inferred *with knowledge of the generative model parameters* and compared with the original spike train $\mathbf{N}$.

1. *Precision-Recall*

The precision-recall performance is evaluated as follows:

- The continuous spike train is binarized with a threshold θ.

- A spike is considered to be detected if the closest detected spike is within a time interval δt around the original spike; a true positive rate is computed (between 0 and 1) as the number of false negatives over the number of actual spikes (see [9]).
- Conversely, a false positive rate (FPR) is computed as the ratio of the number of inferred spikes that do not correspond to a real spike, over the number of actual spikes.
- The threshold is varied from ∞ to θ_{min}, such that θ_{min} corresponds to a maximum acceptable FPR, here taken to 1.
- Precision-recall graphs are plotted, and the area under the curve is computed.

The results are shown in Figure 1b. We observe that the fast-oopsi algorithm performs sometimes very well ($f = 100Hz$, high SNR) sometimes equivalently as con-oopsi and BSD ($f = 100Hz$, lower SNR), but often very poorly ($f = 10Hz$); Such unreliability may be highly detrimental in actual experiments. BSD and con-oopsi compare equivalently, with BSD slightly outperforming con-oopsi in most configurations. Notice that at low sampling rate ($f = 1Hz$), the P-R curve saturates at higher levels for BSD, suggesting that a large fraction of the spikes go undetected with con-oopsi.

2. *Speed*

BSD and fast-oopsi share the same algorithms, albeit with different parameters values; hence they have similar computational cost. The con-oopsi implementation, on the other hand, is slower because the sparse deconvolution has to be performed many times with different values of λ until convergence is reached. In practice, for experiments performed on a MacBook Air 2013, with 1.3 GHz Intel Core i5, we find a 3 to 25-fold increase in computation speed, depending on the array size. Our experiments shows that the number of iterations can be suprisingly large in practice. In particular, if the noise level is underestimated by con-oopsi, the error constraint is tighter and adding the positivity constraint may lead to no solutions at all - yielding many iterations in vain and increased computational time, see Table I. This reflects in the fact that the computing time is largely dependent on whether or not the noise level is provided.

Notice that the exact gain in speed depends on which version of con-oopsi is used (here, Matlab implementation, con-oopsi version of Dec. 2015, with cvx). Although we did not test the PAVA optimizer, we expect a gain of the same order of magnitude between constrained-PAVA and BSD-like PAVA. Such a difference in computation load may prove highly beneficial for real-time inference in high data-throughput recordings, as illustrated in section 6.

N_{frames}	BSD/ fast-oopsi (s)	con-oopsi (s)	con-oopsi σ user-provided
10^4	0.7	2.4	2.1
510^4	2.6	8.2	8.3
210^5	10	84	45
510^5	27	529	128
10^6	49	1594	290

TABLE I: Comparison of BSD, fast-oopsi and con-oopsi computational speed. For con-oopsi, under-estimation of the noise level σ, even for synthetic data, can lead to a large increase in computational time

III. THEORETICAL LIMITS ON THE PRECISION-RECALL AND TEMPORAL RESOLUTION

The fundamental motivations for spike train inference are to denoise the fluorescence signal and to improve the temporal resolution of the neural recording. Theoretically, if the generative model is correct, the convolution kernel is known and the isgnal is noiseless, then perfect retrieval of the spike train in terms of detection and timing can be achieved. Because of the noise, the accuracy is in practice limited by the rise and decay times: some spikes can be missed, have a wrong timing or be split across two successive time bins. These limitations have been characterized quantitatively in [22] in the context of Bayesian inference, when the noise is Poisson-like, τ_r is negligible and without super-resolution. However, no such analysis has been performed for sparse deconvolution algorithms. The BSD package incorporate routines to compute theoretical true and false positive rates and temporal resolution, for any given set of parameters (a, σ, $\mathbf{K}$) that can be extracted from the data.

A. Precision-Recall for isolated spikes

The theoretical false positive and negative rates (FPR, FNR) are first computed within the sparse deconvolution framework with λ_{BSD}. For the false positive rate, the computation was performed in Section 2B: we obtain a probability of false positive rate per time bin:

$$P_{FP} = \Phi\left[z_1 \min\left(1, \frac{a\|K\|}{\sigma(z_1+z_2)}\right)\right] \tag{21}$$

For the false negative rate, we follow a similar reasoning than in Section 2C: we consider a signal of the form $F_i = an_0 K\left[\Delta t i - t_0\right] + \sigma\epsilon_i$ with $t_0 = (i_0-1)\Delta t + \delta t_0$ and $0 \leq \delta t_0 < \Delta t$. Note that we have now relaxed the previously made approximation $K\left[\Delta t i - t_0\right] \approx K\left[\Delta t(i-i_0+1)\right]$ in order to probe the effect of intermittent sampling. We obtain a lower bound [32] for the probability of false negative per spike:

$$\begin{aligned} \hat{N} &= \arg\max_{N\geq 0} \mathcal{L}(\mathbf{N}) \approx an\delta_{i,i_0} \\ \implies n &\sim \max\left[\mathcal{N}\left(n_0\cos\theta(-\delta t_0) - \min\left(z_1\tilde{\sigma}, \frac{z_1}{z_1+z_2}\right), \tilde{\sigma}^2\right), 0\right] \\ \implies P_{FN}(\delta_0) &= \Phi\left[\frac{n_0\cos\theta(-\delta t_0) - \min\left(z_1\tilde{\sigma}, \frac{z_1}{z_1+z_2}\right)}{\tilde{\sigma}}\right] \end{aligned} \tag{22}$$

where

$$\begin{aligned} \cos\theta(\delta t) &= \frac{\sum_{l=-\infty}^{\infty} K\left[\Delta t l\right] K\left[\Delta t l + \delta t\right]}{\sum_{l=-\infty}^{\infty} K\left[\Delta t l\right]^2} \\ \tilde{\sigma} &= \frac{\sigma}{a\|K\|} \end{aligned} \tag{23}$$

Note that the probability depends on δt_0; for instance if $\tau_r = 0$ and $\delta t_0 << \Delta t$, spikes emitted right after a measurement yield low-amplitude fluorescent transients and are thus likely to be missed. Overall, the probability of false positive is given by:

$$P_{FN} = \frac{1}{\Delta t}\int_{\delta t_0=0}^{\Delta t} P_{FN}(\delta t_0) d\delta t_0 \tag{24}$$

We display in Figure 2(a) the true positive rate (TPR) for different sampling rates, as a function of the noise level for $\tau_r = 0.1$, $\tau_d = 0.5$. The FPR is set at 0.01/frame (i.e. $\lambda_{BSD} = \lambda_1$ and $z_1 = 2.366$). An important insight of this graph is that at low sampling rate, the TPR quickly decays with the noise level because spikes emitted shortly after a measurement are often completely missed. Conversely, improving the TPR (with e.g. $z_2 = 2.366$) yields a large number of false positives at low sampling rate.

B. Temporal Resolution

1. *Rough analytical estimate for isolated spikes*

We now estimate the temporal resolution by examining the probability distribution of $\hat{\mathbf{N}} = \arg\min_{\mathbf{N}} \mathcal{L}(\mathbf{N})$ given a single-spike noisy fluorescence signal $F_i = aK\left[\Delta t(i - i_0 + 1)\right] + \sigma\epsilon_i$ [33]. Since the distribution cannot be computed explicitly, we start with a simpler heuristic computation, and study the distribution of the initial negative gradients $-\frac{\partial \mathcal{L}}{\partial N_i}|_{\mathbf{N}=0}$. Consider indeed the gradient descent optimization dynamics. Because of the L_1 penalty, large components N_i tend to grow faster and to screen neighbouring small components, yielding sparse solutions with only few non-zero components; it is therefore likely that the largest components of $\mathbf{N}$ after one gradient descent step (after which $N_i \propto -\frac{\partial \mathcal{L}}{\partial N_i}|_{\mathbf{N}=0}$) remains the largest at the end of the optimization. Hence if the initial negative gradient is larger at position $i_0 + \delta$ than at position i_0, we expect the inferred spike $\hat{\mathbf{N}}$ to be similarly delayed with respect to the true spike position. The probability of such an error can be computed as:

$$\begin{aligned} -\frac{\partial \mathcal{L}}{\partial N_{i_0}}|_{N=0} &= a\|K\|^2 + \sigma \sum_i K\left[\Delta t(i - i_0 + 1)\right]\epsilon_i - \lambda \\ -\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}}|_{N=0} &= a\|K\|^2 \cos\theta(\delta\Delta t) + \sigma \sum_i K\left[\Delta t(i - i_0 - \delta + 1)\right]\epsilon_i - \lambda \\ \Rightarrow \Delta\left[-\frac{\partial \mathcal{L}}{\partial N}\right] &\sim -2a\|K\|^2 \sin^2\frac{\theta(\delta\Delta t)}{2} + 2\sigma\|K\| \sin\frac{\theta(\delta\Delta t)}{2}\mathcal{N}(0,1) \end{aligned} \tag{25}$$

where the angle $\theta(\delta t)$ is defined in Eqn 23.

Thus, the initial gradient at the offset time $i_0 + \delta$ is higher than its value at the spike time i_0 with probability $\Phi\left[\frac{a\|K\| \sin\frac{\theta(\delta\Delta t)}{2}}{\sigma}\right]$, typically resulting in a time-shifted inferred spike. This results in a typical timing error δt on the spike position of the order of:

$$\delta t \;\; s.t. \;\; \sin\frac{\theta(\delta t)}{2} = \frac{\sigma}{a\|K\|} \tag{26}$$

This timing error is a non-trivial function of the kernel K and the noise level. The higher the effective noise level $\frac{\sigma}{a\|K\|}$, the higher δt. The second factor is small for rapidly growing $\theta(\delta t)$, *i.e.* when the overlap between the kernel $K(t)$ and its lagged version $K(t + \delta t)$ is a fast decaying function of δt. Hence, the 'sharper' the kernel, the lower δt. Notice that δt does not depend on λ, suggesting that the temporal resolution is intrinsically limited by the noise, and not by the algorithm.

2. *Response function for isolated spikes*

In a more rigorous fashion, one can estimate the response function to an isolated spike under the assumption that the solution of the optimization is a Dirac, $\hat{N}_i = a n_\delta \delta_{i,i_0+\delta}$. The optimization result is given by the following equation:

$$\begin{aligned} \hat{\mathbf{N}} &= \arg\min_{\mathbf{N}\geq 0} \mathcal{L}(\mathbf{N}) \\ \Longrightarrow \hat{N}_i &\approx a n^\star_{\delta^\star} \delta_{i,i_0+\delta^\star} \\ \delta_\star &= \arg\min_{\delta}\min_{n_\delta} \mathcal{L}(n_\delta, \delta) \end{aligned} \tag{27}$$

For a given fluorescence trace, the inner optimization over n_δ can be carried out analytically, and the outer optimization numerically. By computing the optimal offset $\delta^\star$ for various noise realizations, one can estimate the probability distribution of the inferred spike offsets $P(\delta^\star)$. It is a function of the parameters a, σ, $\|K\|$, z_1, z_2. This computation can be easily generalized to support the super-resolution setting (see Annex C for the analytical details). Some response functions are displayed in Figure 2b-inset for typical calcium indicators. We used the parameters: (i) GCaMP6s: $\tau_r = 180ms$, $\tau_d = 0.55$, (i) GCaMP5k: $\tau_r = 58ms$, $\tau_d = 0.52s$, (i) GCaMP6f: $\tau_r = 25ms$, $\tau_d = 0.38s$,

(i) OGB1-like: $\tau_r = 20ms$, $\tau_d = 80ms$. For the first three set of time constants, the values are deduced from the fluorescence recordings on mice V1 cells reported in [18].

We also display in Figure 2b the width δt (measured by a gaussian fit) as a function of the noise level, for a sampling frequency $f = 60Hz$. As expected, the temporal resolution of the spikes can be lower than the sampling period if the noise is large, and we observe that reporters with large τ_r yield lower temporal resolution.

