Journal of Neuroscience Methods 342 (2020) 108763

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Blind deconvolution for spike inference from fluorescence recordings

Jérôme Tubiana[a], Sébastien Wolf[b,c], Thomas Panier[d], Georges Debregeas[d,*]

[a] Blavatnik School of Computer Science, Tel Aviv University, Israel
[b] Laboratoire de Physique de l'Ecole Normale Supérieure, CNRS UMR 8023 & PSL Research, France
[c] Institut de Biologie de l'Ecole Normale Supérieure, CNRS, INSERM, UMR 8197 & PSL Research, France
[d] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine (IBPS), Laboratoire Jean Perrin (LJP), France

ABSTRACT

The parallel developments of genetically-encoded calcium indicators and fast fluorescence imaging techniques allows one to simultaneously record neural activity of extended neuronal populations *in vivo*. To fully harness the potential of functional imaging, one needs to infer the sequence of action potentials from fluorescence traces. Here we build on recently proposed computational approaches to develop a blind sparse deconvolution (BSD) algorithm based on a generative model for inferring spike trains from fluorescence traces. BSD features, (1) automatic (fully unsupervised) estimation of the hyperparameters, such as spike amplitude, noise level and rise and decay time constants, (2) a novel analytical estimate of the sparsity prior, which yields enhanced robustness and computational speed with respect to existing methods, (3) automatic thresholding for binarizing spikes that maximizes the precision-recall performance, (4) super-resolution capabilities increasing the temporal resolution beyond the fluorescence signal acquisition rate. BSD also uniquely provides theoretically-grounded estimates of the expected performance of the spike reconstruction in terms of precision-recall and temporal accuracy for each recording. The performance of the algorithm is established using synthetic data and through the SpikeFinder challenge, a community-based initiative for spike-rate inference benchmarking based on a collection of joint electrophysiological and fluorescence recordings. Our method outperforms classical sparse deconvolution algorithms in terms of robustness, speed and/or accuracy and performs competitively in the SpikeFinder challenge. This algorithm is modular, easy-to-use and made freely available. Its novel features can thus be incorporated in a straightforward way into existing calcium imaging packages.

1. 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. Compared to standard electrophysiology methods, calcium imaging is non-invasive, allows monitoring extended neuronal networks (up to a few tens of thousands of units) and can be combined with genetic methods in order to target specific neuronal populations. However, calcium imaging only provides a proxy measure of the neuronal activity. The kinetics of the calcium/reporter complexation being relatively slow, a spike-evoked fluorescence transient lasts much longer (0.1-1 s) than the action potential itself (< 5 ms). 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 (Holekamp et al., 2008; Panier et al., 2013; Ahrens et al., 2013; Wolf et al., 2015), these methods need to be fast and unsupervised, as any manual check of the produced inference signals would be prohibitively tedious. Due to the high noise level, simple inference methods such as naive linear deconvolution and Wiener filtering prove inadequate. In the last decade, numerous alternative deconvolution algorithms have thus been proposed (Yaksi and Friedrich, 2006; Holekamp et al., 2008; Sasaki et al., 2008; Vogelstein et al., 2009, 2010; Grewe et al., 2010; Mishchencko et al., 2011; Pnevmatikakis et al., 2013, 2016; Lütcke et al., 2013; Deneux et al., 2016; Theis et al., 2016; Picardo et al., 2016; Friedrich et al., 2017, 2016); among them, a powerful family of algorithms is based on non-negative sparse deconvolution (Vogelstein et al., 2010; Pnevmatikakis et al., 2016; Friedrich et al., 2017). In short, it consists in solving the inverse problem using the *a priori* knowledge that the spikes are sparse and non-negative. This framework, introduced by Vogelstein et al. (2010), 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 that controls the mean spike rate and the time constants characterizing the calcium reporter dynamics.

Despite extensive efforts for automatically adjusting these parameters (Pnevmatikakis et al., 2016; Friedrich et al., 2017), progress are still needed to achieve the adequate robustness of the inference (Kazemipour et al., 2016). Another drawback of current algorithms is that the interpretation of the output can be challenging due to a paucity of theoretical understanding of their performance. No information is

* Corresponding author.
E-mail address: georges.debregeas@upmc.fr (G. Debregeas).

https://doi.org/10.1016/j.jneumeth.2020.108763
Received 3 February 2020; Received in revised form 5 May 2020; Accepted 6 May 2020
Available online 30 May 2020

provided regarding the expected error rate of the inference or the time-precision of the output signal, *i.e.* the probability that a given spike be inferred in advance or delayed with respect to the true spike. Such information would be highly valuable, not only for downstream analysis but also for prior experimental design. In functional imaging, a trade-off has to be made between the sampling frequency, the signal-to-noise ratio and the imaged volume (that in turn sets the number of recorded neurons). Such choice of experimental parameters could be rationalized if one could forsee the achievable performance in terms of spike detection and timing precision for any given configuration. To address these various requirements, the inference algorithm should thus be accurate enough to make the most of the data, while being simple enough to be interpretable. It should provide a robust and unbiased extraction of the experimental parameters, such as signal-to-noise ratio or kernel shape, from the raw fluorescence datasets. It should finally offer a theoretically-grounded estimate of the inference performance for any given value of these parameters.

In the present study, we build on a recently proposed non-negative sparse inference method to develop the so-called *blind sparse deconvolution (BSD)* algorithm. This novel implementation features automatic estimation of the hyperparameters, enhanced speed, similar-to-better reconstruction performances and super-resolution capabilities. We additionally provide thresholding guidelines and theoretical bounds on its performance, in terms of inference efficiency and temporal accuracy as a function of the experimental parameters. These various features are benchmarked on both synthetic and real data, covering a large spectrum of experimental contexts.

2. Materials and methods

2.1. Generative model

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:

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

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: $\langle \varepsilon_i\rangle = 0$, $\langle \varepsilon_i\, \varepsilon_j\rangle = \sigma^2\delta_{i,j}$. The convolution kernel $K(t)$, which reflects the complexation kinetics, is of the form:

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

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 $\max_t K(t) = 1$, 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 detector noise, incorrect baseline estimation, and other modeling errors. Typical SNR values range from below 1 to 10.

This description corresponds to configuration in which reporter sensitivity and acquisition rates allow the detection of individual spikes. In many real situations, this is not the case and one can only detect bursts of spikes over timescales shorter than the sampling window. In this case, the signal N to be inferred is not binary anymore but remains sparse, and it reflects the instantaneous spike rate. The amplitude a is a characteristic scale of the (non-zero) signal, *e.g.* such that $\frac{\text{Var}(N)}{\langle N\rangle} = a$.

2.2. Non-negative sparse deconvolution

We recall first the non-negative sparse deconvolution approach for inferring $N(t)$. We rewrite Eq. (1) as:

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

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 $\mathcal{K}_{ij} = K[\Delta t(i-j+1)]$ 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 $[(j-1)\Delta t, j\Delta t]$ [1]. Note that the first and second lines are not equivalent. The second expression implicitly assumes that:

2 The boundary condition $N(t) = 0$, $\forall\, t < 0$ holds, which is generally true in recordings that start during inactive periods. This simplification can be easily relaxed for inference.[2]
3 One can approximate $K(t_i - t_l) = K(i\Delta t - t_l)$ as $K[i\Delta t - (j_l - 1)\Delta t]$ where $j_l - 1 = \lfloor \frac{t_l}{\Delta t} \rfloor$. This discretization error is negligible when Δt is small,[3] yet it ensures that the matrix $\mathcal{K}$ is translation invariant, *i.e.* $\mathcal{K}_{ij} = K(i-j)$.

From Eq. (3), a naive estimate for N can be written as:

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

In practice however, this approach fails to recover any spike at typical noise level $SNR = 2.5$ as shown in Section 3.1 and illustrated in Fig. 1. To understand this failure, one may reason in the continuous framework for which Eq. (4) reads $\hat{N}(t) \propto \int K^{-1}(\tau)[F(t-\tau) - b]d\tau$. 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 thus involves computing the derivatives of the original signal, and is therefore extremely sensitive to high frequency noise. 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 such filtering smoothes out the fast rise of the spike-evoked fluorescence transients. 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:

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

or equivalently:

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

where λ, $\lambda' = \frac{\lambda}{a}$ are L_1 penalty coefficients that control the sparsity of

[1] The choice of the convention $\mathcal{K}_{ij} = K[\Delta t(i-j+1)]$ instead of $\mathcal{K}_{ij} = K[\Delta t(i-j)]$ 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 occurring *after* measurement F_{i-1} and *before* measurement F_i

[2] 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.

[3] We go beyond this approximation in Section 2.6, when discussing super-resolution.

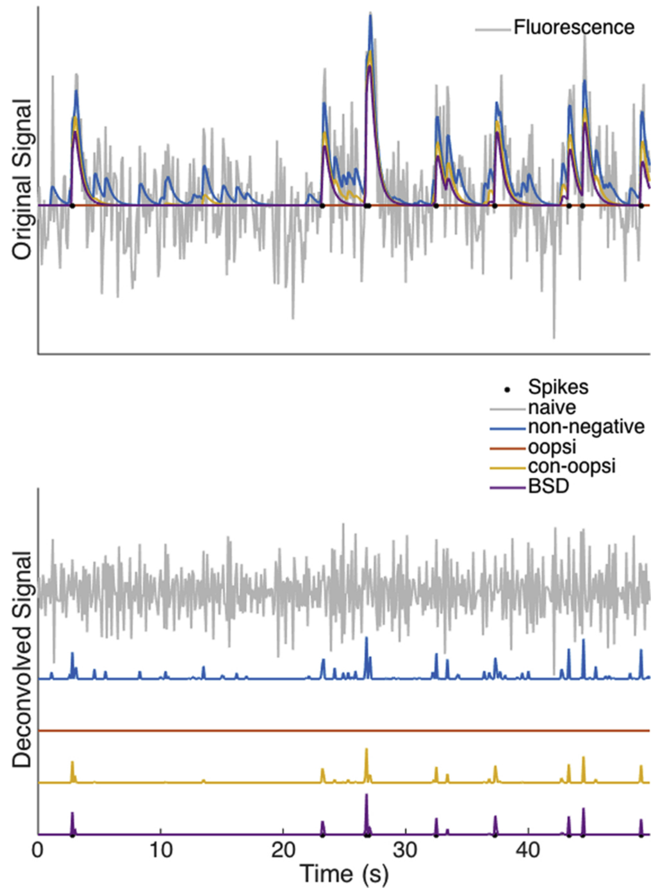

Fig. 1. Example of fluorescence signal and inference results for various deconvolution frameworks. The fluorescence signal was generated using parameters $f = 10\,\text{Hz}$, $\sigma = 0.4$, $a = 1$, $\tau_r = 0.1$, $\tau_d = 0.5$. The naive inference corresponds to the result of Eq. (5).

the optimum (the higher λ, the sparser the optimum). When $\lambda = 0$ and the $N \geq 0$ constraint is relaxed, the optimal value $\widehat{N}$ is exactly given by Eq. (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 $O(T)$ for double-exponential kernels using the interior-point method, see Vogelstein et al. (2010). This unusual linear scaling for a matrix inversion-like operation owes to the fact that $\mathcal{K}^{-1}$ is tridiagonal for double-exponential kernels: $\mathcal{K}_{ij}^{-1} \propto \delta_{ij} - \gamma_1 \delta_{i,j+1} + \gamma_2 \delta_{i,j+2}$, 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 Friedrich et al. (2017), the authors apply the Pool-Adjacent Violator Algorithm originally developed for isotonic regression problems to solve this optimization in an even faster but approximate way.

2.3. Determination of the sparsity prior λ and signal threshold value

The choice of the regularization parameter λ is crucial. If it is too large, the inferred spike train is $\widehat{N} = 0$ and all spikes are missed, whereas if it is too small, noise-induced transients are interpreted as spikes 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 Eq. (7) and drop the primes. We assume for now that all generative model parameters are known.

2.3.1. Existing methods: fast-oopsi and constrained-oopsi

In Vogelstein et al. (2010), 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:

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

This approach thus provides an analytical expression for λ. However, using an exponential approximation instead of Poisson can largely overestimate the threshold required (it is illustrated for a simple example in Appendix I), and result in improper behavior in several realistic experimental conditions as shown in Section 3.1 and illustrated in Fig. 1. To address this issue, a non-analytical method called constrained-oopsi (referred to as con-oopsi in the following) was recently introduced in Pnevmatikakis et al. (2016). The authors propose the following constrained deconvolution:

$$\widehat{N} = \underset{N \geq 0}{\operatorname{argmin}} \sum_i N_i$$
$$\text{subject to } \sum_i [F_i - (\mathcal{K}N)_i]^2 \leq \sigma^2 T \tag{9}$$

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:

$$\widehat{N}(\rho) = \underset{N \geq 0}{\operatorname{argmin}} \left\{ \sum_i N_i + \rho \sum_i [F_i - (\mathcal{K}N)_i]^2 \right\} \tag{10}$$

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:

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

In practice, $\lambda_{con-oopsi}$ is found by alternatively solving Eq. (7) and updating λ, decreasing it if the reconstruction error is too large, and increasing it otherwise. This non-analytical approach performs better than fast-oopsi (see Section 3.1 and Fig. 1).

2.3.2. Blind sparse deconvolution

We propose a different analytical expression for λ, inspired by Selesnick (2012). 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.

2.3.3. 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:

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

Let us first assume that the signal is spikeless, such that $F_i = \sigma\, \varepsilon_i$, where ε_i is a Gaussian white noise. Since $(\mathcal{K}^T\mathcal{K})N > 0$, we have:

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

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

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 $P_{FP} < 0.01$.