We finally examine the impact of the sampling frequency on the temporal resolution. In an experiment with a fixed number of sampled neurons, increasing the sampling rate $f \equiv \Delta t^{-1}$ by a factor s typically comes at the cost of reducing the exposure time τ_e by the same factor s, which in turn increases the noise σ by $\sqrt{s}$; it is therefore not obvious that one would improve the temporal resolution by increasing the sampling frequency. We display in Figure 2c for the same set of calcium indicators, the inverse width δt^{-1} as a function of the sampling rate f, for various signal to noise ratios (SNRs) at a reference frequency 10 Hz. We see that δt^{-1} saturates at a value that depends on the SNR and on the rise and decay constant times. Hence, with GCaMP6s, increasing the frequency beyond 50 Hz does not result in improving the temporal resolution.

The fact that the temporal resolution saturates can be seen from Eqn 26: asymptotically, we have $\|K\| \propto \sqrt{\Delta t}$, and since $\sigma \propto \sqrt{\Delta t}$, the effective noise level $\frac{\sigma}{a\|K\|}$ reaches a well-defined limit - and so does δt.

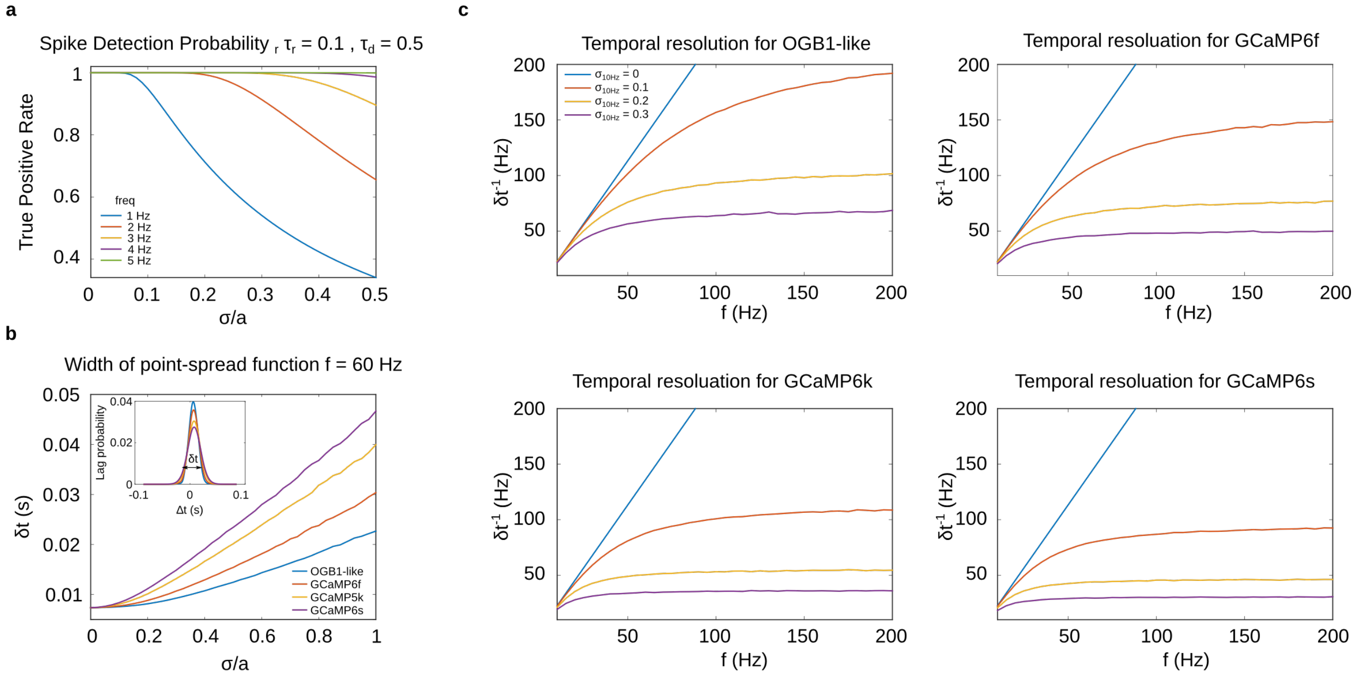

FIG. 3: (a) True positive rate of BSD as a function of the noise level for different sampling frequencies, for a fixed FPR of 0.01/frame, (b) Width of the response function as function of noise for various calcium indicators, at fixed frequency $f = 60Hz$, (c) Width of the response function as a function of the sampling frequency for various calcium indicators and reference noises.

IV. HYPERPARAMETERS LEARNING

All sparse devonvolution methods rely on the knowledge of the generative model's parameters. However, owing to the variability in the calcium reporters intracellular concentration and other biochemical cellular processes, these paramaters may significantly vary from experiment to experiment, and for different neuronal types. In the fast-oopsi implementation, the authors proposed to infer the parameters (a, b, σ, ν) in an iterative way: an initial guess is made, deconvolution is performed, parameters values are then updated based on the deconvolution result, whereas for con-oopsi, the authors propose to estimate σ, K only once. We follow the same iteration-based approach as fast-oopsi, but the parameters are inferred and refined differently; we also add a method to infer and refine the kernel K.

A. Initial estimation of the parameters

We are given a time series of the form $\mathbf{F} = a\mathcal{K}\mathbf{N} + \sigma\epsilon + b$, with unknown a,b,σ, K. In the following, we assume that the baseline is constant or equivalently that the variable baseline has been previously estimated and subtracted from the signal. From the knowledge that N is non-negative and sparse, we deduce that:

- The baseline b is essentially the most often observed value of F; the data histogram is computed, and b is estimated as the center of the interval with highest frequency. Using the median of S also provides a good estimator.
- All activity below the baseline originates from the noise, hence $F' = F[F < b] - b$ follows a half-Gaussian distribution $\min\left[\mathcal{N}\left(0, \sigma^2\right), 0\right]$; it is fitted to deduce σ.

In a similar spirit to con-oopsi, we estimate the convolution kernel K through the signal auto-correlation matrix. Indeed, observe that :

$$A_F(l) \equiv < F_i F_{i+l} > - < F >^2 = a^2 \sum_{j,k} \left[< N_j N_{j+k} > - < N >^2 \right] \sum_i K\left[(i-j-1)\Delta t\right] K\left[(i+l-k-j-1)\Delta t\right] + \sigma^2 \delta_{l,0} \tag{28}$$

Under the assumption that the spiking events N_i are independent, identically distributed Poisson variables, we have $< N_j N_{j+k} > - < N >^2 = a^2 \nu \Delta t \delta_{k,0}$ and Eqn. 28 can be simplified as:

$$\begin{aligned} A_F(l) &= a^2 \nu \Delta t \sum_{j=-\infty}^{\infty} K\left[j\Delta t\right] K\left[(l+j)\Delta t\right] + \sigma^2 \delta_{l,0} \\ \iff \left(A_F(l) - \sigma^2 \delta_{l,0}\right) &\propto \sum_{j=-\infty}^{\infty} K\left[j\Delta t\right] K\left[(l+j)\Delta t\right] \end{aligned} \tag{29}$$

The auto-correlation matrix can be estimated from the data as $\hat{A_F}(l) = \frac{1}{T} \sum_i F_i F_{i+l} - \left[\frac{\sum_i F_i}{T}\right]^2$. Together with the previous estimate of σ, the left-hand side of the equation can thus be estimated [34]. The right-hand side is the overlap between the kernel K and its delayed version $K'(t) = K(t + l\Delta t)$. We can normalize both terms to 1 for $l = 0$, and use a least square fit to estimate K.

Lastly, the spike amplitude a and frequency ν can be deduced from the following equations, that hold under the model assumption:

$$\begin{aligned} < F > &= a \nu \Delta t \sum_i K\left[i\Delta t\right] \\ < F^2 > - < F >^2 &= a^2 \nu \Delta t \sum_i K\left[i\Delta t\right]^2 \end{aligned} \tag{30}$$

Although they yield very good results for synthetic datasets, these estimators can fail in several frequently encountered situations in practice:

- When the neural activity is not sparse, we do not expect b to be the most frequent fluorescence value. An error in the estimation of b can result in a misestimation of σ as well.
- When the neuron displays bursting activity (*i.e.* several spikes in short time intervals), the hypothesis that the N_i are independent usually fails. This may result in overestimating τ_r and/or τ_d.
- In the same situation, Eqn 30 is incorrect and a can be overestimated.
- When the noise exhibits temporal correlation (streaking artefacts in light sheet imaging, small sample drifts, fluctuations in laser intensity, etc.), the white-noise hypothesis does not hold, which may result in a misestimation of τ_r and τ_d.

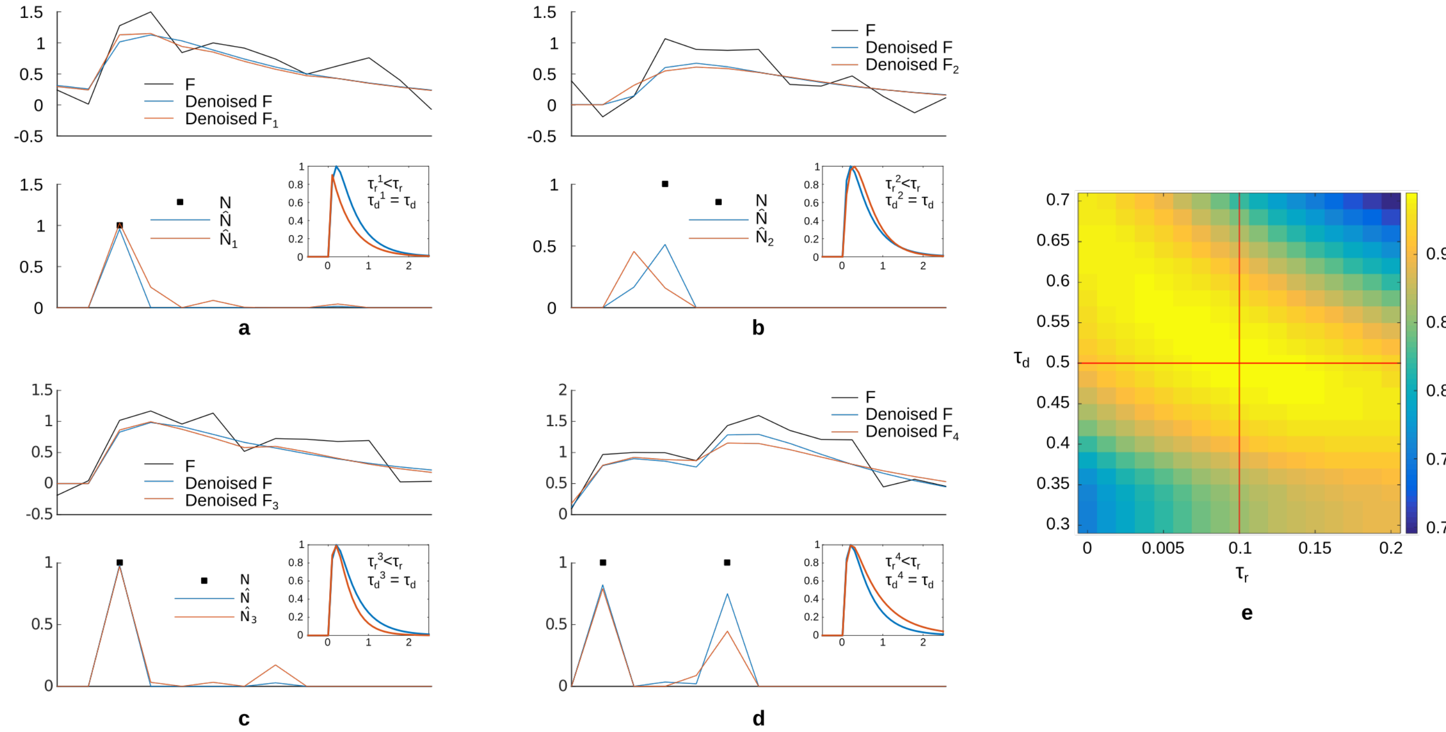

FIG. 4: Left: Example results of spike inference on synthetic data with mismatched convolution kernels. For each of the four figures, a fluorescence signal is generated with a kernel K^0; inference is performed with true parameters (blue curves) and with mismatched parameters (red curves). The true and mismatched kernels K^0 and K are depicted (insets). Systematic errors appear in the spike timings. Right: Area Under Curve classification performance with time tolerance $\delta t = 0s$, as a function of the rise and decay time constants. The parameters used to generate the signal, depicted in red, are typical of a GCaMP6 reporter.

We systematically studied the bias in spike inference that arises when the estimated time constants τ_r and τ_d differ from their true values, τ_r^0 and τ_d^0. As illustrated in Figure 3a-d, inferring the spikes with an incorrect convolution kernel leads to systematic errors. Suppose for instance that $\tau_r < \tau_r^0$ and $\tau_d = \tau_d^0$ (Figure 3a). Then a spike-induced fluorescence transient tends to exhibit a faster initial rise than expected. Hence, from a Bayesian perspective, such a transient is likely to be interpreted as two small consecutive spikes. Such a mismatch in the kernel parameters will therefore show up as duplicate spikes. In general, the nature of the error depends on the kernel mismatch:

- $\tau_r < \tau_r^0, \tau_d = \tau_d^0$ (Figure 3a): the inferred spikes are split in two, to compensate for the smaller rise time than expected for a single spike.
- $\tau_r > \tau_r^0, \tau_d < \tau_d^0$ (Figure 3b): the inferred spikes are in advance, to compensate for the faster rise of the fluorescence signal.
- $\tau_r = \tau_r^0, \tau_d < \tau_d^0$ (Figure 3c): the inferred spikes exhibit 'echos' to compensate for the slower than expected decay of F.
- $\tau_r = \tau_r^0, \tau_d > \tau_d^0$ (Figure 3d): the inferred subsequent spikes are 'screened' (lower amplitude) to compensate for the slower than expected signal decay.

We quantified how a kernel misestimation degrades the decoding performance by evaluating the relative reduction in precision-recall (area under curve) for various offsets of τ_r and τ_d (Figure 3e). Interestingly, some direction of the mismatch vector can be less deleterous: when both $\tau_r > \tau_r^0, \tau_d < \tau_d^0$ or $\tau_r > \tau_r^0, \tau_d < \tau_d^0$, the loss in performance remains modest.

B. Iterative parameter estimation: adaptive blind deconvolution

As previously explained, the initial generative model parameters are not always the true ones. To improve the estimates, let us write the cost function:

$$\mathcal{L}(\mathbf{N}, K, b, \sigma, a) = \frac{1}{2}\|\mathbf{F} - \mathcal{K}\mathbf{N} - \mathbf{b}\|^2 + \lambda_{BSD}(\|K\|, \sigma, a)\|\mathbf{N}\|_1 \tag{31}$$

The spikes are inferred by optimizing $\mathcal{L}$ with respect to $\mathbf{N}$. In order to improve the model parameters, we also optimize over K and b through a coordinate gradient descent:

$$\begin{aligned} \hat{\mathbf{N}}^{(\mathbf{t})} &\leftarrow \arg\min_{\mathbf{N}} \mathcal{L}(\mathbf{N}, \mathbf{K}^{(\mathbf{t-1})}, \mathbf{b}^{(\mathbf{t-1})}, \sigma, \mathbf{a}) \\ \left(K^{(t)}, \mathbf{b}^{(t)}\right) &\leftarrow \arg\min_{K,b} \mathcal{L}(\hat{\mathbf{N}}^{(t)}, K, \mathbf{b}, \sigma, a) \end{aligned} \tag{32}$$

The first optimization step was discussed in Section 1. The second optimization is a parametric temporal regression problem; it can be solved efficiently in $\mathcal{O}\left(\frac{\tau_r+\tau_d}{\Delta t}\right)$ by introducing the cross-correlation $X_\tau = \frac{1}{T}\sum_i F_i \hat{N}_\tau$ and auto-correlation $A_\tau = \frac{1}{T}\sum_i \hat{N}_i \hat{N}_{i-\tau}$ functions up to some cut-off $\tau_m \sim \frac{\tau_r+\tau_d}{\Delta t}$ (see details in Annex B). The point of this step is that if $N^{(t)} = N_0$, then the optimum is exactly K_0 if σ is small or T is large. More generally (N_0, K_0) is a fixed point of the optimization dynamic in the low noise limit, and intuitively, we expect that at finite noise, another fixed point close to (N_0, K_0) exists and can be reached. We show in Annex B that K^0 is the global optimum in the case of isolated spikes and low noise level. The optimization will not necessarily converge to such solution because the function $\mathcal{L}(N, K)$ is not convex, and only local minima are found. In practice, the optimum is usually very close to the original convolution kernel, and is reached if the initial estimate is good enough (Figure 5). The convergence can be improved by thresholding the spikes before updating the kernel, as it prevents false spikes from contributing to the cross-correlation. The iterative process is no longer an optimization but it still converges.

The noise σ and spike amplitude a can be refined as well, using

$$\hat{a} = \frac{\sum_t \hat{N}'_t}{\sum_t 1_{\hat{N}_t>0}} + \frac{\lambda}{\|K\|^2} \tag{33}$$

Where the last term corrects the bias due to the sparse prior (see Eqn. 17)

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_t \left[(F_i - (\mathcal{K}N')_i - b)^2\right]} \tag{34}$$

V. BEYOND THE DISCRETE MODEL: TEMPORAL SUPER-RESOLUTION

Most fluorescent microscopy techniques –two-photon, confocal or scanning light-sheet– involves the sequential scanning of laser beam at different locations within the sample. Hence, for a given dwell time of the laser at each neuron position, there is a trade-off between the sampling rate and the total number of sampled neurons. In other experimental fields, resolution limitations due to recording constraints have been significantly circumvented through signal processing algorithms. For instance, super-resolution microscopy achieves imaging at higher resolution than the diffraction limit [23–27], and compressed sensing applied to MRI allows to drastically reduce the number of measurements required to reach a given resolution [28]. These algorithms rely on the hypothesis that the original signal is sparse in a certain basis - it is therefore tempting to apply them to our problem, given that neural spikes are sparse in the canonical basis. This possibility had been discussed in the context of bayesian inference [4]. Temporal resolution was shown to be slightly improved in very specific settings, *i.e.* when using prior knowledge of inputs (stimulus) and spiking history dependence of the neuronal activity. In this section, we demonstrate on synthetic data how the blind sparse deconvolution framework can be extended to support super-resolution in more generic contexts.

A. Qualitative analysis

We start off with a qualitative analysis and consider the fluorescence signal produced by an isolated single spike of amplitude n:

$$F_i = anK(\Delta t i - t_0) + \sigma \epsilon_i + b \tag{35}$$

Denoting $j = \lfloor \frac{t_0}{\Delta t} \rfloor$, $\tau = t_0 - j\Delta t \in [0, \Delta t]$, $\lambda_d = e^{-\frac{t}{\tau_d}}$, $\lambda_r = e^{-\frac{t}{\tau_r}}$, and assuming for simplicity that $b = 0$ and that K is unnormalized, we write, for $i > j$:

$$\begin{aligned} F_i &= an \left[\lambda_d^{i-j-\frac{\tau}{\Delta t}} - \lambda_r^{i-j-\frac{\tau}{\Delta t}} \right] + \sigma \epsilon_i \\ \Longleftrightarrow F_i &= an\lambda_d^{-\frac{\tau}{\Delta t}} \lambda_d^{i-j} - an\lambda_r^{-\frac{\tau}{\Delta t}} \lambda_r^{i-j} + \sigma \epsilon_i \end{aligned} \tag{36}$$

Thus, the observed fluorescence is a double exponential with non-equal coefficients of the form $f(i) = A\lambda_d^i + B\lambda_r^i$. Fitting the coefficients with a least-square method yields estimates of $an\lambda_d^{-\frac{\tau}{\Delta t}}$ and $an\lambda_r^{-\frac{\tau}{\Delta t}}$, which can be converted to estimates of τ and n. Thus, it is possible in principle to find the exact spike position in the noiseless case, if we know a priori that the signal contains a single spike. Notice that this is possible only if $\lambda_r > 0$, *i.e.* $\tau_r > 0$; if $\tau_r = 0$, the observed fluorescence is a single exponential of amplitude $an\lambda_d^{-\frac{\tau}{\Delta t}}$, and we cannot recover both n and τ without ambiguity [35]. In the case of a noisy signal, we expect that super-resolution can be achieved only if $\frac{\tau_r}{\Delta t}$ is large enough with respect to some function of σ. Notice also that if multiple spikes occur within the same time bin, the observed fluorescence transient is still a double exponential with non-equal coefficients, and it cannot be distinguished from the one produced by a single large spike at some average position. More generally, resolving two spikes in the same time bin would require the use of more complex convolution kernels.

B. Generative model

With these limitations in mind, we now extend the deconvolution framework to implement super-resolution. The fluorescence signal is constructed using a discrete generative model at a fine-grained time scale $\frac{\Delta t}{s}$, where s is a non-zero integer, which is then down-sampled by the same factor s. This yields the following generative model:

$$\begin{aligned} F_k^s &= a \sum_{j=1}^{sT} K \left[\frac{\Delta t}{s}(k - j + 1) \right] N_j^s + b + \sigma \epsilon_k^s \\ F_i \equiv F_{is}^s &= a \sum_{j=1}^{sT} K \left[i\Delta t - (j-1)\frac{\Delta t}{s} \right] N_j^s + b + \sigma \epsilon_i \\ \Leftrightarrow \mathbf{F} &= a\mathcal{K}\mathbf{N^s} + \mathbf{b} + \sigma \epsilon \end{aligned} \tag{37}$$

where F_i is the fluorescence measurement at $t_i = i\Delta t$ and $N_j^s = \int_{(j-1)\frac{\Delta t}{s}}^{j\frac{\Delta t}{s}} N(t)dt$ is the number of spikes emitted in the time interval $[(j-1)\frac{\Delta t}{s}, j\frac{\Delta t}{s}]$. [36] The convolution matrix $\mathcal{K}$ is now rectangular, of size $T \times sT$. It is not translation invariant anymore with respect to the spikes index j as the norm of the transient, $\|K_j\| = \sqrt{\sum_i \mathcal{K}_{ij}^2}$ now depends on j. Indeed, writing $j = (p-1)s + r$, we have:

$$\|K_j\| = \sqrt{\sum_{i=1}^{T} K \left[i\Delta t - (j-1)\frac{\Delta t}{s} \right]^2} = \sqrt{\sum_{i=1}^{T} K \left[(i-p)\Delta t + \frac{s+1-r}{s}\Delta t \right]^2} \approx \sqrt{\sum_{k=-\infty}^{\infty} K \left[k\Delta t + \frac{s+1-r}{s}\Delta t \right]^2} = f(r) \tag{38}$$

Typically, spikes occurring right after a fluorescence measurement (small r) have smaller $\|K_j\|$ than spikes occurring right before a measurement (large r).

C. Sparse Deconvolution

A sparse deconvolution algorithm is applied to estimate the spikes N^s:

$$\hat{\mathbf{N}}^s = \arg\min_{N^s \geq 0} \left\{ \frac{1}{2} \sum_{i=1}^{T} [F_i - a(\mathcal{K}N)_i - b]^2 + \sum_{j=1}^{Ts} \lambda^j N_j^s \right\} \tag{39}$$

Notice that, although $\mathcal{K}$ is not invertible anymore, the optimum is still well-defined because of the sparsity penalty and non-negativity constraint. Compared to Eqn. 6, the main difference is that λ is not uniform anymore: $\lambda_j \propto \|K_j\|$. This property has an important consequence, as can be seen by considering the limit case $\tau_r = 0$, $\sigma << a$. As discussed previously, a transient observed for $i \geq i_0$ can be interpreted either as a small spike right before the i_0 measurement, or a 'large' one right after the $i_0 - 1$ measurement. Thus, using a constant λ would systematically select the small spike interpretation, *i.e.* the inferred spike train would be systematically delayed with respect to the original spike train. This behavior is not desirable, and we would rather have both solutions to be degenerate global optima. This can be achieved by setting λ to a smaller value right after the $i_0 - 1$ measurement. We show in Annex C that both efficient noise filtering and unbiased estimation of spike timing for isolated spikes can be obtained with the following expression for λ^j_{BSD}:

$$\lambda^j_{BSD} = z_1 \|K_j\| \min\left(\sigma, \frac{a \frac{\sum_{r=1}^{s} \|K_r\|}{s}}{z1 + z2}\right) \tag{40}$$

In practice, the optimization can also be performed efficiently using the interior-point method [5]. Adding a small L_2 penalty $\sum_j \mu_j N_j$, with $\mu_j \propto \|K_j\|^2$ often provides better conditioning of the hessian, and faster convergence. It also ensures the unicity of the solution, in particular when $\tau_r = 0$. Figure 5a shows an example of reconstruction of a signal generated at $f_0 = 20Hz$, and sampled at $5Hz$. We observe a good agreement with the original spike train; we observe in particular that, in spite of the sparse sampling, the onsets of the green and dark curves transients are very close to one another.