2.3.4. Single spike signal

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

$$N_i^0 = \delta_{i,i_0}$$
$$F_i = \mathrm{a}\,\mathrm{K}[\Delta t(i - i_0 + 1)] + \sigma\varepsilon_i \tag{15}$$

The gradient writes:

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

We look for an optimum of the form $\widehat{N_i} = \mathrm{a}\, \mathrm{n}\delta_{i,i_0}$. The optimization with respect to n gives:

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

where the last line assumes that $\sum_{i=1}^T K^2[(i-i_0)\Delta t] \approx \sum_{i=-\infty}^{\infty} K^2[i\Delta t] \equiv \|K\|^2$, 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.

Importantly, the effective noise 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 even signals with low SNR can be efficiently inferred provided that the spike-induced fluorescence transient is sufficiently well sampled.

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

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

Therefore, setting $\lambda \le \lambda_2 \equiv \|K\|^2 a - z_2\sigma\|K\|$, 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. 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 $\lambda = \frac{z_1 a\,\|K\|}{z_1+z_2} = \lambda_1(\sigma_{\max})$.

2.3.5. Sparsity prior for BSD

To summarize, in our blind sparse deconvolution (BSD) algorithm, the sparsity prior is set analytically as:

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

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.

We expect the sparse deconvolution to perform more consistently with λ_{BSD} than with λ_{oopsi}. Indeed, 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 $\lambda_1 = 0.51$; using λ_{oopsi} therefore results in multiple noise-induced false spikes. Conversely, for $\sigma = 0.25$ and $\nu = 0.1$ Hz, we have $\lambda_1 = 1.25$, $\lambda_2 = 3.39$ $\lambda_{oopsi} = 6.25$; in this case, λ_{oopsi} is too large and most, if not all the spikes are missed.

Robust performances are also expected using $\lambda_{con-oopsi}$, although a slightly larger FPR is expected compared to λ_{BSD}. Indeed, in the absence of spikes, $N = 0$ satisfies the reconstruction constraint in the large T limit and is correctly found by the algorithm. In the presence of spikes, we expect $\lambda_{con-oopsi}$ to be slightly lower than λ_{BSD}, resulting in small overfitting of the noise. Indeed, as soon as $\lambda > 0$, the spikes are on average underestimated, see Eq. (17); therefore, any 'good' choice of λ that perfectly filters the noise also underestimates the reconstructed trace $K \quad N$, yielding a reconstruction error $\sum_{i=1}^T (F_i - \mathrm{a}\,\mathrm{nK}[(i-i_0)\Delta t])^2 > \sigma^2 T$ and is therefore not a valid solution for the constrained optimization. Instead, λ is further decreased until false (noise-induced) spikes appear and reduce the reconstruction error below $\sigma^2 T$.

2.3.6. 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 Sections 2.3.3 and 2.3.4 show 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:

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

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.

2.4. Theoretical limits on the precision–recall and temporal resolution

We now present a similar analysis to derive theoretical estimates of the true and false positive rates, and of the temporal accuracy of the predicted spikes trains as function of the generative model parameters. The corresponding scripts are also implemented in the BSD package, and can be directly applied for a given fluorescence recording once the generative model parameters have been inferred.

2.4.1. 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 2.3.3: we obtain a probability of false positive rate per time bin:

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

For the false negative rate, we follow a similar reasoning as in Section 2.3.4: we consider a signal of the form $F_i = an_0K[\Delta t\, i - t_0] + \sigma\varepsilon_i$ with $t_0 = (i_0-1)\Delta t + \delta t_0$ and $0 \le \delta t_0 < \Delta t$. Note that we have now relaxed the previously made approximation

$K[\Delta t\, i - t_0] \approx K[\Delta t(i - i_0 + 1)]$ in order to probe the effect of intermittent sampling. We obtain a lower bound [4] for the probability of false negative per spike:

$$\begin{aligned}\widehat{N} &= \underset{N \geq 0}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{N}) \approx \mathrm{a\, n}\delta_{i,i_0} \\ &\Rightarrow n \sim \max[\mathcal{N}(n_0 \cos\theta(-\delta t_0) - \min(z_1\tilde{\sigma}, \frac{z_1}{z_1 + z_2}), \tilde{\sigma}^2), 0] \\ &\Rightarrow P_{\mathrm{FN}}(\delta_0) = \Phi[\frac{n_0 \cos\theta(-\delta t_0) - \min(z_1\tilde{\sigma}, \frac{z_1}{z_1+z_2})}{\tilde{\sigma}}]\end{aligned} \tag{22}$$

where

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

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 negative is given by:

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

2.4.2. Temporal resolution for isolated spikes

Intuitively, the temporal resolution depends on three factors: the signal-to-noise ratio, the sampling rate, the shape of the fluorescence kernel. It is characterized by the point-spread function (PSF) of the inferred spikes with respect to the true spikes, namely the conditional average given a fluorescence signal with a single-spike at $t = 0$:

$$\mathrm{PSF}_\tau = \mathbb{E}[\widehat{N}_{t_0+\tau} | N_t^0 = \delta_{t,t0}] \tag{25}$$

where $\widehat{\mathbf{N}} = \mathrm{argmin}_{\mathbf{N}}\, \mathcal{L}(\mathbf{N})$ is the spike train inferred from the fluorescence signal see Eq. (12) and the expectation is taken over the Gaussian noise realizations, see Eq. (1). The distribution of $\widehat{\mathbf{N}}$ is not tractable and, therefore, the PSF cannot be derived by exact analytical computation. Instead, we use two heuristics to obtain analytical insights into the width of the PSF and an efficient numerical approximation of the PSF.

For the analytical computation, we focus on the distribution of the initial negative gradients $-\frac{\partial \mathcal{L}}{\partial N_i}|_{\mathbf{N}=0}$ rather than the one of $\widehat{\mathbf{N}}$. Consider indeed the gradient descent optimization dynamics: because of the L_1 penalty, large components N_i tend to grow faster and to screen neighboring 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 $\widehat{\mathbf{N}}$ to be similarly delayed with respect to the true spike position. The probability of such an error can be computed as:

$$\begin{aligned}-\frac{\partial \mathcal{L}}{\partial N_{i_0}}|_{N=0} &= a||K||^2 + \sigma \sum_i K[\Delta t(i - i_0 + 1)]\varepsilon_i - \lambda \\ -\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}}|_{N=0} &= a||K||^2 \cos\theta(\delta\Delta t) + \sigma \sum_i K[\Delta t(i - i_0 - \delta + 1)]\varepsilon_i - \lambda \\ \Rightarrow \Delta\left[-\frac{\partial \mathcal{L}}{\partial N}\right] &\sim -2a||K||^2 \sin^2\frac{\theta(\delta\Delta t)}{2} + 2\sigma||K|| \sin\frac{\theta(\delta\Delta t)}{2} \mathcal{N}(0, 1)\end{aligned} \tag{26}$$

where the angle $\theta(\delta t)$ is defined in Eq. (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:

$$\delta t \;\; s.\, t. \quad \sin\frac{\theta(\delta t)}{2} = \frac{\sigma}{a||K||} \tag{27}$$

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.

For the numerical approximation of the PSF, we restrict the optimization search space to solutions of the form $\widehat{N}_i(\tau, n) = \mathrm{a\, n}\delta_{i,i_0+\tau}$. Using this simplification, the optimization over n can be carried out analytically (similarly as Eq. (22)) for each τ, and PSF_τ is given by the probability that the optimal solution is located at τ. After rearrangement, we obtain:

$$\mathrm{PSF}_\tau \approx P(X_\tau \geq 0 \;\&\; X_\tau \geq X_{\tau'} \quad \forall\, \tau') \tag{28}$$

where $X_\tau, \tau \in \mathbb{Z}$ is a Gaussian vector with mean $\mathbb{E}[X_\tau] = \cos(\theta(\tau\Delta t)) - \min\left[\frac{z_1\sigma}{a\,||K||}, \frac{z_1}{z_1+z_2}\right]$ and covariance $\mathrm{Cov}(X_\tau, X_{\tau'}) = \frac{\sigma}{a}\cos\theta(\Delta t|\tau - \tau'|)$.

Eq. (28) is efficiently evaluated by Monte Carlo, with values of $|\tau| \leq 20$ and is valid for the discrete generative model; the computation can be generalized for the continuous generative model, see Appendix D. The final formula is given by:

$$\mathrm{PSF}(\delta t) = \frac{1}{\Delta t^2} \int_{\delta t_0 = 0}^{\Delta t} \sum_{\tau \in \mathbb{Z}} \mathrm{PSF}_\tau(\delta t_0)\, \mathbb{1}_{\tau\Delta t - \delta t_0 < \delta t} \tag{29}$$

where $PSF_\tau(\delta t_0)$ is the discrete PSF of Eq. (28), but with a term cos $(\theta(\tau\Delta t - \delta t_0))$ instead of $\cos(\theta(\tau\Delta t))$ in the mean value of X_τ.

2.5. 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.

2.5.1. Initial estimation of the parameters

We are given a time series of the form $\mathbf{F} = a\mathcal{K}\,\mathbf{N} + \sigma\varepsilon + b$, 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 $\min[\mathcal{N}(0, \sigma^2), 0]$; it is fitted to deduce σ.

[4] 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 at different time steps are non-zero

We estimate the convolution kernel K through the signal auto-correlation matrix. Indeed, observe that:

$$A_F(l) \equiv \langle F_i F_{i+l}\rangle - \langle F\rangle^2 = a^2 \sum_{j,k} [\langle N_j N_{j+k}\rangle - \langle N\rangle^2] \sum_i K[(i-j-1)\Delta t]K[(i+l-k-j-1)\Delta t] + \sigma^2\delta_{l,0} \tag{30}$$

Under the assumption that the spiking events N_i are independent, identically distributed Poisson variables, we have $\langle N_j N_{j+k}\rangle - \langle N\rangle^2 = a^2\nu\Delta t\delta_{k,0}$ and Eq. (30) can be simplified as:

$$A_F(l) = a^2\nu\Delta t \sum_{j=-\infty}^{\infty} K[j\Delta t]K[(l+j)\Delta t] + \sigma^2\delta_{l,0}$$
$$\Leftrightarrow (A_F(l) - \sigma^2\delta_{l,0}) \propto \sum_{j=-\infty}^{\infty} K[j\Delta t]K[(l+j)\Delta t] \tag{31}$$

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. In practice, $\hat{A_F}$ obtained is not necessarily positive definite, 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 $\min(\sigma^2, \lambda_{min})$ instead of σ^2, where λ_{min} is the smallest eigenvalue of the Toepliz autocorrelation matrix. 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:

$$\langle F\rangle = a\nu\Delta t \sum_i K[i\Delta t]$$
$$\langle F^2\rangle - \langle F\rangle^2 = a^2\nu\Delta t \sum_i K[i\Delta t]^2 \tag{32}$$

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, Eq. (32) 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.

When the estimated time constants τ_r and τ_d differ from their true values, τ_r^0 and τ_d^0, systematic estimation errors arise. Suppose for instance that $\tau_r < \tau_r^0$ and $\tau_d = \tau_d^0$. 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. Hence, inferred spikes will tend to be duplicated. In general, the nature of the error depends on the kernel mismatch; some simulation results are presented in Appendix B. These results highlight the need to refine the kernel parameters estimators.

2.5.2. Iterative parameter estimation: adaptive blind deconvolution

At fixed hyperparameters, the cost function to minimize is the sum of a reconstruction error and a sparse penalty on N:

$$\mathcal{L}(\mathbf{N}, K, b, \sigma, a) = \frac{1}{2}\|\,\mathbf{F} - \mathcal{K}\,\mathbf{N} - \mathbf{b}\|^2 + \lambda_{\mathrm{BSD}}(\|K\|, \sigma, a)\|\mathbf{N}\|_1 \tag{33}$$

We can further refine the model hyperparameters by jointly minimizing this cost function with respect to both the spike train and the hyperparameters. In other words, we look for the convolution kernel that achieves the best trade-off between inferred spikes sparsity and reconstruction error. Such optimization yields non-trivial kernels, since a very sharp kernel (τ_r, $\tau_d \to 0$) would give a perfect reconstruction but dense spikes, whereas a wide kernel would give very sparse spikes but poor reconstruction. Furthermore, it can be shown that such sparsity-reconstruction trade-off maximization, which is also featured in Sparse Dictionary Learning (Mairal et al., 2009) or Blind Deconvolution in image deblurring (Freeman et al., 2009) is equivalent to a Maximum A Posteriori optimization of the likelihood, assuming a gaussian noise, a sparse prior for spikes and a flat prior for the hyperparameters (Freeman et al., 2009). In practice, we optimize over K and b iteratively through the following coordinate descent algorithm:

$$\widehat{\mathbf{N}}^{(t)} \longleftarrow \operatorname*{argmin}_{\mathbf{N}} \mathcal{L}(\mathbf{N}, K^{(t-1)}, \mathbf{b}^{(t-1)}, \sigma, a)$$
$$(K^{(t)}, \mathbf{b}^{(t)}) \longleftarrow \operatorname*{argmin}_{K,b} \mathcal{L}(\widehat{\mathbf{N}}^{(t)}, K, \mathbf{b}, \sigma, a) \tag{34}$$

The first optimization step was discussed in Section 2.2. The second optimization is a parametric temporal regression problem; it can be solved efficiently in $O\left(\frac{\tau_r+\tau_d}{\Delta t}\right)$ by introducing the cross-correlation $X_\tau = \frac{1}{T}\sum_i F_i \widehat{N}_{i-\tau}$ and auto-correlation $A_\tau = \frac{1}{T}\sum_i \widehat{N}_i \widehat{N}_{i-\tau}$ functions up to some cut-off $\tau_m \sim \frac{\tau_r+\tau_d}{\Delta t}$ (see details in Appendix C). The purpose 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 Appendix C 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. 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

$$\hat{a} = \frac{\sum_t \widehat{N}_t'}{\sum_t 1_{\widehat{N}_t > 0}} + \frac{\lambda}{\|K\|^2} \tag{35}$$

where the last term corrects the bias due to the sparse prior (see Eq. (17))

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_t [(F_i - (\mathcal{K}N')_i - b)^2]} \tag{36}$$

2.6. Super-resolution

Most fluorescent microscopy techniques – two-photon, confocal or light-sheet – involves the sequential scanning of a 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 (Rust et al., 2006; Huang et al., 2008; Fernández-Suárez and Ting, 2008; Heilemann et al., 2008; Hell, 2009), and compressed sensing applied to MRI allows to drastically reduce the number of measurements required to reach a given resolution (Lustig et al., 2008). 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 (Vogelstein et al., 2009). 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 extend the Blind Sparse Deconvolution framework to super-resolution, i.e. we develop a method to infer spiking events timing with a temporal resolution beyond the sampling rate.