D. Numerical experiments

We now test our algorithm on synthetic data generated using the model 37 at $f_0 = 500Hz$, with $\tau_r = 0.1$, $\tau_d = 0.5$, spike frequency $\nu = 2Hz$. The fluorescence signal is down sampled to recording frequencies ranging from $f = 1Hz$ to $500Hz$. Spike trains are inferred with and without super-resolution. For super-resolution, we use $s = \frac{f_0}{f}$, in order to reconstruct a spike train at the original frequency f_0. For the case without super-resolution, an inferred spike train $\hat{N}$ is first obtained at the sampling frequency f; to compare it with the original spike train at f_0, we construct a signal $\hat{N}^S$ at sampling frequency f_0 inferred signal by splitting evenly the spike counts:

$$N^S_{(i-1)s+r} \equiv \int_{(i-1)\Delta t + \frac{(r-1)\Delta t}{s}}^{i\Delta t + \frac{r\Delta t}{s}} N(t)dt \approx \frac{1}{s} \int_{(i-1)\Delta t}^{i\Delta t} N(t)dt = \frac{1}{s} N_i \forall r \in [1, s], \forall i \in [1, T] \tag{41}$$

Once the signal is inferred, we measure the performance of the reconstruction in terms of temporal accuracy. A simple measure would be the cross-correlation between the spikes and the inferred spikes:

$$X_\tau = \frac{1}{T} \sum_{i=1}^{T} N_i \hat{N}_{i+\tau} \tag{42}$$

The faster X_τ decays to 0 as $|\tau|$ increases, the more accurate the reconstruction. This simple measure works well for independent spikes, but it does not produce the expected results when the spikes are temporally correlated: in the best case scenario, the spike is perfectly recovered $N_i = \hat{N}_i \; \forall i$, and we would have $X_\tau = \frac{1}{T} \sum_{i=1}^{T} N_i N_{i+\tau} \equiv A_\tau$. Since real spike trains may have significant temporal correlations, we introduce an estimator suited both for synthetic and real data sets, whose output does not depend on the spikes correlation. We define a response function R_τ through the temporal regression model:

$$\hat{N}_t = \sum_{\tau=-m}^{m} R_\tau N(t - \tau) + \epsilon \tag{43}$$

It can be estimated as:

$$\mathbf{R} = \mathcal{A}^{-1}\mathbf{X}$$
$$\mathcal{A}_{ij} = A(i-j) \tag{44}$$

where R_τ, X_τ are indexed formally as vectors $\mathbf{R}, \mathbf{X}$. R_τ is estimated for various sampling frequencies and noise levels, and results are depicted in Figure 5 b,c. They demonstrate that super-resolution is perfectly workable at small noise levels, and that significant resolution gain can be achieved at intermediate noise level typical of actual experimental conditions. For instance, at $f = 10Hz$, $SNR = 5$, the response function width is $\sim 2\times$ smaller than without super-resolution. Figure 5c shows that significant gain in resolution can be achieved as soon as the $f \gtrsim 4Hz$. This behavior is expected, as no gain can be achieved when $f\tau_r$ is small, see Section 5 A.

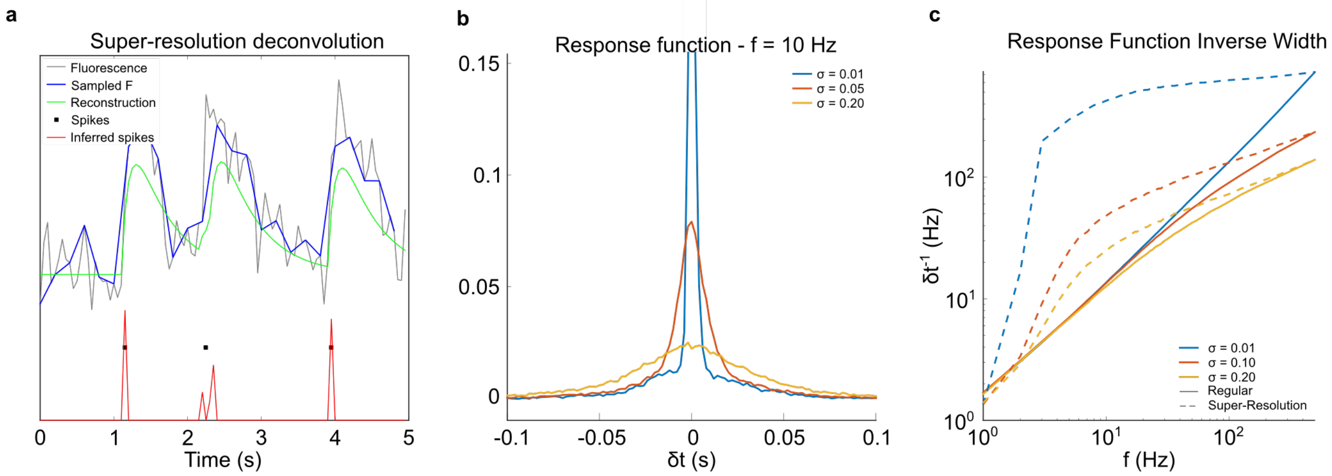

FIG. 5: (a) Example of super-resolution inference: a fluorescence signal is generated at $f_0 = 20$ Hz, sampled at 5 Hz and reconstructed at 20 Hz. Parameters: $\tau_r = 0.1$, $\tau_d = 0.5$, $\sigma = 0.2$, $a = 1$. (b) Response function at 10Hz for various noises. Notice a smaller width than the sampling interval 0.1s (c) Inverse width δt^{-1} of the point-spread function, as function of the sampling frequency for regular reconstruction (full) and SR reconstruction (dotted).

VI. EXPERIMENTS ON REAL DATA SETS

We now apply our Blind Sparse Deconvolution algorithm to real data, and compare its performance with con-oopsi.

A. Joint Electrophysiology and Fluorescence recordings on mice data

We first test it on joint electrophysiology and fluorescence recordings, as the former can be used as a ground truth for comparison. Recordings were performed on mice visual cortex at the Svoboda laboratory [17, 18], using the GCaMP5k, GCaMP6s and GCaMP6f probes, see summary in Table II. [37].For all datasets, the fluorescence is recorded at $60Hz$, and the electrophysiology at $10kHz$. For each neuron, the fluorescence signal is computed by simple averaging of the raw voxel-scale fluorescence over a region of interest. A baseline is computed using a moving percentile (window: $10s$, quantile $q = 0.15$); it is subtracted to the fluorescence trace to remove the temporal fluctuations of the baseline. The resulting pre-processed signal is then fed to the following deconvolution algorithms:

- con-oopsi.
- BSD without iterative parameter estimation.
- adaptive BSD (up to 200 iterations). Each neuron is endowed with its own rise and decay times constants.
- BSD with super-resolution and adaptive parameters (up to 200 iterations). We use $s = 5$, i.e. spikes are reconstructed at 300 Hz.

- con-oopsi, BSD, and super-resolution BSD with 'ground-truth' parameters learnt using the spike positions, see below.

For each fluorescence probe, we investigate the spike detection precision-recall and temporal accuracy. For the former, we use the area under curve metric defined in Section 2 (Time tolerance: $1/60s$); the AUC is computed separately for each recording, and a weighted average is calculated, with weights equal to the number of spikes per recording. For the latter, we estimate for each recording the response-function and compute the average; an offset and a width are estimated through a gaussian fit. The results are summarized in Tables III, IV and Figure 6.

The inference is also performed with 'ground truth' parameters τ_r, τ_d, b for each neuron. The latter are obtained by minimizing the square error, using the knowledge of the position of the spikes provided by electrophysiology, (discretized at 60Hz): [38]

$$E = \sum_i \left(F_i - \sum_j K\left[(i-j+1)\Delta t\right] a_j N_j^0 - b \right)^2 \tag{45}$$
$$(\tau_r, \tau_d, b) = \arg\min E\left(\tau_r, \tau_d, b, \mathbf{a}\right)$$

As expected from section 3A, the performance of the inference strongly depend on the inferred kernel. For the precision-recall, the main finding is that adaptive BSD always outperforms non-adaptive BSD and con-oopsi. Indeed, the uncorrelated spikes hypothesis used to estimate the kernel can be very wrong in practice, as illustrated in Figure 3a. Using the initial kernel estimates, con-oopsi and BSD misinterpret the fluorescence burst as being evoked by just a few spikes and a slow convolution kernel, whereas adaptive BSD correctly recovers the fast kernel and thus the true spike train. For all indicators, we found that inferred parameters are very similar to their ground truth values τ_r, τ_d (figure 3b), such that using the latter for the inference yields little to no improvement in performance. When using the same kernel parameters, con-oopsi and BSD in general display comparable performance. In the absence of adaptive kernel estimation, the sparsity prior λ_{BSD}, which depends on $\|K\|$, can be misestimated, resulting in weaker performance than con-oopsi. This advocates for the use of adaptive parameter estimation for robust inference.

In terms of temporal resolution, we find that GCamP6f reporter yields higher accuracy than GCaMP5k and GCaMP6s owings to its smaller rise and decay time. The measured temporal resolution can be compared quantitatively against the theoretical width depicted in Figure 2b, using the average measured noise to signal levels. The overall agreement is good, but the experimental width are $\sim 30\%$ larger than the theoretical bounds. This is likely due to the fact that real spikes are not isolated, and neighboring spikes tend to decrease the temporal accuracy.

We find that the super-resolution improves the temporal resolution by $\sim 30\%$ for both GCaMP6f and GCaMP5k, a gain which can be considered as very significant given that it is here obtained with no prior knowledge of the convolution kernel. A more modest gain is obtained for GCaMP6s, likely due to the higher noise and larger rise time associated with this dataset.

	GCaMP5k	GCaMP6s	GCaMP6f
Number of recordings	9	21	37
Total number of spikes	2735	2103	4535
Average τ_r	0.058	0.078	0.024
Average τ_d	0.52	0.88	0.38
Median Noise level $\frac{\sigma}{a}$	0.35	0.40	0.30

TABLE II: Data sets summary

Algorithm	GCaMP5k	GCaMP6s	GCaMP6f
ground-truth con-oopsi	0.592	0.524	0.832
ground-truth BSD	0.595	0.519	0.824
con-oopsi	0.346	0.424	0.743
BSD	0.349	0.335	0.698
adaptive BSD	**0.569**	**0.570**	**0.819**

TABLE III: Spike detection performance (AUC)

Algorithm	GCaMP5k	GCaMP6s	GCaMP6f
Adaptive, no SR	13 ± 22 ms	1.9 ± 23 ms	4.3 ± 16 ms
Adaptive, SR	8.4 ± 16 ms (-29%)	4.0 ± 20 ms (-11%)	1.4 ± 12 ms (-27%)
Ground truth, no SR	14 ± 22 ms	9.5 ± 26 ms	4.6 ± 16 ms
Ground truth, SR	8.7 ± 16 ms (-30%)	4.3 ± 22 ms (-19 %)	1.3 ± 11 ms (-33%)

TABLE IV: Temporal accuracy of various algorithms: offset μ and width σ of the response function (ms). The relative gain in root square distance $\sqrt{\mu^2 + \sigma^2}$ is displayed. The frame rate interval is 16.6ms