2.6.1. Qualitative analysis

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

$$F_i = \mathrm{a}\,\mathrm{n}K(\Delta t\, \mathrm{i} - t_0) + \sigma\varepsilon_i + b \tag{37}$$

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

$$\begin{aligned} F_i \quad &= \mathrm{a}\,\mathrm{n}\left[\lambda_d^{\,i-j-\frac{\delta t}{\Delta t}} - \lambda_r^{\,i-j-\frac{\delta t}{\Delta t}}\right] + \sigma\varepsilon_i \\ \Leftrightarrow F_i \ &= \mathrm{a}\,\mathrm{n}\lambda_d^{-\frac{\delta t}{\Delta t}}\lambda_d^{\,i-j} - \mathrm{a}\,\mathrm{n}\lambda_r^{-\frac{\delta t}{\Delta t}}\lambda_r^{\,i-j} + \sigma\varepsilon_i \end{aligned} \tag{38}$$

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 $\mathrm{a}\,\mathrm{n}\lambda_d^{-\frac{\delta t}{\Delta t}}$ and $\mathrm{a}\,\mathrm{n}\lambda_r^{-\frac{\delta t}{\Delta t}}$, which can be converted to estimates of δt 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 $\mathrm{a}\,\mathrm{n}\lambda_d^{-\frac{\tau}{\Delta t}}$, and we cannot recover both n and τ without ambiguity.[5] 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.

2.6.2. 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:

$$\begin{aligned} F_k^s \quad &= a\sum_{j=1}^{\mathrm{sT}} K\left[\frac{\Delta t}{s}(k - j + 1)\right]N_j^s + b + \sigma\varepsilon_k^s \\ F_i \quad \equiv F_{\mathrm{i\,s}}^s &= a\sum_{j=1}^{\mathrm{sT}} K\left[i\Delta t - (j-1)\frac{\Delta t}{s}\right]N_j^s + b + \sigma\varepsilon_i \\ \Leftrightarrow\ \mathbf{F} \quad &= a\mathcal{K}\mathbf{N^s} + \mathbf{b} + \sigma\varepsilon \end{aligned} \tag{39}$$

where F_i is the fluorescence measurement at $t_i = i\Delta t$ and $N_j^s = \int_{(j-1)\frac{\Delta t}{s}}^{j\frac{\Delta t}{s}} N(t)\mathrm{d}t$ is the number of spikes emitted in the time interval $[(j-1)\frac{\Delta t}{s}, j\frac{\Delta t}{s}]$.[6] The convolution matrix $\mathcal{K}$ is now rectangular, of size $T \times s\ T$. It is not translation invariant anymore with respect to the spikes index j as the norm of the transient, $\|K_j\| = \sqrt{\sum_i \mathcal{K}_{\mathrm{ij}}^2}$ now depends on j. Indeed, writing $j = (p-1)s + r$, we have:

$$\begin{aligned} \|K_j\| &= \sqrt{\sum_{i=1}^{T} K\left[i\Delta t - (j-1)\frac{\Delta t}{s}\right]^2} \\ &= \sqrt{\sum_{i=1}^{T} K\left[(i-p)\Delta t + \frac{s+1-r}{s}\Delta t\right]^2} \\ &\approx \sqrt{\sum_{k=-\infty}^{\infty} K\left[k\Delta t + \frac{s+1-r}{s}\Delta t\right]^2} = f(r) \end{aligned} \tag{40}$$

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

2.6.3. Sparse deconvolution

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

$$\hat{\mathbf{N}}^s = \underset{N^s \geq 0}{\mathrm{argmin}}\left\{\frac{1}{2}\sum_{i=1}^{T}[F_i - a(\mathcal{K}N)_i - b]^2 + \sum_{j=1}^{\mathrm{Ts}}\lambda^j N_j^s\right\} \tag{41}$$

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 Eq. (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 $\tau_r = 0$, $\sigma \ll a$. 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 Appendix E that both efficient noise filtering and unbiased estimation of spike timing for isolated spikes can be obtained with the following expression for λ_{BSD}^j:

$$\lambda_{\mathrm{BSD}}^j = z_1\|K_j\|\min\left(\sigma, \frac{a\frac{\sum_{r=1}^{s}\| K_r \|}{s}}{z1 + z2}\right) \tag{42}$$

In practice, the optimization can also be performed efficiently using the interior-point method (Vogelstein et al., 2010). Adding a small L_2 penalty $\sum_j \mu_j N_j^2$, 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 $\tau_r = 0$. The kernel inference can also be adapted efficiently to support super-resolution, see Appendix F.

2.7. Methods for performance evaluation

The algorithm is evaluated based on three criteria: whether the spikes are correctly detected or not; if they are detected, whether their timing is accurate; and whether the inferred generative model parameters match the ground truth values.

2.7.1. Spike detection

For the spike detection assessment, we adopt the SpikeFinder main metric (Deneux et al., 2016), i.e. the Pearson correlation between the ground-truth spike train discretized at a frequency f_{eval} and the inferred spike train resampled at the same frequency. Resampling is performed as follows: (i) if $f_{\mathrm{eval}} = f/s$ for integer s, s consecutive frames are summed; (ii) if $f_{\mathrm{eval}} = f \times s$ for integer s, we write $N'_j = N_{[j/s]}/s$. (iii) otherwise, (*e.g.* for SpikeFinder), the signal is first resampled to the closest multiplier or divider of f_{eval} by linear interpolation, then resampled using (i) or (ii). When $f_{\mathrm{eval}} = f$, the metric penalizes equally an undetected spike and a spike that is detected either in advance or delayed with respect to the true spike; when $f_{\mathrm{eval}} < f$, the metric is

[5] In Vogelstein et al. (2009), the authors assume $\tau_r = 0$ and that a is fixed.

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

tolerant to timing errors of order $1/f_{\text{eval}}$.

2.7.2. Spike timing

The accuracy of the timing is assessed by measuring the point spread function (PSF), *i.e* the average (over noise) of the inferred spike train in presence of an isolated single spike:

$$\mathrm{PSF}_\tau = \mathbb{E}[\widehat{N}_{t_0+\tau}|N_t^0 = \delta_{t,t_0}] \tag{43}$$

where $\tau \in \mathbb{Z}$ is the lag. Ideal point spread functions are centered around 0 and decay fast to ~ 0 with $|\tau|$. [7] . Then, the timing error can be characterized by the mean and standard deviation of a Gaussian curve fitted the point spread function.[8]

A naive estimator of *PSF* would be the cross-correlation between the ground-truth spikes and the inferred spikes $X_\tau = \frac{1}{T}\sum_{i=1}^{T} N_i^0 \widehat{N}_{i+\tau}$. Such estimator is correct when the spike train are Poisson-distributed, but is biased when the spikes are temporally correlated: in the best case scenario where the spike is perfectly recovered $\widehat{N}_i = N_i^0 \quad \forall\, i$, we obtain the auto-correlation of the spike trains $X_\tau = \frac{1}{T}\sum_{i=1}^{T} N_i^0 N_{i+\tau}^0 \equiv A_\tau$. To overcome this issue, we estimate the PSF as the kernel of a linear temporal regression model:

$$\hat{N}_t = \sum_{\tau=-m}^{m} \mathrm{PSF}_\tau N^0(t-\tau) + \varepsilon \tag{44}$$

The least square estimator for the PSF is then:

$$\mathbf{PSF} = \mathcal{A}^{-1}\mathbf{X}$$
$$\mathcal{A}_{\text{ij}} = A(i-j) \tag{45}$$

where PSF_τ, X_τ are indexed formally as vectors $\mathbf{R}$, $\mathbf{X}$.

This estimator is defined for N^0, $\widehat{N}$ with identical sampling rate; if $\widehat{N}$ has lower sampling rate, it is first upsampled using the above procedure.

2.7.3. Generative model parameters

Finally, we assess the accuracy of the blind generative model parameters inference. For simulated train spikes, ground truth generative model parameters exist and they can be directly compared with the inferred ones. For real fluorescence data with joint electrophysiological recordings, we derive 'ground truth' generative parameters using the knowledge of the spike positions. The kernel parameters and the baseline are obtained by minimizing the following mean square error:

$$\mathrm{MSE} = \sum_i \left(F_i - \sum_j K[(i-j+1)\Delta t]a_j N_j^0 - b\right)^2$$
$$(\tau_r^{\mathrm{GT}}, \tau_d^{\mathrm{GT}}, b^{\mathrm{GT}}) = \mathrm{argmin}_{\tau_r,\tau_d,b,\mathbf{a}\geq 0}\,\mathrm{MSE}(\tau_r, \tau_d, b, \mathbf{a}) \tag{46}$$

where the ground-truth spikes are discretized at the fluorescence sampling frequency. Note that we relax the hypothesis that all transients have the same amplitude a, and optimize over all the amplitudes $a_j \geq 0$. This is particularly important for spike bursts, where strong non-linear effects are observed. In practice, the optimization is carried out by using the exact same algorithm as for the fully blind setting, but with a position-dependent sparsity prior λ_i that takes the value 0 for positions where a ground truth spike is present and a large value (*e.g.* 20) elsewhere. Once inference is performed, the transient amplitude a is computed as the average transient amplitude $\left(\sum_j a_j\right)/(\sum_i N_i^0)$.

[7] We have $\lim_{\tau\to\pm\infty} \mathrm{PSF}_\tau > 0$ due to false positives

[8] This estimator is more stable with respect to noise and finite sample size than directly computing the first and second moments of the PSF.

3. Results

The blind sparse deconvolution (BSD) method, whose algorithmic details were presented in the preceding section, allows for unsupervised spike inference, *i.e.* both the algorithm hyperparameters – sparsity prior λ and generative model parameters – kernel time constants τ_r, τ_d, transient amplitude a, noise level σ - are automatically evaluated. BSD can infer spike trains at or beyond the fluorescence sampling rate. Importantly, the expected performances of BSD can be predicted as well and this possibility is integrated in the publicly available program.

Section 3.1 is dedicated to simulated data; we demonstrate that the choice of sparsity prior outperforms other methods in terms of spike and/or computational speed and that the kernel parameters can be accurately recovered using our iterative approach. We also demonstrate super-resolution capabilities for a wide range of experimental conditions. In Section 3.3, we apply BSD to the SpikeFinder contest, a collection of joint electrophysiological and fluorescence recordings. We show that (i) our choice of sparsity prior outperforms others (ii) the kernel inference is accurate, allows robust performance even in the absence of training data and can improve spike detection performance when interneuron variability is important. (iii) The predicted temporal errors are consistent with empirical errors and that integrating it into the prediction can improve spike detection. (iv) Super-resolution significantly increases the temporal accuracy for some datasets. Overall, our best submission is competitive across all datasets with state-of-the-art Machine Learning algorithms, while not requiring any training data or hyperparameter fine-tuning. In Section 3.4, we show that BSD scales well to large-scale zebrafish recordings. Finally, Section 3.2 is dedicated to experimental design: we use BSD to predict the expected accuracy for various standard calcium reporters and imaging parameters.

3.1. Simulated data

3.1.1. Spike detection accuracy

In BSD, the sparsity prior λ_{BSD} is computed analytically and allows one to simultaneously minimize, in a tractable way, both the false-positive rate (FPR) and false-negative rate (FNR) (see Methods). In contrast, the expression λ_{oopsi} used in the fast-oopsi algorithm (Vogelstein et al., 2010) offers no guarantee that either is small in all situations (see Methods). This issue has motivated the recent development of the constrained-oopsi algorithm (Pnevmatikakis et al., 2016) where the sparsity prior is determined iteratively. Fig. 1 illustrates the strong impact of the chosen value of the sparsity prior on the inference performance, as it compares the results of the four inference algorithms (BSD, oopsi, con-oopsi, and non-negative, *i.e.* with $\lambda = 0$) for a signal with $\sigma = 0.4$, $f = 10$ Hz. 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. Here, BSD infers slightly less false spikes than con-oopsi in this particular configuration. As a baseline, we also show the non-negative deconvolution without any sparsity prior, which is significantly better than naive deconvolution but shows significantly more false positives than con-oopsi and BSD. One may notice that the BSD's reconstructed signal is systematically lower than the original signal. This difference directly derives from Eq. (17), which indicates that a spike of amplitude a is reconstructed with an amplitude $a - \frac{\lambda}{\|K\|^2}$. However, this systematic bias in the reconstructed signal does not impact the quality of the inferred spike signal.