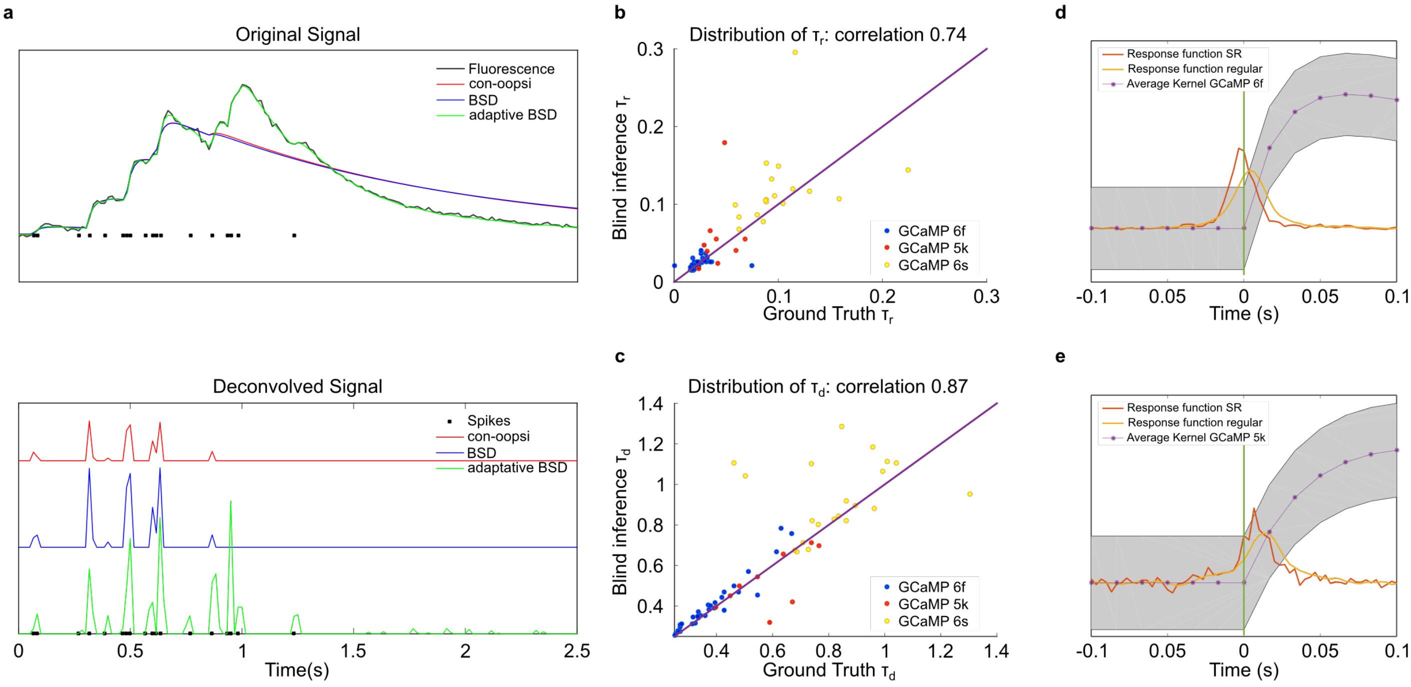

FIG. 6: (a) Results of con-oopsi, BSD and adaptive BSD on joint GCaMP6f fluorescence and EP recordings of mice [17, 18]. (b,c) Distribution of inferred rise and decay time constants (d,e) Average response function for GCaMP6f/GCaMP5k recordings as function of the time to the spike onset. The inference is performed with or without super-resolution. The discrete kernel with median rise time, decay time and signal to noise ratio is displayed for comparison.

B. Light-sheet Imaging of Zebrafish

Compared to standard fluorescence microscopy technique, such as confocal or two-photon point-scanning technique, light-sheet imaging allows for a parallelization of the recording, yielding ~ 100-fold increase in data-throughput [19**?** **?**]. When applied to zebrafish larvae, this enables simultaneous recording of the quasi-entirety of the neurons (~ 100,000 units) at typically 1 brain/second. The BSD algorithm might prove to be particularly useful for such experiments, as the size of individual datasets precludes supervision. Furthermore, the gain in speed with respect to con-oopsi should also be beneficial as it may allow one to carry out the spike inference on the fly.

To illustrate this latter claim, we test con-oopsi and BSD inference algorithms on a typical whole-brain recording, consisting of 1,800 successive volumetric stacks sampled at 1 stack/second, each of them comprising 20 z-sections. The experiment is performed on a 5 dpf larva expressing the GCaMP5 reporter panneurally. Voxels were clustered by a factor X. Hence, 255463 fluorescence traces encompassing the brain volume are processed independently. The baseline is computed as described before and the spike deconvolution is then carried out using both BSD and con-oopsi on an Intel Xeon Phi (28 cores) computer. In line with our observations of Section 2D2, we find that BSD achieves a 7-fold increase in speed compared to con-oopsi, see table V. Under these experimental conditions, the computation time with BSD match the duration of the experiment itself (20 minutes), making possible real-time spike inference. Importantly, the computation time per voxel is fairly stable with BSD, whereas some voxels use up to 200 times more time to be processed than others with con-oopsi.

These brain-scale simultaneous recordings allow one to compute the correlation of neuronal pairs activity, which might then be used to extract information regarding the large-scale functional organization of the brain. In this context, we examine whether the correlation statistics of the spike-inferred signals may be significantly different from the one computed using the raw DFF signals. For this purpose, we use a 2D recording acquired at 20 frame/second for 20 minutes in a 5dpf-old zebrafish larva expressing the genetically encoded indicator GCaMP3 (elavl3:GCaMP3). Automatic segmentation allowed us to identify 8082 individual neurons or neuropil regions of similar area, and the inference is then carried out on the ROI-averaged fluorescence traces. The rise and decay times are inferred for all neurons (see supplementary Figure). The average values of these two time-constants are then used to perform spike inference.

Figure 7a displays the time-averaged image of the brain section. Fluorescent traces and associated inferred spike trains for 5 representative neurons located in various brain regions are shown in Figure 7b. As expected, the deconvolved spike trace appear much sparser and less noisy than the original fluorescent signal. The pair-wise correlations, corrected for uniform coherent noise, are then computed for both the raw DFF signal and the inferred spike traces. We find the correlation distribution to be much more peaked after deconvolution (Figure 7c) which reflects in the more uniform appearance of the associated correlation matrix (Figure 7d).

This difference may have two possible origins. First, it may reflect the gain in temporal precision brought along by the spike inference, which may reduce the correlation of neuronal pairs that tend to discharge coherently (due to common inputs for instance), but with a slight systematic time-lag. A second explanation is related to the denoising property of the inference. In light-sheet imaging, the noise tends to display significant spatial correlation. This is notably due to the motion of small absorbing objects such as red cells that project elongated shadows and produce characteristic streaking features. Provided that these artifacts have characteristic timescales distinct from the spike-induced fluorescent transient, they are not interpreted as actual spike by BSD. This latter interpretation is confirmed by the fact that the highly negatively correlated pairs in the raw fluorescence signals are mostly confined within thin bands aligned along the beam direction (Figure 7e). For the same neuronal pairs, the correlation value computed from the inferred signal is thus largely reduced (Figure 7f).

Algorithm	Total run time	Average run time per voxel
con-oopsi	124 minutes	0.38s (min: 0.31 s, max: 110s)
BSD	18 minutes	0.051 s

TABLE V: Time for perfoming deconvolution on voxelated data

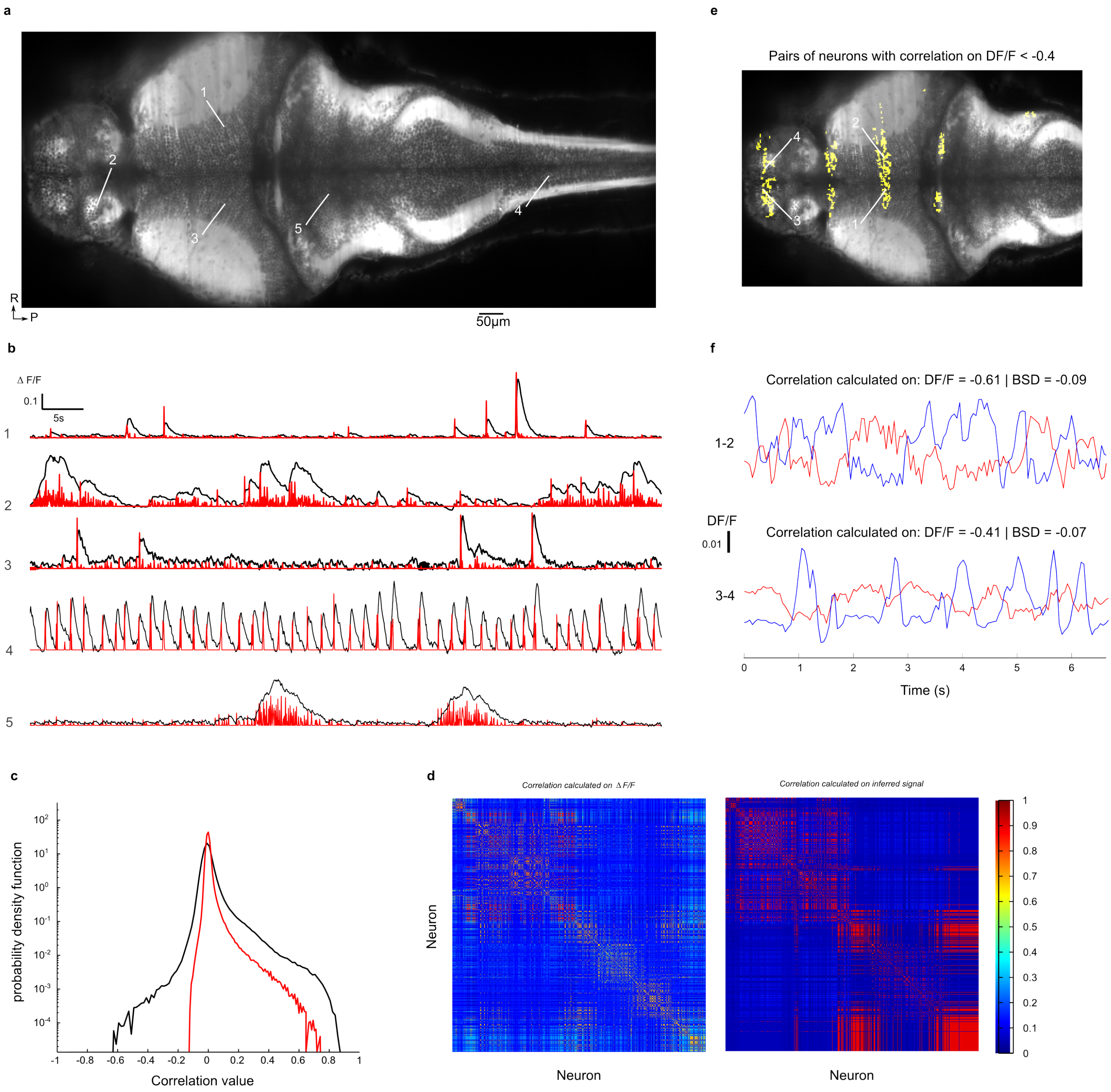

FIG. 7: (a) Bottom: Individual traces of 5 neurons recorded at 20Hz from 6dpf larvae, in black curve DF/F, in red resulting signal from BSD deconvolution algorithm. Top: Time-averaged image of a brain slice of the larva, the white arrows give the location of the 5 neurons. (b) Distribution of pair-wise correlations of DF/F (black) and signal after BSD deconvolution. The data were obtained from a 20Hz, 20 min long experiment on a 6dpf larva. (c) Time-averaged image of a brain slice of the larva . In yellow, pairs of neurons that display a correlation on DF/F inferior to -0.4 . (d) Top : Pair of neuron DF/F traces that display a pair-wise correlation calculated on DF/F of -0.61 and a pair-wise correlation calculated after BSD deconvolution of -0.09. Down : Pair of neuron DF/F traces that display a pair-wise correlation calculated on DF/F of -0.41 and a pair-wise correlation calculated after BSD deconvolution of -0.07. (e) Correlation matrix computed from DF/F. (f) Correlation matrix computed from the signal after deconvolution

Conclusion

One may expect that spike inference algorithms become an important tool for functional imaging experiments, as they allow one to recover a signal closer to the underlying spiking activity with reduced noise and better temporal resolution. However, current implementations suffer from various limitations, including long computational time, the need for arbitrary setting model parameters, the sensitivity of the outcome to experimental conditions and, more generally, a paucity of theoretical understanding of the expected performances. These drawbacks have hampered the generalization of these algorithmic methods, such that their use remain in practice relatively limited among neuroscientists.

Here we introduced a novel non-negative sparse algorithm, named Blind Sparse Deconvolution, which was designed to specifically address these issues. Compared to previous works, this fully unsupervised algorithm features higher computational speed, accuracy and adaptability. Moreover, it incorporates theoretically-grounded framework to derive estimates of the expected deconvolution performance in terms of temporal accuracy and precision-recall of the inferred spike train. These information may be used before recording as guidelines for experimental design, or a posteriori to estimate error rates. BSD also features temporal super-resolution, which we showed to significantly increase the temporal resolution on real data. As both calcium reporters and imaging methods will gain in sensitivity and speed, this capability may provide sufficient temporal precision to investigate, using calcium imaging, the role of spike-timing in extended networks dynamics. This package stands as a complement to libraries that address other challenges of calcium imaging processing, such as the spatial filtering.

[1] E. Yaksi and R. W. Friedrich, "Reconstruction of firing rate changes across neuronal populations by temporally deconvolved ca2+ imaging," *Nature Methods*, vol. 3, no. 5, pp. 377–383, 2006.

[2] 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.

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

[4] J. T. Vogelstein, B. O. Watson, A. M. Packer, R. Yuste, B. Jedynak, and L. Paninski, "Spike inference from calcium imaging using sequential monte carlo methods," *Biophysical journal*, vol. 97, no. 2, pp. 636–655, 2009.

[5] J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *Journal of neurophysiology*, vol. 104, no. 6, pp. 3691–3704, 2010.

[6] 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.

[7] Y. Mishchencko, J. T. Vogelstein, and L. Paninski, "A bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data," *The Annals of Applied Statistics*, pp. 1229–1261, 2011.

[8] E. A. Pnevmatikakis, J. Merel, A. Pakman, and L. Paninski, "Bayesian spike inference from calcium imaging data," in *Asilomar Conference on Signals, Systems and Computers, 2013*, 2013.

[9] H. Lütcke, F. Gerhard, F. Zenke, W. Gerstner, and F. Helmchen, "Inference of neuronal network spike dynamics and topology from calcium imaging data.," *Frontiers in neural circuits*, vol. 7, p. 201, jan 2013.

[10] T. Deneux, A. Kaszas, G. Szalay, G. Katona, T. Lakner, A. Grinvald, B. Rózsa, and I. Vanzetta, "Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo," *Nature Communications*, vol. 7, 2016.

[11] L. Theis, P. Berens, E. Froudarakis, J. Reimer, M. R. Rosón, T. Baden, T. Euler, A. S. Tolias, and M. Bethge, "Benchmarking spike rate inference in population calcium imaging," *Neuron*, vol. 90, no. 3, pp. 471–482, 2016.

[12] M. A. Picardo, J. Merel, K. A. Katlowitz, D. Vallentin, D. E. Okobi, S. E. Benezra, R. C. Clary, E. A. Pnevmatikakis, L. Paninski, and M. A. Long, "Population-level representation of a temporal sequence underlying song production in the zebra finch," *Neuron*, vol. 90, no. 4, pp. 866–876, 2016.

[13] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, *et al.*, "Simultaneous denoising, deconvolution, and demixing of calcium imaging data," *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.

[14] J. Friedrich, P. Zhou, and L. Paninski, "Fast active set methods for online deconvolution of calcium imaging data," *arXiv preprint arXiv:1609.00639*, 2016.

[15] J. Friedrich, W. Yang, D. Soudry, Y. Mu, M. B. Ahrens, R. Yuste, D. S. Peterka, and L. Paninski, "Multi-scale approaches for high-speed imaging and analysis of large neural populations," *bioRxiv*, p. 091132, 2016.

[16] A. Kazemipour, J. Liu, P. Kanold, M. Wu, and B. Babadi, "Efficient Estimation of Compressible State-Space Models with Application to Calcium Signal Deconvolution," oct 2016.

[17] J. Akerboom, T.-W. Chen, T. J. Wardill, L. Tian, J. S. Marvin, S. Mutlu, N. C. Calderón, F. Esposti, B. G. Borghuis, X. R. Sun, *et al.*, "Optimization of a gcamp calcium indicator for neural activity imaging," *The Journal of Neuroscience*,

vol. 32, no. 40, pp. 13819–13840, 2012.

[18] T.-W. Chen, T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R. Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman, *et al.*, "Ultrasensitive fluorescent proteins for imaging neuronal activity," *Nature*, vol. 499, no. 7458, pp. 295–300, 2013.

[19] T. Panier, S. A. Romano, R. Olive, T. Pietri, G. Sumbre, R. Candelier, and G. Debrégeas, "Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy," *Frontiers in neural circuits*, vol. 7, 2013.

[20] S. Wolf, W. Supatto, G. Debrégeas, P. Mahou, S. G. Kruglik, J.-M. Sintes, E. Beaurepaire, and R. Candelier, "Whole-brain functional imaging with two-photon light-sheet microscopy," *Nature methods*, vol. 12, no. 5, pp. 379–380, 2015.

[21] I. Selesnick, "Sparse deconvolution (an mm algorithm).," *Connexions*, 2012.

[22] B. A. Wilt, J. E. Fitzgerald, and M. J. Schnitzer, "Photon Shot Noise Limits on Optical Detection of Neuronal Spikes and Estimation of Spike Timing," *Biophysical Journal*, vol. 104, no. 1, pp. 51–62, 2013.

[23] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature methods*, vol. 3, no. 10, pp. 793–796, 2006.

[24] B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy," *Science*, vol. 319, no. 5864, pp. 810–813, 2008.

[25] M. Fernández-Suárez and A. Y. Ting, "Fluorescent probes for super-resolution imaging in living cells," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 12, pp. 929–943, 2008.

[26] M. Heilemann, S. Van De Linde, M. Schüttpelz, R. Kasper, B. Seefeldt, A. Mukherjee, P. Tinnefeld, and M. Sauer, "Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes," *Angewandte Chemie International Edition*, vol. 47, no. 33, pp. 6172–6176, 2008.

[27] S. W. Hell, "Microscopy and its focal switch," *Nature methods*, vol. 6, no. 1, pp. 24–32, 2009.

[28] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing mri," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[29] The choice of the convention $\mathcal{K}_{ij} = K\left[\Delta t(i-j+1)\right]$ instead of $\mathcal{K}_{ij} = K\left[\Delta t(i-j)\right]$ ensures that $\mathcal{K}_{ij} > 0 \forall j \geq i, \mathcal{K}_{ij} = 0 \forall j < i$. Thus, N_i is the count of spikes occuring *after* measurement F_{i-1} and *before* measurement F_i

[30] Indeed, $\exists \alpha, \beta, \forall t > 0, \sum_{t_l} K(t-t_l) = \sum_{t_l>0} K(t-t_l) + \alpha e^{-\frac{t}{\tau_r}} + \beta e^{-\frac{t}{\tau_d}}$. We assume here that $\alpha = \beta = 0$ but we could treat them as unknown variables to be inferred

[31] We go beyond this approximation in section 6, when discussing super-resolution

[32] We approximate the FNR as the probability that the Dirac solution at $i = i_0$ is zero; both probabilities are not strictly equal because there is a small probability that this solution is zero but other solutions are non-zero

[33] We stick with the discrete generative model for simplicity in this subsection

[34] In practice, the auto-correlation matrix obtained is not necessarily definite positive, because the estimate of σ can be incorrect - this can lead to very bad estimates of τ_r, τ_d. To mitigate this issue, we subtract $\min(\sigma, \lambda_{min})$, where λ_{min} is the smallest eigenvalue .

[35] In [4], the authors assume $\tau_r = 0$ and that a is fixed

[36] The spikes occurring between measure $i-1$ and measure i are the $N_{(i-1)+r} \forall r \in [1, s]$

[37] Available online at https://crcns.org/

[38] Note that we relax the hypothesis that all transients have the same amplitude a, and optimize over all the amplitudes a_j. This is particularly important for spike bursts, where strong non-linear effects are observed.

Annex A: Stability of the single spike solution and the half-spike problem

In Section 2, we have not studied the stability of the single-spike solution. We study it here, and discuss when it is a global optimum. Assuming $\hat{N}_i = \delta_{i,i_0} \max\left\{a - \frac{\lambda}{\sum_i K^2(t)} + \sigma \frac{\sum_i K(t)\epsilon_i}{\sum_i K^2(t)}, 0\right\} > 0$ and looking for the stability of the solution, w.r.t the other coordinates, we find:

$$-\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}} = \sigma \sum_i \left\{K(\Delta t(i - i_0 - \delta)) - K(\Delta t(i - i_0) \cos(\theta(\delta \Delta t)\right\} \epsilon_i - \lambda(1 - \cos(\theta(\delta \Delta t))) \tag{46}$$

$$P\left(-\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}} > 0\right) = \Phi\left[\frac{\lambda \tan \frac{\theta(\delta \Delta t)}{2}}{\sigma \|K\|}\right] \tag{47}$$

Where $\Phi(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$. Therefore, the Dirac solution is stable only if the above probability is small enough for all values of δ. Far away from the spike $\delta \to \infty$, the angle $\theta_\delta \to \frac{\pi}{2}$ and we recover $P = \Phi\left[\frac{\lambda}{\sigma\|K\|}\right]$, as in the spikeless signal. On the other hand, the smaller δ, the smaller θ_δ and the probability is higher. For $\lambda = \lambda_{BSD}$ and low noise, the above probability reduces to $P = \Phi\left[z_1 \tan\frac{\theta_\delta}{2}\right]$; the Dirac solution can become unstable. In practice, the result depends on the level of noise: for low σ, the optimum remains close to the Dirac solution, whereas for high noise, we can find 'half-spikes' solutions, of the form $N_i = \frac{an}{2}(\delta_{i,i_0} + \delta_{i,i_0\pm 1})$

Annex B: Kernel inference: proof of convergence and fast algorithm

We prove here that for isolated spikes and small noise, the cost function $\mathcal{L}(\mathbf{N},\mathbf{K}) = \frac{1}{2}\|\mathbf{F} - \mathcal{K}\mathbf{N}\|^2 + \lambda(\mathbf{K})1^T\mathbf{N}$ admits solution $K = K^0$ as local minimum. Denoting $\hat{\mathbf{N}} = \arg\min_{N\geq 0}\mathcal{L}(\mathbf{N},\mathbf{K})$

For a signal with a single spike $F_i = aK\left[\Delta t(i - i_0 + 1)\right] + \sigma\epsilon_i$, if the noise is small and K is close enough to K_0, we have: $\hat{N}_i = an\delta_{i,i_0}$, $\lambda = z\sigma\|K\|$. Optimizing over n yields:

$$\begin{aligned}
n &= \max\left\{\frac{\cos\phi_K\|K^0\|}{\|K\|} + \frac{\sigma}{a\|K\|}(\tilde{\epsilon}_1 - z), 0\right\}\\
\mathcal{L}(\hat{\mathbf{N}},\mathbf{K}) &= \frac{1}{2}a^2\|K^0\|^2(1-\cos^2\theta_K) + za\sigma\|K^0\|\cos\theta_K\\
&+ \frac{\sigma^2}{2}(\sum_i \epsilon_i^2 - \tilde{\epsilon}_1^2 - z^2) + \sigma^2 z\tilde{\epsilon}_1 - a\sigma\|K^0\|\sqrt{2(1-\cos\phi_K)}\tilde{\epsilon}_2
\end{aligned} \tag{48}$$

Where:

$$\begin{aligned}
\|K\| &= \sqrt{\sum_l K(l\Delta t)^2}\\
\|K^0\| &= \sqrt{\sum_l K^0(l\Delta t)^2}\\
\cos\phi_K &= \frac{\sum_i K^0\left[(i-i_0+1)\Delta t\right]K\left[(i-i_0+1)\Delta t\right]}{\|K\|\|K^0\|}
\end{aligned} \tag{49}$$

Note that we recover Eqn. 17 when $K = K^0$. For a signal of multiple isolated spikes $F_i = a\sum_l K\left[\Delta t(i - i_l + 1)\right] + \sigma\epsilon_i$, with $|i_l - i_l'| >> \frac{\tau_d+\tau_r}{\Delta t}$, a similar solution $\hat{N}_i = \sum_l an_l\delta_{i,i_l}$ can be derived, and $\mathcal{L}$ is self averaging:

$$\mathcal{L}(\hat{\mathbf{N}},\mathbf{K}) \propto <\mathcal{L}(\hat{\mathbf{N}},\mathbf{K})> \propto \frac{1}{2}a^2\|K^0\|^2(1-\cos^2\phi_K) + za\sigma\|K^0\|\cos\phi_K + \text{Constant} \tag{50}$$

Hence, the function depends on $\mathbf{K}$ only through $\cos\theta_K$. One can check that when $\frac{z\sigma}{a\|K\|} < 1$, the minimum is reached at $\cos\phi_K = 1$, *i.e.* $K = K^0$. This concludes the proof. Although we can not prove more about the radius of convergence, good convergence was achieved in practice after starting from the initialization.