The performances of the four algorithms are 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 $t = 10{,}000$ s. The signal is generated according to the discrete model Eq. (3) with a fixed transient amplitude $a = 1$ and variable sampling frequency f and noise level σ. We use a double exponential kernel K with $\tau_r = 0.1$, $\tau_d = 0.5$ (see Eq. (2)), akin to a GCaMP6 reporter. Spike trains $\widehat{\mathbf{N}}$ are inferred *with knowledge of the generative model parameters* and compared

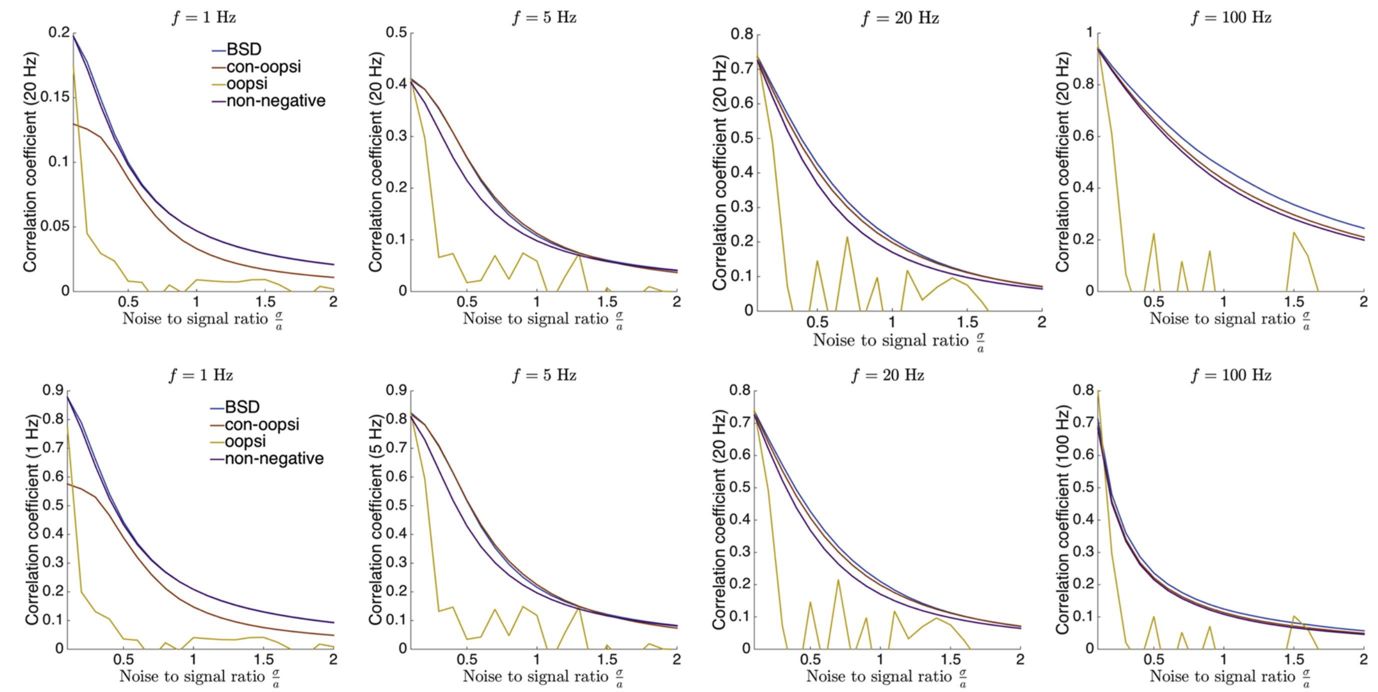

Fig. 2. Comparison of the reconstruction performances on synthetic data between BSD, con-oopsi, oopsi, and non-negative deconvolution. For each algorithm, the correlation between the inferred and ground truth spike train is shown as a function of the noise-to-signal ratio. The different plots correspond to various sampling frequencies f and evaluation frequency (indicated on the y axis legend).

with the original spike train **N** using the metric defined in Section 2.7.1. We show in Fig. 2 the correlation as function of the signal-to-noise ratio for various sampling frequencies, and at evaluation frequencies $f_{\text{eval}} = f$ (top row) and $f_{\text{eval}} = 10$ Hz (bottom row).

We observe that the fast-oopsi algorithm sometimes performs very well ($f = 100$ Hz, high SNR) sometimes equivalently as con-oopsi and BSD ($f = 100$ Hz, lower SNR), but often very poorly ($f = 20$ Hz, $\sigma > 0.2$); Such unreliability may be highly detrimental in actual experiments. BSD and con-oopsi yield comparable results, with BSD slightly outperforming con-oopsi in most configurations and particularly at low sampling rate ($f = 1$ Hz). The non-negative deconvolution without sparsity prior baseline is more stable than oopsi (as was previously reported in Pachitariu et al. (2018)) and performs equivalently as BSD and con-oopsi at high SNR, but, as expected, has a significantly higher false positive rate than BSD and con-oopsi at low SNR. Similar results are found when increasing the firing rate (see same simulations for $\nu = 1$ Hz, Supplementary Fig. 11); then differences tend to vanish at large firing rates (same experiment for $\nu = 5$ Hz, Supplementary Fig. 11).

We also compare the computational cost of the various algorithms. BSD, non-negative deconvolution and fast-oopsi all share the same core algorithm and therefore have similar computational cost. In contrast, the con-oopsi implementation 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 1. 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

Table 1
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.

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

(Friedrich et al., 2017), 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 3.4.

3.1.2. Kernel inference accuracy

When deconvolution is performed with an incorrect kernel, systematic biases arise in the inferred spike trains and the accuracy decreases both in terms of spike timings and spike detections: for instance, spikes may be split into two time frames if the rise time is too short, or some spikes may be missed if the decay time is too long, see Appendix B and Fig. 10 for a detailed study. We now relax the assumption that the generative model parameters are known and attempt to retrieve them from the raw fluorescence recordings. For long recordings of Poisson-distributed spike trains with sufficiently high signal-to-noise ratio and purely white Gaussian noise, the task is relatively easy. Indeed, the autocorrelation function follows exactly Eq. (31), and the initial kernel estimation is excellent. However, most real datasets feature bursts of spikes, temporally correlated noise, artifacts, limited spike counts and low signal-to-noise ratio. Under these conditions, Eq. (31) becomes incorrect.

We thus build a more realistic collection of synthetic fluorescence traces as follows: for each recording of the SpikeFinder dataset (see next section), we generate five synthetic fluorescence recordings according to the generative model of Eq. (1) using the actual spike train measured

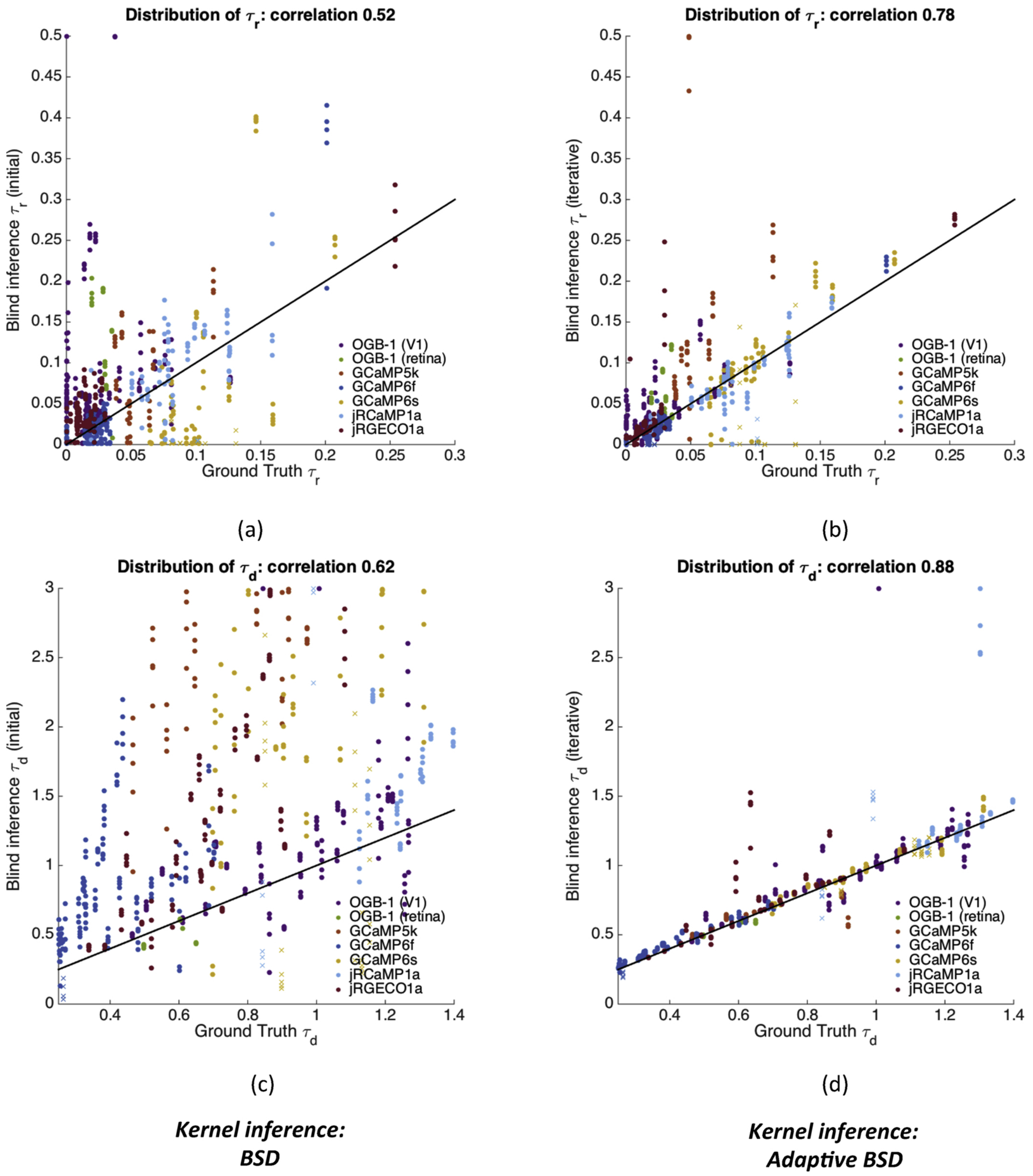

Fig. 3. Kernel inference benchmark for realistic simulated data For each recording of the SpikeFinder datasets (see Section 3.3), we generate five simulated fluorescence traces according to Eq. (3) using the actual spike trains, and signal-to-noise ratios and kernel parameters estimated, see Section 3.3 for derivation of the generative model parameters. Then, BSD and adaptive BSD are applied to infer the spike trains and the kernel parameters. Panels (a),(b): ground-truth rise and decay time constants vs inferred values using BSD. Panels (c),(d): Same, with adaptive BSD. In all panels, crosses denote spike recordings with fewer than 20 spikes.

by electrophysiology and the rise and decay time constants and signal-to-noise ratio inferred from the recording (see next section). This synthetic dataset has diverse spike counts, signal-to-noise ratios, sampling rates, kernel parameters and spiking patterns. We jointly infer the spike trains and kernel parameters using either BSD or adaptive BSD (iterative parameter estimation). Results are shown in Fig. 3. Panels (a) and (c) show the outcome of the initial kernel estimation, compared with the ground truth kernel values. We find that the decay time is often vastly overestimated. This systematic bias reflects the existence of large temporal correlations in the experimental spike signal whereas

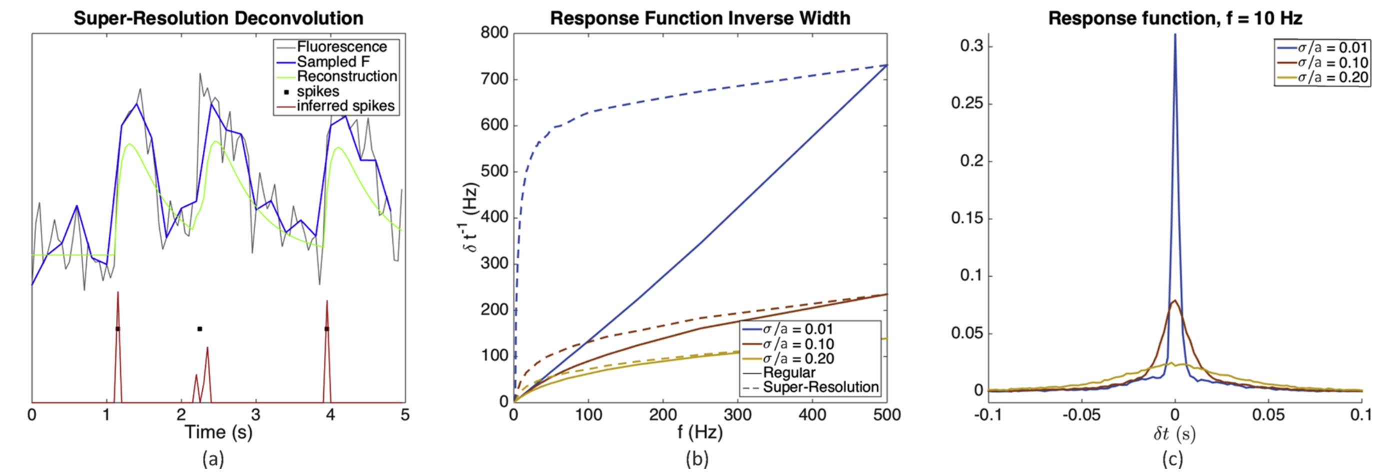

Fig. 4. Super-resolution for synthetic data. (a) Example of super-resolution inference: a fluorescence signal is generated at $f_0 = 20$ Hz, sampled at 5 Hz and reconstructed at 20 Hz. Parameters: $\tau_r = 0.1$, $\tau_d = 0.5$, $\sigma = 0.2$, $a = 1$. (b) point spread function at 10Hz for various noises. Notice a smaller width than the sampling interval 0.1 s (c) Inverse width δt^{-1} of the point-spread function, as function of the sampling frequency for regular reconstruction (full) and SR reconstruction (dotted).

the inference model assumes none. The algorithm thus tends to confuse spike bursts with long fluorescence transients, and the kernels are therefore poorly inferred at first. However, after kernel refinement (Panels (b) and (d)), the inferred kernel becomes very similar to its experimental counterpart in most experiments. Upon closer inspection, we can identify the main sources of error for the algorithm. Coarse errors arise when the algorithm starts from largely incorrect initial kernel values and ends up trapped in a wrong local minimum of the cost function. These errors can be easily avoided by providing realistic bounds for the rise and decay time constants. For all the remaining cases, three main factors affect the quality of inference parameter, see Supplementary Fig. 13: the spike count (higher is better), the effective signal-to-noise ratio (higher is better) and the spike 'burstiness', i.e. the deviations from Poisson-distributed spikes (lower is better).