In practice, the optimization with respect to K can be performed efficiently using standard temporal regression tricks. Observe that:

$$\begin{aligned}
\frac{1}{2}\|F - KN\|^2 &= \frac{1}{2}(F^TF - 2F^T\mathcal{K}N + N^T\mathcal{K}^T\mathcal{K}N)\\
&= \frac{1}{2}\left(F^TF - 2\text{Trace}\left[\mathcal{K}NF^T\right] + \text{Trace}\left[(\mathcal{K}^T\mathcal{K})NN^T\right]\right)\\
&= \frac{1}{2}\left\{\sum_{i=1}^{T}F_i^2 - 2\sum_{l=0}^{\infty}K(\Delta t(l+1))\left(\sum_{i=1}^{T-l}F_{i+l}N_i\right)\right.\\
&+ \sum_{l=-\infty}^{\infty}\left(\sum_{i=\max(1,1-l)}^{\min(T,T-l)}N_{i+l}N_i\right)\left(\sum_{j=-\infty}^{\infty}K\left[\Delta t j\right]K\left[\Delta t(j+l)\right]\right)\\
&\left.- \sum_{j=T+1}^{\infty}\left(\sum_i K\left[(j-i+1)\Delta t\right]N_i\right)^2\right\}
\end{aligned} \tag{51}$$

To go from the second line to the third line, we used the translation invariance property of $\mathcal{K}$, the causality of $\mathcal{K}$ ($\mathcal{K}_{ij} = 0 \forall j \geq i$) and wrote $\sum_{l=1}^{T} \mathcal{K}_{li}\mathcal{K}_{lj} = \sum_{l=-\infty}^{\infty} \mathcal{K}_{li}\mathcal{K}_{lj} - \sum_{l=T+1}^{\infty} \mathcal{K}_{li}\mathcal{K}_{lj}$ Hence, $\mathcal{L}(\mathbf{N}, \mathbf{K})$ depends on N and F only through:

- the sums $S_1 = \sum_{i=1}^{T} N_i$ and $S_2 = \sum_{i=1}^{T} F_i^2$
- the unnormalized cross-correlation between fluorescence and inferred spikes $X(l) = \sum_{i=1}^{T-l} F_{i+l} N_i$.
- the unnormalized autocorrelation function of the inferred spikes $A(l) = \sum_{i=\max(1,1-l)}^{\min(T,T-l)} N_{i+l} N_i$.
- the boundary term $\sum_{j=T+1}^{\infty} \left(\sum_{i=1}^{T} K\left[(j-i+1)\Delta t\right] N_i \right)^2$

The first three terms can be precomputed in $\mathcal{O}(T)$ once for all, and the second and third up to a cutoff $l_{\max} \sim \lfloor 5 \frac{\tau_r + \tau_d}{\Delta t} \rfloor$, such that $K(l_{\max}) << 1$. The last one can be computed in $\mathcal{O}(l_{\max})$, by noting that after T, the convolved spikes is a double exponential, with coefficients depending on the $\sim l_{\max}$ last time bins. Overall, the cost function can be evaluated in $\mathcal{O}(l_{\max})$ and optimized efficiently.

Annex C: Detailed computations for the response function estimation

We assume a noisy single spike signal, $F_i = aK\left[\Delta t i - t_0\right] + \sigma\epsilon_i$, where we write formally $t_0 = \Delta t\left(i_0 - 1 + r_0\right)$, with $r_0 \in [0, 1[$; *i.e.* the spike is emitted before measurement i_0. The likelihood27 becomes:

$$\begin{aligned}
N_i &= an\delta_{i,i_0+\delta} \\
\mathcal{L}(n,\delta) &= \frac{1}{2}\sum_i F_i^2 + \frac{a^2}{2}\left\{-2n\sum_i K\left[\Delta t(i-i_0+1-r)\right]K\left[\Delta t(i-i_0+1-\delta)\right]\right. \\
&\left. +n^2\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]^2\right\} - a\sigma\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]\epsilon_i + \lambda an \\
n^\delta = \arg\max_{n\geq 0}\mathcal{L}(n,\delta) &= \frac{\min\left\{\sum_i K\left[\Delta t(i-i_0+1-r)\right]K\left[\Delta t(i-i_0+1-\delta)\right] + \frac{\sigma}{a}\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]\epsilon_i - \frac{\lambda}{a}, 0\right\}}{\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]^2} \\
\mathcal{L}(n^\delta,\delta) = \frac{1}{2}\sum_i F_i^2 &+ \frac{\min\left\{\sum_i K\left[\Delta t(i-i_0+1-r)\right]K\left[\Delta t(i-i_0+1-\delta)\right] + \frac{\sigma}{a}\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]\epsilon_i - \frac{\lambda}{a}, 0\right\}}{2\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]^2}
\end{aligned} \tag{52}$$

In the last expression, the term $\rho_{r,\delta} = \sum_i K\left[\Delta t(i-i_0+1-r)\right]K\left[\Delta t(i-i_0+1-\delta)\right]$ can be computed analytically for all δ and r and is independent of i_0; the term $\sum_i K\left[\Delta t(i-i_0+1-\delta)\right]^2$ is the usual $\|K\|^2$ and the term involving noise can be rewritten by introducing new, correlated gaussian noises:

$$\begin{aligned}
\tilde{\epsilon}_\delta &= \sum_i K\left[\Delta t(i-i_0+1-\delta)\right]\epsilon_i \\
< \tilde{\epsilon}_\delta > &= 0 \\
< \tilde{\epsilon}_\delta \tilde{\epsilon}_{\delta'} > &= \sum_i K\left[\Delta t(i-i_0+1-\delta)\right]K\left[\Delta t(i-i_0+1-\delta')\right] \\
\mathcal{L}(n^\delta,\delta) &= \frac{1}{2}\sum_i F_i^2 + \frac{\min\left\{\rho_{r,\delta} + \frac{\sigma}{a}\tilde{\epsilon}_\delta - \frac{\lambda}{a}, 0\right\}}{2\|K\|^2}
\end{aligned} \tag{53}$$

For a given r and noise realization, we can thus compute the optimal δ - and by Monte Carlo averaging, we obtain an estimate of the probability distribution $P(\delta|r)$. To obtain a response function in continuous time, it is then transformed into a continuous piecewise-constant probability density through: $P^c(\delta^c \in \mathbb{R}|r) = \frac{P(\lfloor\delta^c\rfloor|r)}{\Delta t}$

And the overall response function is obtained by averaging over r, yielding:

$$R(\delta^c) = \int_{r=0}^{1} P^c(\delta^c + r|r)$$

In practice, R and r are computed over a discrete grid of the form $k\frac{\Delta t}{s}$.

For the super-resolution case, the computation is almost the same; the only difference being that we reconstruct the spikes with a thinner resolution.

Annex D: Proof of unbiased estimation for super-resolution

We show here that the choice $\lambda_j = z\sigma\|K_j\|$ is best suited for an unbiased (in time) reconstruction of the spikes. We consider again the single-spike setting, with a single spike of a at position $k_0 = (i_0 - 1)s + r_0$, for which $F_i = aK\left[\Delta t(i - i_0) + \frac{\Delta t(2-r_0)}{s}\right] + \sigma\epsilon_i$.

We now look for optima of $\mathcal{L}(\mathbf{N},\mathbf{K})$ of the form $\hat{N}_{(i-1)s+r} = an\delta_{i,i_0+\Delta}\delta_{r,r_0+\delta}$. Note that instead of doing this computation, we can simply observe that it is a special case of Annex B, using reference kernel $K^0(t) \equiv K(t)$, and measurement kernel $K(t) \equiv K(t - \Delta\Delta t - \frac{\delta\Delta t}{s})$.

$$\begin{aligned} n &= \max\left\{\frac{\|K_{r_0}\|}{\|K_{r_0+\delta}\|}\cos\theta_{\Delta,\delta} + \frac{\sigma}{a}\|K_{r_0+\delta}\|(\tilde{\epsilon_1} - z)\right\} \\ \mathcal{L}(\hat{\mathbf{N}},\mathbf{K}) &= \frac{1}{2}\|K_{r_0}\|^2(1-\cos^2\theta_{\Delta,\delta}) + za\sigma\|K_r\|\cos\theta_{\Delta,\delta} \\ &+ \frac{\sigma^2}{2}(\sum_i \epsilon_i^2 - \tilde{\epsilon_1}^2 - z^2) + \sigma^2 z\tilde{\epsilon}_1 - a\sigma\|K_r\|\sqrt{2(1-\cos\theta_{\Delta,\delta})}\tilde{\epsilon}_2 \end{aligned} \tag{54}$$

Where:

$$\begin{aligned} \|K_r\| &= \sqrt{\sum_l K\left[l\Delta t + \frac{\Delta t(s+1-r)}{s}\right]^2} \\ \cos\theta_{\Delta,\delta} &= \frac{\sum_i K\left[i\Delta t + \frac{\Delta t(s+1-r)}{s}\right] K\left[(i+\Delta)\Delta t + \frac{\Delta t(s+1-r')}{s}\right]}{\|K_r\|\|K_{r'}\|} \end{aligned} \tag{55}$$

And $\tilde{\epsilon_1}$, $\tilde{\epsilon_2}$ are gaussian noises of variance unity (see Annex B). Thus, the optimum over Δ, δ is with highest probability $\delta = \Delta = 0$, and the estimator is unbiased. Note that this result is expected: using the equivalence with a LASSO regression developed in Sec. 5, we know that the coefficients (here, the spikes) are correctly estimated with a uniform λ only when the features (Here, K) are normalized to unity $\sum_i K_{ij}^2 = 1 \forall j$.

Annex E: Kernel inference in the super-resolution setting

Since the convolution matrix $\mathcal{K}$ is not fully translation invariant in the super-resolution setting, the estimation of the kernel is slightly different. For the initial estimation, Eqn. 28 becomes:

$$\begin{aligned} A_F(l) - \sigma^2\delta_{l,0} &= a^2\nu\frac{\Delta t}{s}\sum_{k=1}^{Ts} K\left[\Delta t i - \frac{\Delta t}{s}(k-1)\right] K\left[\Delta t(i+l) + \frac{\Delta t}{s}(k-1)\right] \\ &= a^2\nu\Delta t\sum_{j=1}^{T}\frac{1}{s}\left(\sum_{r=1}^{s} K\left[\Delta t(i-j-1) + \frac{\Delta t(s+1-r)}{s}\right] K\left[\Delta t(i+l-j-1) + \frac{\Delta t(s+1-r)}{s}\right]\right) \\ &\approx a^2\nu\Delta t\sum_{m=-\infty}^{\infty}\frac{1}{s}\left(\sum_{r=1}^{s} K\left[\Delta t m + \frac{\Delta t(s+1-r)}{s}\right] K\left[\Delta t(m+l) + \frac{\Delta t(s+1-r)}{s}\right]\right) \end{aligned} \tag{56}$$

For $s > 1$, this formula is different from Eqn. 28. It can be shown (see Annex E) that the right-hand side has a well-defined limit when $s \to \infty$, ie in the continuous setting.