3.1.3. Super-resolution

Fig. 4a shows an example of reconstruction of a signal generated at $f_0 = 20$ Hz, and sampled at 5 Hz. We observe a good agreement with the original spike train. In particular, it appears that in spite of the sparse sampling, the onsets of the green and dark curves transients are very close to one another.

We test our algorithm on synthetic datasets generated using the model Eq. (39) at $f_0 = 500$ Hz, with $\tau_r = 0.1$, $\tau_d = 0.5$, spike frequency $\nu = 2$ Hz. The fluorescence signal is down sampled to recording frequencies ranging from $f = 1$ Hz to 500 Hz. Spike trains are inferred with and without super-resolution. For super-resolution, we use a frequency gain $s = \frac{f_0}{f}$ in order to reconstruct a spike train at the original frequency f_0. We perform the spike inference for various sampling frequencies and noise levels, and we estimate the point spread function of the inferred spike train (*i.e.* the average response to a single spike, see the method in Section 2.7.2). Results are depicted in Fig. 4(b) and (c). They demonstrate that super-resolution is perfectly workable at small noise levels, and that a significant resolution gain can be achieved at intermediate noise level typical of actual experimental conditions. For instance, at $f = 10$ Hz, $SNR = 5$, the point spread function width is $\sim 2\times$ smaller than without super-resolution. Fig. 4(c) shows that the gain in resolution becomes significant as soon as $f \gtrsim 4$ Hz.

3.2. Theoretical limits of calcium reporter accuracy

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 signal 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 (see Appendix B and Fig. 10). These limitations have been characterized quantitatively in Wilt et al. (2013) 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.

We derived in Section 2.4 theoretical bounds of performance of our deconvolution algorithm; we present hereafter the main findings. We display in Fig. 5a 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 false-positive rate (FPR) is set at 0.01/frame (i.e. $\lambda_{BSD} = \lambda_1$ and $z_1 = 2.366$, see Methods). 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. Some point spread functions are displayed in Fig. 5b-inset for typical calcium indicators. We used the parameters: (i) GCaMP6s: $\tau_r = 180$ ms, $\tau_d = 0.55$, (ii) GCaMP5k: $\tau_r = 58$ ms, $\tau_d = 0.52$ s, (iii) GCaMP6f: $\tau_r = 25$ ms, $\tau_d = 0.38$ s, (iv) OGB1-like: $\tau_r = 20$ ms, $\tau_d = 80$ ms. For the 3 first sets of time constants, the values are deduced from the fluorescence recordings on mice V1 cells reported in Chen et al. (2013).

We also display in Fig. 5b the width δt of the point spread function (extracted using a gaussian fit) as a function of the noise level, for a sampling frequency $f = 60$ Hz. 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 $\sqrt{s}$. Therefore, there is no guarantee that increasing the sampling frequency improves the time-resolution. We display in Fig. 5c 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 of 10 Hz. We see that δt^{-1} saturates at a value that depends on the SNR and on the rise and decay constant times. For instance, with GCaMP6s and $SNR_{10Hz} = 5$, increasing the frequency beyond 50Hz does not result in improving the temporal resolution.

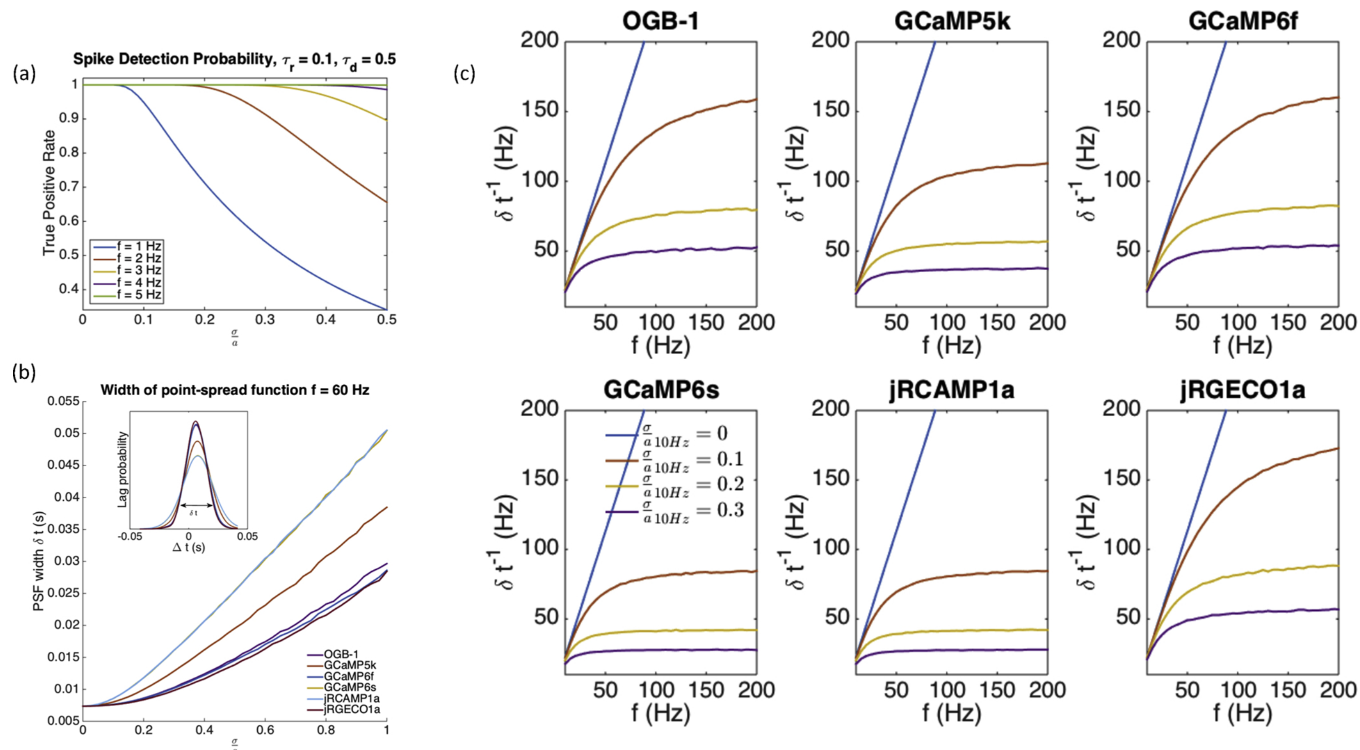

Fig. 5. Theoretical limits of calcium reporters. (a) True positive rate of BSD as a function of the noise level for different sampling rates, for a fixed FPR of 0.01/frame, (b) Width of the point spread function as function of the noise for various calcium indicators, at fixed frequency f = 60 Hz, (c) Width of the point spread function as a function of the sampling frequency for various calcium indicators and reference noises.

The observation that the temporal resolution saturates at high sampling rate can be understood by examining Eq. (27): 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.

3.3. Joint electrophysiological and fluorescence recordings: the SpikeFinder contest

SpikeFinder (http://spikefinder.codeneuro.org/) (Berens et al., 2018) is a public contest for spikes train inference from calcium recordings. It consists of a compilation of 10 datasets of joint electrophysiological and calcium recordings from mouse V1 or retina, for six different calcium probes (OGB-1,GCaMP5k,GCaMP6s,GCaMP6f, jRCaMP1a,jRGECO1a) at various sampling rate, see Table 2. The data was compiled from various sources (Theis et al., 2016; Akerboom et al., 2012; Chen et al., 2013; Dana et al., 2016). All the recordings are upsampled to f = 100 Hz, and the metric used to assess performance is the correlation between the inferred spike train and the electrophysiological spike train, computed after downsampling at 20 Hz. Dozens of algorithms based on Supervised Machine Learning, generative models and others have been benchmarked on SpikeFinder, and the results are publicy available.

We benchmark several inference algorithms and postprocessing variants on the SpikeFinder dataset using the following pipeline:

- **Preprocessing.** For all datasets, we started from the raw data, before upsampling or detrending. Indeed, when the fluorescence is upsampled, the noise becomes temporally correlated between time frames and the white noise assumption used to compute the sparsity prior λ is violated. Then, each fluorescence recording is normalized (zero mean and unit variance). The variable component of the baseline is further removed by substracting a moving percentile (quantile q = 0.15, variable window size $\in$[10s, 60s] adjusted by validation). Such baseline slow modulations are unrelated to the calcium signal but reflect experimental artefact such as photobleaching or minute axial motions of the specimen.
- **Initial values of rise and decay time.** We tested four initialization procedures: **Blind**: no information is provided and the automated initial kernel estimation described in Section 2.5.1 is used; **Literature** parameters are taken from the literature, see values and

Table 2
SpikeFinder data sets summary. For recording-specific variables, the median value across recordings was provided. The rise and decay time constants and signal-to-noise ratios, as defined by Eq. (2), are obtained by a parametric linear regression of the true spikes against the fluorescence, see Section 2.7.3. ⋆ The rise time for datasets 3 and 5 are unreliable because of important delays between the fluorescence and electrophysiological recordings.

Dataset	1	2	3	4	5	6	7	8	9	10
Calcium indicator	OGB-1	OGB-1	GCaMP6s	OGB-1	GCaMP6s	GCaMP5k	GCaMP6f	GCaMP6s	jRCAMP1a	jRGECO1a
Brain region	V1	V1	V1	Retina	V1	V1	V1	V1	V1	V1
Number of recordings	11	21	13	6	9	9	37	21	20	27
Spikes per recording	1012	525	1240	1304	872	363	127	96	75	232
Sampling rate (Hz)	40	12	60	8	60	50	60	60	15	30
Rise time (s)	0.02	0.02	0.00⋆	0.03	0.00⋆	0.05	0.02	0.10	0.09	0.01
Decay Time (s)	0.95	1.00	1.10	0.55	0.99	0.65	0.33	0.97	1.32	0.68
Noise level $\frac{\sigma}{a}$	1.06	0.68	0.83	1.06	0.90	0.41	0.26	0.21	0.39	0.29

sources in Supplementary Table VI; **Ground Truth** for each recording, we compute ground truth rise and decay time constants by temporal regression of the fluorescence from the true spikes, see Section 2.7.3. **Train Set** for each dataset, we use the median values of the ground truth rise and decay time as initial values, see values in Table 2. We also use the minimum and maximum values found in the dataset as bounds for adaptive BSD. Note that the ground-truth initialization is not a valid algorithm as it requires prior knowledge of the spikes for each neuron. We thys only use it as a reference for validating the kernel inference and as an upper bound of performance.

- **Inference Algorithm** We tested four sparse deconvolution variants: non-negative deconvolution (*i.e.* with $\lambda = 0$), con-oopsi, BSD and adaptive BSD.
- **Postprocessing.** The inferred signal is upsampled to 100 Hz by linear interpolation, and we add a temporal offset (adjusted on the training set). It is mostly relevant for datasets 3 and 5 to account for a delay between the fluorescence time and electrophysiological time, in agreement with other submissions (Berens et al., 2018). Then, the inferred spike is optionally convolved with the point-spread function computed in Section 2.4.2, using the parameters inferred from the recording. This can be viewed as a poor man (but computationally efficient) version of posterior averaging: one first computes the most likely spike train given the fluorescence using BSD, and then convolve it with the point-spread function (PSF) to account for the temporal uncertainty, see examples in Fig. 6

Quantitative results are displayed in Table 3 and selected examples are shown in Fig. 6.

Sparsity prior. As expected, introducing a sparsity prior reduces the false detection rate (see Fig. 6) and significantly improves the correlation regardless of the kernel choice. At fixed kernel parameters, BSD and con-oopsi are virtually equivalent but the former is faster.

Kernel Inference. When no training data is available, adaptive kernel inference is critical for performance as was found in the simulated data. Even when training data is available, adaptive kernel inference is always equivalent or better to using fixed kernels derived from training data as it takes into account inter-neuron variability. The most important gains is found for dataset 8, where non-linear effects result in large variability of decay rates. The quantitative results are corroborated by Fig. 7, which compares the rise and decay time constants and transient amplitude a inferred by the blind algorithm with their ground truth counterparts. Adaptive BSD successfully captures the rise and decay time constants variability both across experiments, and across neurons in a given experiment, notably for dataset 8 where the decay time can vary by more than 2-fold. When the training data is of low quality as for data sets 3 and 5, blind inference actually outperforms the ground-truth. Importantly, although our best submission was obtained with the train set initial value, the blind and literature initialization follow closely: BSD almost does not require any training data. We note however that the correlations are slightly smaller than the values found using synthetic data with similar parameters (see Fig. 3). This can be explained by artifacts, temporally correlated noise and non-linear effects. The inferred transient amplitudes also correlate with the ground-truth values, but less accurately. For the OGB-1 datasets, which have low signal-to-noise ratio, a is systematically overestimated because the algorithm frequently misses isolated spikes and can only detect bursts of consecutive spikes, see Fig. 6 panel (c); it therefore confuses the latter with individual spikes. Such confusion is expected and inevitable: for instance, in dataset 4, only bursts of more than 10 spikes produce visually detectable transients. For dataset 8 (GCaMP6s), we find conversely that some transients amplitudes are largely underestimated, because the algorithm infers many small (but above threshold) spurious spikes due to artifacts or non-linear effects, and confuses them with real spikes.

Temporal resolution. We assess the temporal accuracy of the algorithms as follows: for each recording, we computed the empirical point-spread function (PSF) between the ground-truth spikes and the inferred spikes using Eq. (45). An average PSF is then computed for each dataset by weighted average of the PSFs (normalized to max = 1, then weighted by the number of spikes). We also predicted for each dataset a theoretical Point-Spread Function, using Eq. (29) and either the parameters of Table 2 or the parameters inferred by BSD for each recording. Results are shown in Table 4 and Fig. 8. Empirical PSF and the fluorescence kernel for the last five datasets are displayed in Fig. 8. The PSF have smaller width than the kernel, *i.e.* the temporal resolution is significantly improved by deconvolution; it is in general of order Δt (the inter-frame period) rather than τ_d (the calcium reporter decay time), see Table 2. The results hold even in the absence of information about the kernel: in most datasets, the temporal resolution of blind adaptive BSD is very close if not equal to the one of BSD with ground-truth kernel parameters. Importantly, we find an excellent overall agreement between the empirical PSF and the predicted one using the train set parameters: except for datasets 3 and 5 which have unreliable ground-truth, the predicted widths are notably more accurate than naive estimates such as the rise time τ_r or the sampling rate Δt. The prediction is overly pessimistic for dataset 4 because the PSF is computed for an isolated spikes, whereas only spike bursts are detected in practice. The offsets are accurate for datasets 6–10 but not for datasets 1–5, probably due to delays between the electrophysiological and fluorescence recordings. When the inferred parameters are used, the prediction is less accurate owing to the difficulty to estimate the signal-to-noise ratio for single spikes.

Interestingly, we found that convolving the inferred spike train with the point-spread function computed using the inferred parameters significantly improves the accuracy for all algorithms and kernel initializations for the first five datasets, and has little to no impact for the last five datasets. Indeed at low signal-to-noise ratio, fluorescence transients may appear in advance (Fig. 6b) or delayed (Fig. 6c) with respect to the action potential and the corresponding point-spread function spans across several time windows. In contrast, datasets 6,7,8,10 have higher signal-to-noise ratio, hence thinner point-spread function (see Fig. 6(d)) and convolution does not change significantly the prediction at 20 Hz. The small performance drop for dataset 8 can be explained by the overall underestimation of the signal-to-noise ratio.

Super-resolution We tested super-resolution for the datasets 6–10 which have the largest signal-to-noise ratios. The same pipeline is used, but BSD is applied with super-resolution upsamplings factors of 2 (datasets 6,7,8), 6 (dataset 9) and 3 (dataset 10). Fig. 8 shows for each dataset the empirical point-spread function of the inferred spikes with (yellow) and without (red) super-resolution. Super-resolution reduces both the offset and the width of the point-spread function for all datasets, see Table 4. The most significant improvement is for dataset 9, which has high a signal-to-noise ratio and a relatively low sampling rate. Overall, super-resolution is relevant when the main source of temporal error is the sampling rate rather than the noise level.

Comparison with other submissions. Our algorithm compares positively with other algorithms that have been tested during the competition. According to the current leaderboard, it performs similarly as Team 6 (OASIS Friedrich et al., 2017), and is above Team 7 (Suite2p Pachitariu et al., 2017) of the original contest. Both teams also used a non-negative deconvolution framework, followed by a re-convolution by a PSF-like kernel. Both teams used the training set to determine hyperparameter values: Team 6 for the convolution kernel, sparsity prior and PSF and Team 7 for the convolution kernel and the PSF; for Team 7 the sparsity prior $\lambda = 0$ was used as in the non-negative deconvolution baseline presented here. In contrast, our approach does not require any training set; the fluorescence kernel, sparsity prior and PSF are determined automatically for each neuron; it can therefore be applied to any experimental configuration.

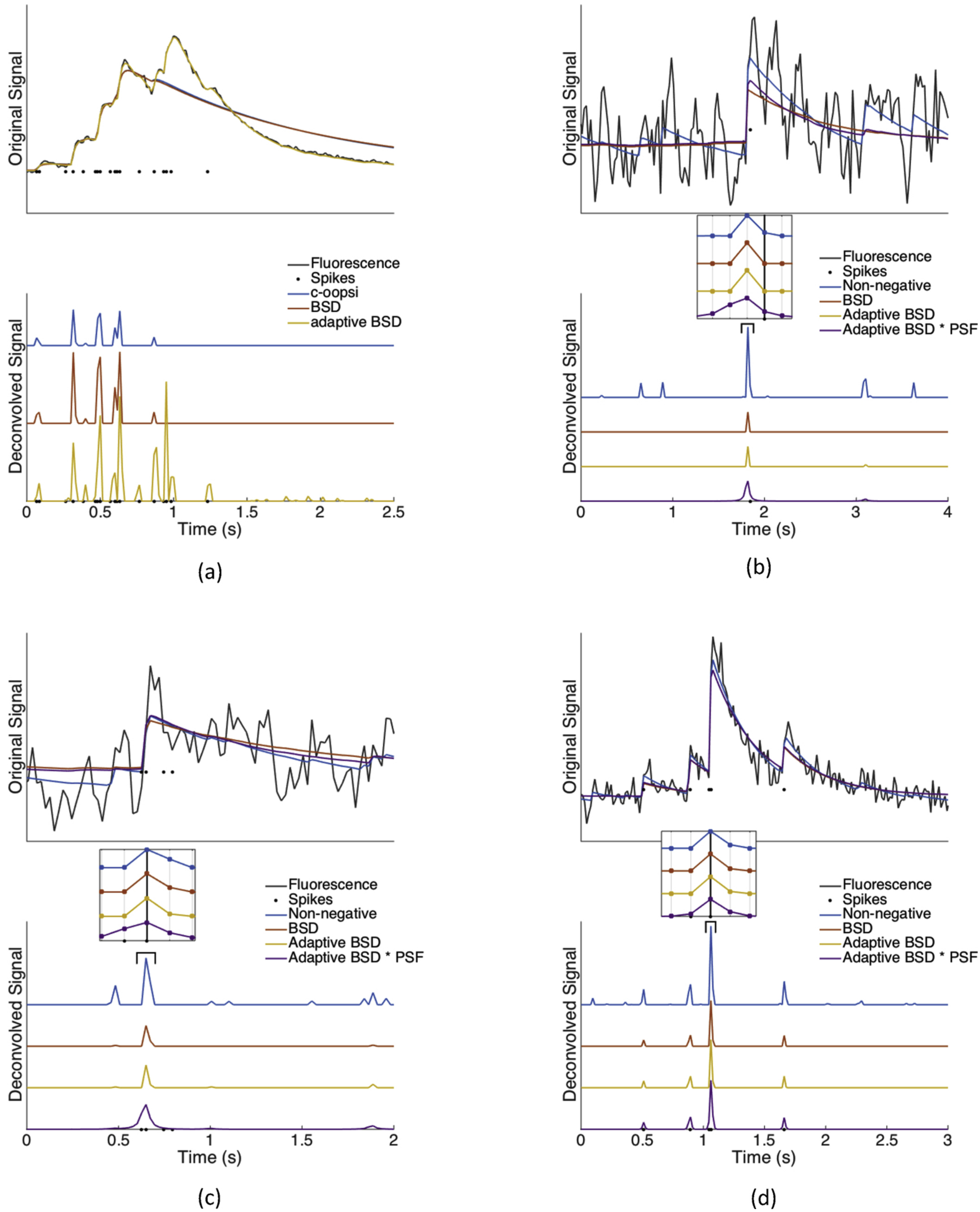

Fig. 6. Selected examples from the SpikeFinder datasets. Four fluorescence traces from the SpikeFinder dataset and the corresponding inferred and ground truth spike trains for various algorithms. (a): GCaMP6s, $f = 60Hz$ (dataset 8), blind initial kernel. (b), (c): OGB-1 $f = 40Hz$ (dataset 1), train set initial kernel (d) GCaMP6f $f = 60Hz$ (dataset 7), train set initial kernel. For panels (b),(c),(d), insets show the same inferred spike trains, zoomed-in on a spike to illustrate timing errors.

Table 3
Correlation scores for the SpikeFinder data (train sets), for various choices of inference algorithms, postprocessings (convolved with the Point-Spread function or not) and initial kernel values: Blind = no information at all; Literature = using parameters derived from the literature, see Supplementary Table VI; Train set = using parameter derived from the train set, see Table 2; Ground-truth = using different parameters for each neuron, derived using the knowledge of the spike positions. The best score for each dataset and configuration is indicated in bold.

Dataset Calcium Indicator	1 OGB-1	2 OGB-1	3 GCaMP6s	4 OGB-1	5 GCaMP6s	6 GCaMP5k	7 GCaMP6f	8 GCaMP6s	9 jRCAMP1a	10 jRGECO1a	Mean Score	
Blind	non-negative	0.378	0.439	0.223	0.416	0.189	0.439	0.649	0.610	0.579	0.760	0.468
	con-oopsi	0.370	0.452	0.253	**0.551**	0.213	0.444	0.659	0.619	0.586	0.728	0.487
	BSD	0.425	0.439	0.277	0.545	0.279	0.533	0.707	0.595	0.583	0.747	0.513
	adaptive BSD	0.451	0.450	0.294	0.542	0.332	**0.621**	**0.747**	**0.686**	**0.625**	**0.835**	0.558
	non-negative * PSF	0.405	0.457	0.333	0.442	0.291	0.464	0.631	0.547	0.526	0.747	0.484
	BSD * PSF	0.450	0.466	0.410	0.544	0.344	0.554	0.681	0.575	0.553	0.758	0.533
	adaptive BSD * PSF	**0.487**	**0.466**	**0.424**	0.537	**0.413**	0.613	0.740	0.674	0.559	0.820	**0.573**
Literature	non-negative	0.372	0.391	0.288	0.419	0.312	0.562	0.605	0.646	0.580	0.760	0.493
	con-oopsi	0.372	0.427	0.288	0.541	0.312	0.562	0.627	0.651	0.585	0.794	0.516
	BSD	0.396	0.435	0.291	0.525	0.305	0.565	0.625	0.649	0.601	0.800	0.519
	adaptive BSD	0.451	0.459	0.453	0.541	0.420	**0.611**	**0.733**	**0.643**	**0.613**	**0.837**	0.576
	non-negative * PSF	0.421	0.398	0.396	0.451	0.411	0.547	0.584	0.601	0.537	0.733	0.508
	BSD * PSF	0.440	0.439	0.368	0.529	0.391	0.549	0.628	0.641	0.572	0.788	0.535
	adaptive BSD * PSF	**0.486**	**0.473**	**0.493**	**0.546**	**0.497**	0.600	0.725	0.620	0.577	0.823	**0.584**
Train Set	non-negative	0.456	0.457	0.441	0.442	0.403	0.600	0.701	0.654	0.588	0.794	0.554
	con-oopsi	0.441	0.458	0.440	**0.550**	0.405	0.600	0.715	0.654	0.588	0.811	0.564
	BSD	0.446	0.458	0.449	0.539	0.398	0.598	0.734	0.641	0.611	0.803	0.568
	adaptive BSD	0.446	0.459	0.453	0.540	0.420	**0.621**	**0.737**	**0.705**	**0.618**	**0.835**	0.583
	non-negative * PSF	0.488	0.467	0.485	0.473	0.479	0.598	0.681	0.641	0.536	0.777	0.563
	BSD * PSF	**0.497**	0.470	0.491	0.543	0.475	0.599	0.717	0.645	0.582	0.802	0.582
	adaptive BSD * PSF	0.495	**0.475**	**0.493**	0.546	**0.498**	0.613	0.730	0.684	0.574	0.817	**0.593**
Ground Truth	non-negative	0.461	0.460	0.270	0.444	0.395	0.613	0.723	0.681	0.592	0.777	0.542
	con-oopsi	0.461	0.462	0.270	0.550	0.395	0.613	0.741	0.681	0.592	0.789	0.555
	BSD	0.446	0.456	0.269	0.539	0.390	0.610	0.747	0.671	0.609	0.790	0.553

3.4. Light-sheet Imaging of Zebrafish

Compared to standard fluorescence microscopy techniques, such as confocal or two-photon epifluorescence microscopy, light-sheet imaging allows for a parallelization of the recording, yielding ~100-fold increase in data-throughput (Panier et al., 2013; Ahrens et al., 2013; Wolf et al., 2015). 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, as well as MLspike, one of the top performing algorithms of the SpikeFinder contest (Deneux et al., 2016) on a typical whole-brain recording, consisting of 1800 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 pan-neurally. After segmentation, 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 3.1.1, we find that BSD achieves a 7-fold increase in speed compared to con-oopsi, and a X-fold increase compared to MLspike, see Table 5. Under these

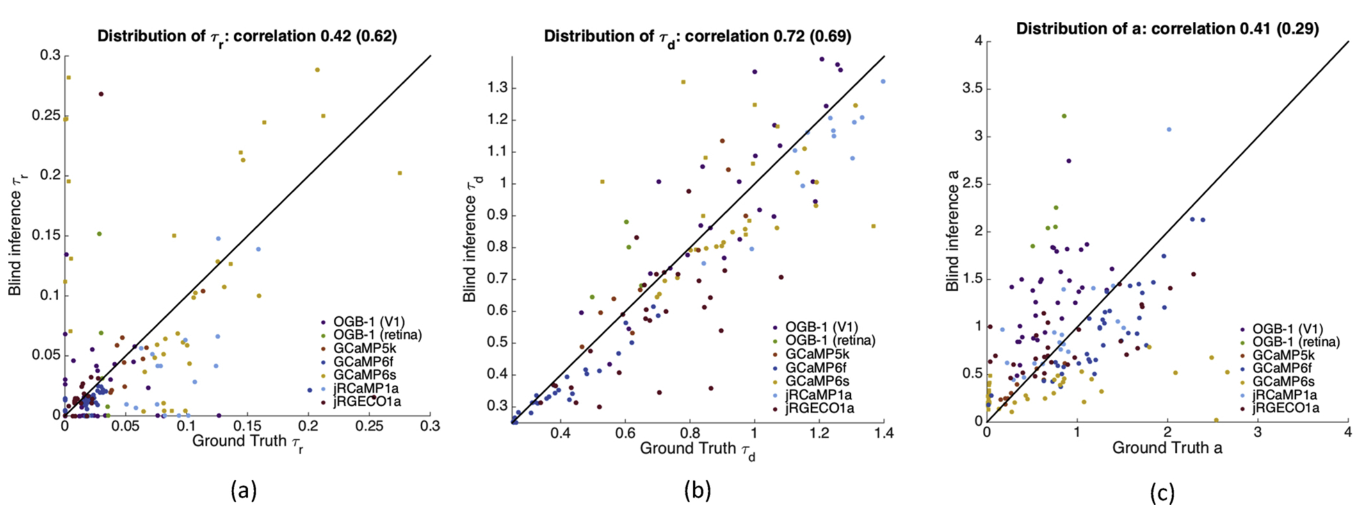

Fig. 7. Evaluation of the generative models parameter inference for the ten SpikeFinder datasets. For each recording, we derive 'ground-truth' generative model parameters τ_r,τ_d,a by a temporal regression of the fluorescence against the spikes measured by electrophysiology, see Section 2.7.3. The values are compared to blind estimates obtained using adaptive BSD. (a) Rise time. (b) Decay time. (c) Transient amplitude a (each fluorescence trace is normalized to unit variance). Recordings from datasets 3 and 5 are represented as squares, and the others as disks. The correlation coefficient are computed either using all datasets or excluding datasets 3 and 5 (values in parenthesis), for which the ground-truth is inaccurate due to offset between electrophysiological and fluorescence recordings.