Similarly, the iterative kernel update Eqn. 51 is different:

$$\begin{aligned}\frac{1}{2}\|F-KN\|^2 = \frac{1}{2}\Bigg\{&\sum_{i=1}^{T}F_i^2 - 2\sum_{l=-\infty}^{\infty}\sum_{r=1}^{s}K\left[\Delta t l+\frac{s+1-r}{s}\right]\left(\sum_{i=1}^{T}F_{i+l}N_{s(i-1)+r}\right)\\ &+\sum_{l=-\infty}^{\infty}\sum_{r=1}^{s}\sum_{r'=1}^{s}\left(\sum_{m=-\infty}^{\infty}K\left[\Delta t m+\frac{\Delta t(s+1-r)}{s}\right]K\left[\Delta t(m+l)+\frac{\Delta t(s+1-r')}{s}\right]\right)\left(\sum_{i=1}^{T}N_{(i+l-1)s+r}N_{(i-1)s+r'}\right)\Bigg\}\end{aligned} \tag{57}$$

The sparsity penalty becomes:

$$\lambda^T N = \sum_{j=1}^{sT}\lambda_j N_j = z\sigma\sum_{r=1}^{s}\sqrt{\sum_{m=-\infty}^{\infty}K\left[m\Delta t+\frac{(s+1-r)\Delta t}{s}\right]}\left(\sum_{i=1}^{T}N_{(i-1)s+r}\right) \tag{58}$$

Hence, $\mathcal{L}(\mathbf{N},\mathbf{K})$ now depends on $\mathbf{F}$ and $\mathbf{N}$ through the following quantities:

- the sum $S_2 = \sum_{i=1}^{T}$
- the sums vector $S_1(r) = \left(\sum_{i=1}^{T} N_{(i-1)s+r}\right)$
- the cross-correlation matrix $X(l,r) = \left(\sum_{i=1}^{T} F_{i+l}N_{s(i-1)+r}\right)$
- the autocorrelation tensor $A(l,r,r') = \left(\sum_{i=1}^{T} N_{(i+l-1)s+r}N_{(i-1)s+r'}\right)$

Altogether, the cost function can be evaluated relatively fast. Note that the complexity of the kernel optimization is now $\mathcal{O}(l_{\max}s^2)$.

Annex F: Various explicit formulas for the double exponential kernel

Various useful formulas for blind sparse deconvolution are consigned, here for double exponential kernels.

Kernel normalization. We normalize K such that $\max_{t\geq 0} K(t) = 1$. This gives:

$$\begin{aligned}K(t) &= \frac{1}{M(\tau_r,\tau_d)}\left[e^{-\frac{t}{\tau_d}} - e^{-\frac{t}{\tau_r}}\right]1_{t\geq 0}\\ M(\tau_r,\tau_d) &= \left(\frac{\tau_r}{\tau_d}\right)^{-\frac{\tau_r}{\tau_d-\tau_r}} - \left(\frac{\tau_r}{\tau_d}\right)^{\frac{\tau_d}{\tau_d-\tau_r}}\end{aligned} \tag{59}$$

Kernel norms. The L_1 and L_2 norms are computed as follow:

$$\begin{aligned}\lambda_d &= e^{-\frac{\Delta t}{\tau_d}}\\ \lambda_r &= e^{-\frac{\Delta t}{\tau_r}}\\ \|K\| \equiv \sqrt{\sum_{i=-\infty}^{\infty}K\left[\Delta t i\right]^2} &= \frac{1}{M(\tau_r,\tau_d)}\sqrt{\frac{\lambda_d^2}{1-\lambda_d^2} - \frac{2\lambda_d\lambda_r}{1-\lambda_d\lambda_r} + \frac{\lambda_r^2}{1-\lambda_r^2}}\\ \|K\|_1 \equiv \sum_{i=-\infty}^{\infty}K\left[\Delta t i\right] &= \frac{1}{M(\tau_r,\tau_d)}\left(\frac{\lambda_d}{1-\lambda_d} + \frac{\lambda_r}{1-\lambda_r}\right)\end{aligned} \tag{60}$$

Kernel norms for super-resolution. The L_1 and L_2 norms for a spike emitted at time $(j-1)s+r$, $r \in [1,s]$ are given by:

$$\begin{aligned}\|K_r\| &= \frac{1}{M(\tau_r,\tau_d)}\sqrt{\sum_i K\left[\Delta i + \frac{\Delta t(s+1-r)}{s}\right]^2} = \sqrt{\frac{\lambda_d^{2^{\frac{s+1-r}{s}}}}{1-\lambda_d^2} - 2\frac{(\lambda_d\lambda_r)^{\frac{s+1-r}{s}}}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{2^{\frac{s+1-r}{s}}}}{1-\lambda_r^2}} \\ \|K_r\|_1 &= \frac{1}{M(\tau_r,\tau_d)}\frac{\lambda_d^{\frac{s+1-r}{s}}}{1-\lambda_d} + \frac{\lambda_r^{\frac{s+1-r}{s}}}{1-\lambda_r}\end{aligned} \tag{61}$$

Kernel overlaps Useful for assessing temporal uncertainty and for kernel estimation

$$\cos\theta(\delta\Delta t) \equiv \frac{\sum_{i=-\infty}^{+\infty} K\left[i\Delta t\right] K\left[(i+\delta)\Delta t\right]}{\|K\|^2} = \frac{\frac{\lambda_d^{2+\delta}}{1-\lambda_d^2} - \frac{(\lambda_d^\delta+\lambda_r^\delta)\lambda_d\lambda_r}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{2+\delta}}{1-\lambda_r^2}}{\frac{\lambda_d^2}{1-\lambda_d^2} - \frac{2\lambda_d\lambda_r}{1-\lambda_d\lambda_r} + \frac{\lambda_r^2}{1-\lambda_r^2}} \tag{62}$$

Boundary term The estimation of the kernel involves the computation of the following boundary term:

$$\begin{aligned}&\sum_{j=T+1}^{\infty}\left(\sum_{i=1}^{T} K\left[(j-i+1)\Delta t\right] N_i\right)^2 = \frac{1}{M(\tau_r,\tau_d)^2}\sum_{j=T+1}^{\infty}\left(\lambda_d^{j-T}\left[\sum_i \lambda_d^{T-i+1}N_i\right] - \lambda_r^{j-T}\left[\sum_i \lambda_r^{T-i+1}N_i\right]\right)^2 \\ &= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^4\left(\sum_i \lambda_d^{T-i}N_i\right)^2}{1-\lambda_d^2} - \frac{2(\lambda_d\lambda_r)^2\left(\sum_i \lambda_r^{T-i}N_i\right)\left(\sum_i \lambda_d^{T-i}N_i\right)}{1-\lambda_d\lambda_r} + \frac{\lambda_r^4\left(\sum_i \lambda_r^{T-i}N_i\right)^2}{1-\lambda_r^2}\right)\end{aligned} \tag{63}$$

Kernel overlaps for super-resolution Useful for assessing temporal uncertainty and for kernel estimation

$$\begin{aligned}A_K(l,r_1,r_2) &= \sum_i K\left[(i+l+\frac{s+1-r_1}{s})\Delta t\right] K\left[(i+\frac{s+1-r_2}{s})\Delta t\right] \\ &= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^{l+\frac{2(s+1)-r_1-r_2}{s}}}{1-\lambda_d^2} - \frac{\lambda_d^{l+\frac{s+1-r_1}{s}}\lambda_r^{\frac{s+1-r_2}{s}} + \lambda_r^{l+\frac{s+1-r_1}{s}}\lambda_d^{\frac{s+1-r_2}{s}}}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{l+\frac{2(s+1)-r_1-r_2}{s}}}{1-\lambda_r^2}\right)\end{aligned} \tag{64}$$

In particular:

$$\begin{aligned}\frac{1}{s}\sum_{r=1}^{s} A_K(0,r,r) &= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\phi_s(\lambda_d^2)}{1-\lambda_d^2} - \frac{\phi_s(\lambda_d\lambda_r)}{1-\lambda_d\lambda_r} + \frac{\phi_s(\lambda_r^2)}{1-\lambda_r^2}\right) \\ \phi_s(x) &= \frac{1-x}{s(-1+x^{-\frac{1}{s}})}\end{aligned} \tag{65}$$

Boundary-term for super-resolution

$$\begin{aligned}&\sum_{i=T+1}^{\infty}\left(\sum_{j=1}^{sT} K\left[i\Delta t - \frac{(j-1)\Delta t}{s}\right] N_j\right)^2 = \frac{1}{M(\tau_r,\tau_d)^2}\sum_{i=T+1}^{\infty}\left(\lambda_d^{i-T}\left[\sum_i \lambda_d^{\frac{Ts-j+1}{s}}N_j\right] - \lambda_r^{i-T}\left[\sum_i \lambda_r^{\frac{Ts-j+1}{s}}N_j\right]\right)^2 \\ &= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^{\frac{4}{s}}\left(\sum_j \lambda_d^{\frac{Ts-j}{s}}N_j\right)^2}{1-\lambda_d^2} - \frac{2(\lambda_d\lambda_r)^{\frac{2}{s}}\left(\sum_j \lambda_r^{\frac{Ts-j}{s}}N_j\right)\left(\sum_j \lambda_d^{\frac{Ts-j}{s}}N_j\right)}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{\frac{4}{s}}\left(\sum_j \lambda_r^{\frac{Ts-j}{s}}N_j\right)^2}{1-\lambda_r^2}\right)\end{aligned} \tag{66}$$

Annex G: Heterogeneity in rise and decay time constants in Zebrafish

Application of BSD to zebrafish data yields heterogeneous distributions of rise and decay times. This means that different regions show different patterns of fluorescence bursts. We see that the heterogeneities have a spatial structure: in particular neurons in X tend to have longer time constants, whereas neurons in have shorter time constants. The two possible explanations are that the spike patterns are different in these regions (*e.g.*, regular vs sparse spike trains), and/or that the expression of GCaMP is significatively different. Overall, they motivate the use of heterogeneous time constants.

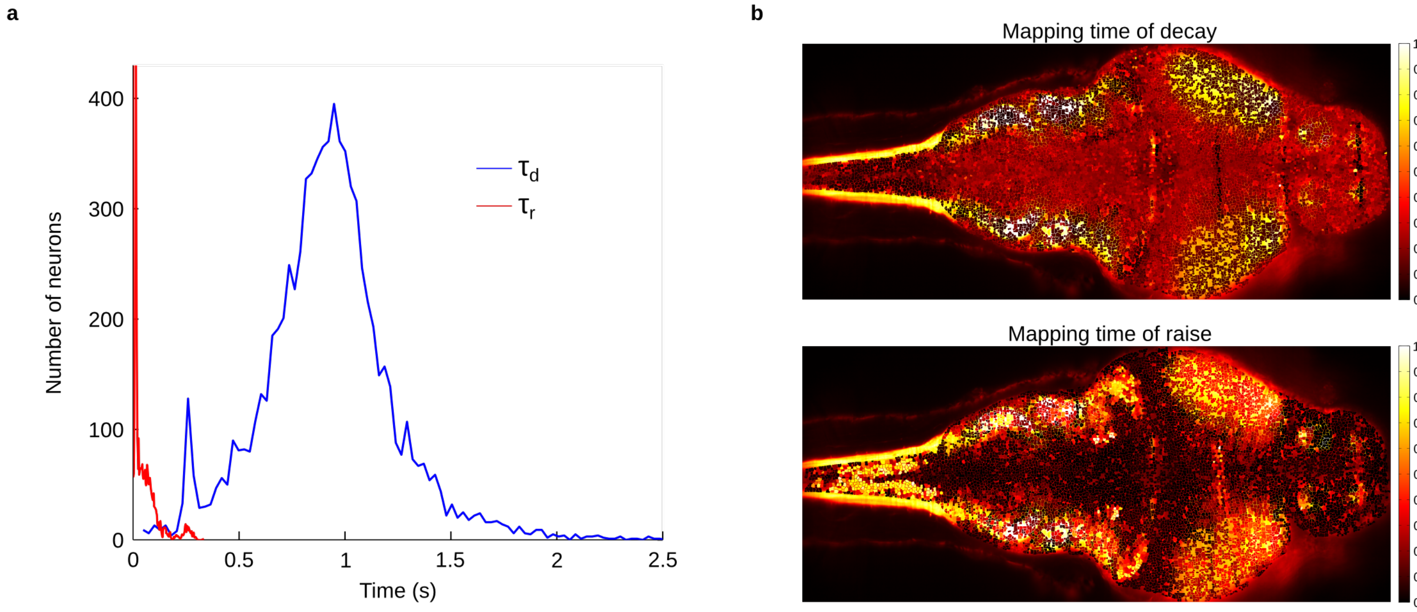

FIG. 8: (a) Distribution of rise and decay time. (b) Mapping of rise and decay time across a neurons