Table 4
Empirical and predicted point-spread functions for the SpikeFinder challenge. For each dataset and each algorithm, the point-spread function is fitted with a Gaussian distribution of mean μ and standard deviation σ. the latter are displayed as $\mu \pm \sigma$ in ms. For super-resolution BSD, we also indicate in parenthesis the relative gain in mean square temporal error $\sqrt{\mu^2 + \sigma^2}$ compared to regular BSD.

Dataset Calcium indicator	1 OGB-1	2 OGB-1	3 GCaMP6s	4 OGB-1	5 GCaMP6s	6 GCaMP5k	7 GCaMP6f	8 GCaMP6s	9 jRCAMP1a	10 jRGECO1a
BSD (train set)	−1 ± 36	−36 ± 66	115 ± 194	−34 ± 96	142 ± 117	4 ± 19	3 ± 12	−4 ± 16	14 ± 39	13 ± 22
adaptive BSD (blind)	−1 ± 37	−42 ± 67	3 ± 136	−54 ± 111	31 ± 72	5 ± 19	5 ± 13	5 ± 19	35 ± 46	14 ± 23
adaptive BSD (train set)	−2 ± 36	−37 ± 66	98 ± 175	−31 ± 96	121 ± 93	5 ± 19	6 ± 13	1 ± 14	32 ± 46	16 ± 23
adaptive BSD + SR (train set)	/	/	/	/	/	2 ± 16 (-16%)	4 ± 11 (-22%)	−3 ± 11 (-15%)	19 ± 30 (-37%)	14 ± 19 (-14%)
BSD (ground truth)	−8 ± 33	−40 ± 67	15 ± 129	−32 ± 87	132 ± 115	4 ± 19	3 ± 12	−6 ± 18	11 ± 42	14 ± 22
BSD + SR (ground truth)	/	/	/	/	/	0 ± 16 (-16%)	0 ± 11 (-15%)	−9 ± 16 (-2%)	−13 ± 25 (-36%)	10 ± 16 (-28%)
Predicted (train set)	9 ± 43	14 ± 66	0 ± 19	23 ± 167	0 ± 17	9 ± 19	7 ± 9	8 ± 12	28 ± 49	9 ± 17
Predicted (blind)	5 ± 24	3 ± 46	3 ± 25	1 ± 60	2 ± 19	9 ± 16	5 ± 10	8 ± 18	24 ± 50	8 ± 19

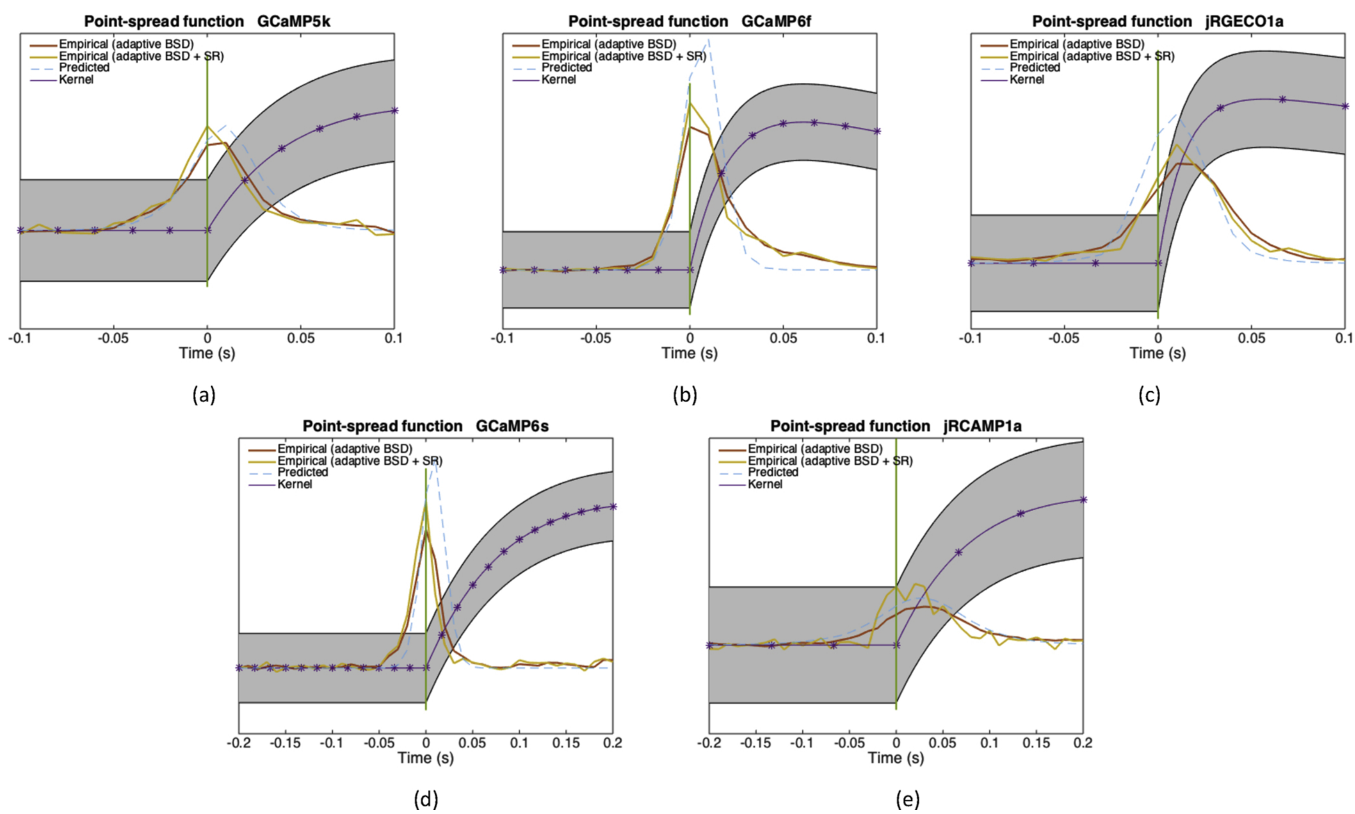

Fig. 8. Temporal accuracy of BSD for datasets 6–10 of SpikeFinder. For each dataset, spikes are inferred with adaptive BSD either using super-resolution or not. After resampling to 100 Hz, empirical point-spread function are computed and averaged for each dataset. Dashed blue curves denote the predicted point-spread function using Eq. (29) and Table 2. The corresponding fluorescence kernel is displayed for comparison. Asterisks indicate fluorescence measurements.

Table 5
Time for perfoming deconvolution on voxelated data.

Algorithm	Total run time	Average run time per voxel
con-oopsi	124 min	0.38s (min: 0.31 s, max: 110 s)
MLspike	120 min	0.37 s
BSD	18 min	0.051 s

experimental conditions, the computation time with BSD matches the duration of the experiment itself (20 min), and is thus compatible with 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 DF/F signals. For this purpose, we use a 2D recording acquired at 20 frame/second for 20 min 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 Appendix H). The average values of these two time-constants are then used to perform spike inference.

Fig. 9a 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 Fig. 9b. As expected, the deconvolved spike trace appear much sparser

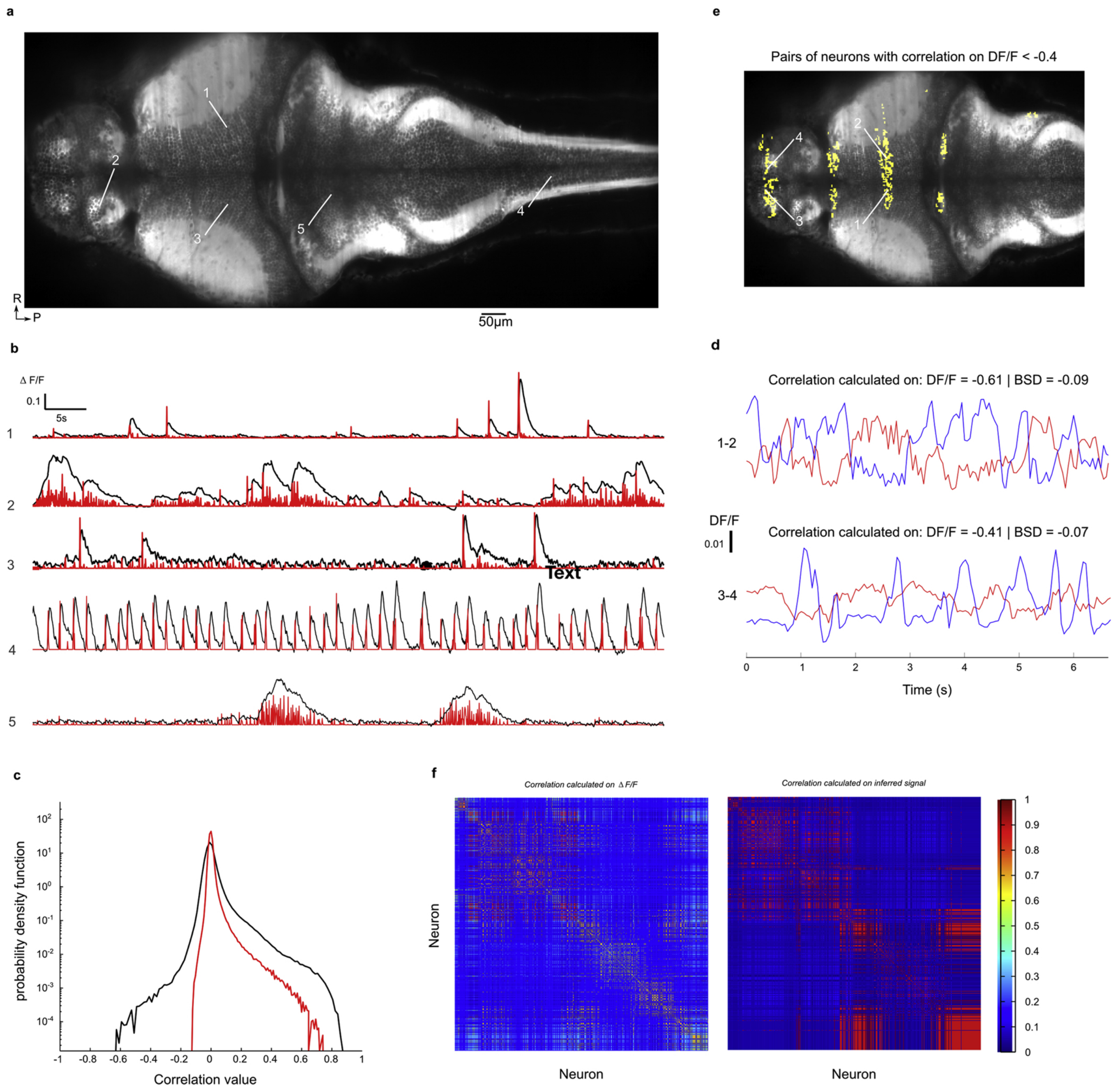

Fig. 9. (a) Bottom: Individual traces of 5 neurons recorded at 20Hz from 6dpf larvae, in black curve DF/F, in red resulting signal from BSD deconvolution algorithm. Top: Time-averaged image of a brain slice of the larva, the white arrows give the location of the 5 neurons. (b) Distribution of pair-wise correlations of DF/F (black) and signal after BSD deconvolution. The data were obtained from a 20 Hz, 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.

and less noisy than the original fluorescent signal. The pair-wise correlations, corrected for uniform coherent noise, are then computed for both the raw DF/F signal and the inferred spike traces. We find the correlation distribution to be much more peaked after deconvolution (Fig. 9c) which reflects in the more uniform appearance of the associated correlation matrix (Fig. 9d).

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

(Fig. 9e). For the same neuronal pairs, the correlation value computed from the inferred signal is thus largely reduced (Fig. 9f).

4. Discussion

The last few years have seen the release of numerous spike inference algorithms (Berens et al., 2018). Their increasing complexity make them poorly interpretable and their performance, beyond the specific conditions for which they have been optimized, can thus be difficult to predict. This is particularly true for supervised machine learning approaches. Although they can offer excellent results on datasets on which they have been trained (Theis et al., 2016) – which requires the availability of a ground truth, *i.e.* a simultaneous electrophysiological measurement of the actual spike train – they become less reliable when generalized to other experimental conditions and/or calcium reporters.

In this context, forward generative models, such as non-negative sparse deconvolution, offer a more robust and tractable solution to this problem. These models are based on explicit hypothesis regarding the form of the fluorescence response kernel, the statistics of the spike train and the noise signal, and are thus less prone to systematic bias. However, their performance can still be very sensitive to the way the different model parameters are set, as exemplified by the unreliable results that we obtained with the oopsi algorithm. Due to a paucity of theoretical understanding of the expected performances, these failures are generally impossible to anticipate. The broad implementation of inference methods in functional imaging laboratories will thus depend on the robustness of these algorithms as much as on their optimal performance. Neuroscientists need algorithms that are not only efficient and fast, to accommodate the rapidly growing size of calcium imaging datasets, but that also provide them with a reliable way to assess the quality of the inferred signals.

Here we introduced a novel non-negative sparse algorithm, named blind sparse deconvolution (BSD), which was designed to specifically address these issues. This fully unsupervised algorithm features state-of-the-art computational speed, accuracy and adaptability while incorporating a 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, allowing one to choose, in a given experimental context and for a given calcium reporter, the recording rate that will provide the optimal temporal resolution. They also can be used *a posteriori* to estimate error rates and thus provide bounds on the reliability of the inferred spike trains.

One of the main assets of BSD, compared to other generative models, owes to the fact that most model parameters are analytically derived, in particular the sparsity prior. This allows a tractablity of the algorithm, but also a gain in speed compared to other approaches (such as constrained-oopsi) that require recursive evaluations of the sparsity prior.

The main output of the algorithm is a continuous signal that approximates the spike-evoked calcium influx for each recorded neuron. This signal can be directlty used to characterize *e.g.* the tuning properties of a sensory-responsive neuron through correlation with the sensory input. BSD also provides automatic theoretically-grounded thresholding and thus enables to binarize the signal into active and non-active periods. Although such binarization comes at a cost of a loss of information, as revealed by the cross-correlation with the actual spike train, it can be necessary for the implementation of graphical circuit inference approaches, such as RBM (Cocco et al., 2018) or Ising models, (Meshulam et al., 2017; Cocco et al., 2017; Posani et al., 2018). These models aim at interpreting the collective dynamics of large neuronal ensembles by infering effective interactions between neurons using the measured pairwise firing statistics. In this particular context, the knowledge of the temporal PSF, as offered by BSD, is also very beneficial as it indicates the minimal time bin over which the pairwise correlations can be robustly evaluated.

In calcium imaging, the temporal resolution is generally thought to be limited by the recording rate. However, at high enough SNR, one may in principle overcome this limit. BSD thus introduces temporal super-resolution, which was shown to significantly increase the temporal accuracy of inferred spikes on real data, yielding a temporal resolution better than the recording period. As both calcium reporters and imaging methods will gain in sensitivity and speed, this capability may help to reveal spatio-temporal short term dynamics, such as activity propagating waves, or to investigate the role of spike-timing in neural coding.

Although we demonstrated that BSD provides consistent results over a large spectrum of reporters and experimental conditions, this algorithm would benefit from embedding more features that would address specific conditions. In particular, we noticed that the performance tends to degrade when neurons display sustained periods of bursting activity. First, the moving percentile method may provide inconsistent estimate of the slowly drifting baseline in this case. [We have shown however that the iterative baseline estimation allows to partially correct this issue]. Second, kernel inference and SNR estimate become less reliable as bursts of activity are mistaken for individual spiking events. One way around would consist in constraining the inferred parameters based on their values estimated on neurons exhibiting sparser activity. A second aspect, ignored in the present implementation, is the non-linearity of the fluorescence response of usual calcium reporters. Since the predicted signal is continuous, this could be accounted for a posteriori in a straightforward way provided that an experimental characterization of the fluorescence vs spike-rate relationship is available. Importantly, the modularity of the different features introduced in BSD – sparsity parameter estimation, iterative kernel iteration and PSF estimation – makes it straighforward to incorporate them into complete calcium imaging packages such as CaIman (Giovannucci et al., 2019) Suite2P (Pachitariu et al., 2017) that address other challenges of calcium imaging processing, such as spatial filtering.

Availability

The spike inference program BSD and its companion program for evaluating its accuracy are implemented in MATLAB and both available at https://github.com/jertubiana/BSD. By design, it is straightforward to infer spike trains and evaluate a posteriori the precision–recall and temporal accuracy for each recording. Then, users can optionally convolve the inferred spikes with the predicted point-spread function to account for the uncertainty on the spike location. Tutorial scripts are provided for all use cases and the scripts for generating all figures of this article and reproducing the SpikeFinder experiments are also made available.

Authors' contribution

Jérôme Tubiana: conceptualization, methodology, software, formal analysis, investigation, writing – original draft, visualization.

Sébastien Wolf: conceptualization, validation, investigation, writing – original draft, visualization.

Thomas Panier: investigation.

Georges Debregeas: conceptualization, validation, resources, writing – original draft, writing – review & editing, supervision, project administration, funding acquisition.

Conflict of interest

None declared.

Acknowledgements

J.T. acknowledges partial support by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. This work was funded by the Human Frontier Science Program (RGP0060/2017) and the Fondation pour la Recherche Médicale (SPF201809007064).

Appendix A. Stability of the single spike solution and the half-spike problem

In Section 3, we have not studied the stability of the single-spike solution. We study it here, and discuss when it is a global optimum. Assuming $\widehat{N_i} = \delta_{i,i_0}\max\left\{a - \frac{\lambda}{\sum_i K^2(t)} + \sigma\frac{\sum_i K(t)\varepsilon_i}{\sum_i K^2(t)}, 0\right\} > 0$ and looking for the stability of the solution, w.r.t the other coordinates, we find:

$$-\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}} = \sigma \sum_i \{K(\Delta t(i-i_0-\delta)) - K(\Delta t(i-i_0)\cos(\theta(\delta\Delta t)\}\varepsilon_i - \lambda(1-\cos(\theta(\delta\Delta t))) \tag{47}$$

$$P\left(-\frac{\partial \mathcal{L}}{\partial N_{i_0+\delta}} > 0\right) = \Phi\left[\frac{\lambda \tan\frac{\theta(\delta\Delta t)}{2}}{\sigma||K||}\right] \tag{48}$$

where $\Phi(x) = \int_x^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\mathrm{d}u$. 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 $N_i = \frac{\mathrm{an}}{2}(\delta_{i,i_0} + \delta_{i,i_0\pm 1})$

Appendix B. Impact of kernel parameters mismatch on inference

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 Fig. 10a–d, inferring the spikes with an incorrect convolution kernel leads to systematic errors.The nature of the error depends on the kernel mismatch:

- $\tau_r < \tau_r^0$, $\tau_d = \tau_d^0$ (Fig. 10a): the inferred spikes are split in two, to compensate for the smaller rise time than expected for a single spike.
- $\tau_r > \tau_r^0$, $\tau_d < \tau_d^0$ (Fig. 10b): the inferred spikes are in advance, to compensate for the faster rise of the fluorescence signal.
- $\tau_r = \tau_r^0$, $\tau_d < \tau_d^0$ (Fig. 10c): the inferred spikes exhibit 'echos' to compensate for the slower than expected decay of F.
- $\tau_r = \tau_r^0$, $\tau_d > \tau_d^0$ (Fig. 10d): the inferred subsequent spikes are 'screened' (lower amplitude) to compensate for the slower than expected signal decay.

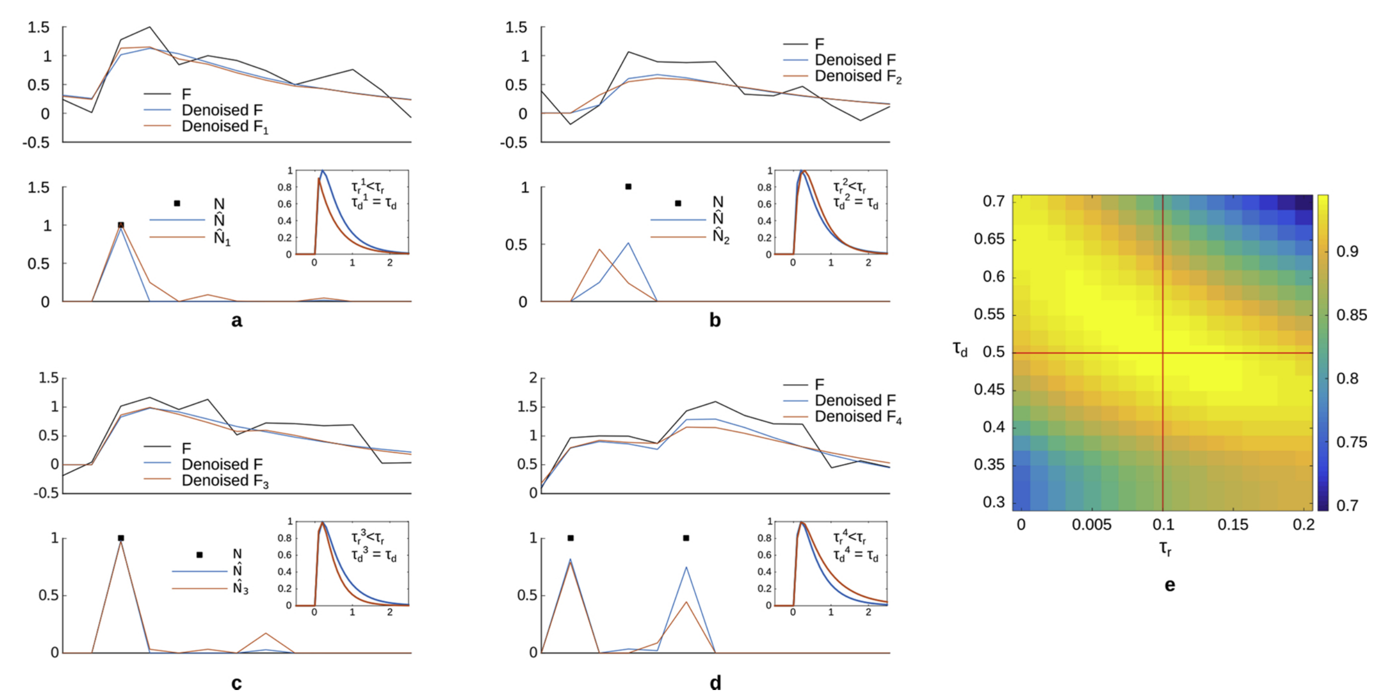

Fig. 10. Left: Example results of spike inference on synthetic data with mismatched convolution kernels. For each of the four figures, a fluorescence signal is generated with a kernel K^0; inference is performed with true parameters (blue curves) and with mismatched parameters (red curves). The true 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 $\delta t = 0$ s, 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 (Fig. 10e). 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. These findings motivate the use of parameter refinement.

Appendix C. 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 $\widehat{\mathbf{N}} = \mathrm{argmin}_{N\geq 0}\,\mathcal{L}(\mathbf{N}, \mathbf{K})$.

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

$$
\begin{aligned}
n \quad &= \max\left\{\frac{\cos\phi_K\,\|K^0\|}{\|K\|} + \frac{\sigma}{a\|K\|}(\tilde{\varepsilon}_1 - z),\, 0\right\}\\
\mathcal{L}(\hat{\mathbf{N}}, \mathbf{K}) \quad &= \frac{1}{2}a^2\|K^0\|^2(1 - \cos^2\theta_K) + \mathrm{z}\,\mathrm{a}\sigma\|K^0\|\cos\theta_K\\
&+ \frac{\sigma^2}{2}\left(\sum_i \varepsilon_i^2 - \tilde{\varepsilon}_1^2 - z^2\right) + \sigma^2 z\tilde{\varepsilon n}_1 - a\sigma\|K^0\|\sqrt{2(1 - \cos\phi_K)}\,\tilde{\varepsilon}_2
\end{aligned}
\tag{49}
$$

where

$$
\begin{aligned}
\|K\| &= \sqrt{\sum_l K(l\Delta t)^2}\\
\|K^0\| &= \sqrt{\sum_l K^0(l\Delta t)^2}\\
\cos\phi_K &= \frac{\sum_i K^0[(i - i_0 + 1)\Delta t]K[(i - i_0 + 1)\Delta t]}{\|K\|\|K^0\|}
\end{aligned}
\tag{50}
$$

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

$$
\mathcal{L}(\widehat{\mathbf{N}}, \mathbf{K}) \propto \; < \mathcal{L}(\widehat{\mathbf{N}}, \mathbf{K}) > \; \propto \frac{1}{2}a^2\|K^0\|^2(1 - \cos^2\phi_K) + \mathrm{z}\,\mathrm{a}\sigma\|K^0\|\cos\phi_K + \text{Constant}
\tag{51}
$$

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 cannot 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:

$$
\begin{aligned}
\frac{1}{2}\|F - \mathrm{K}\,\mathrm{N}\|^2 &= \frac{1}{2}(F^TF - 2F^T\mathcal{K}N + N^T\mathcal{K}^T\mathcal{K}N)\\
&= \frac{1}{2}(F^TF - 2\mathrm{Trace}[\mathcal{K}NF^T] + \mathrm{Trace}[(\mathcal{K}^T\mathcal{K})NN^T])\\
&= \frac{1}{2}\left\{\sum_{i=1}^{T} F_i^2 - 2\sum_{l=0}^{\infty} K(\Delta t(l + 1))\left(\sum_{i=1}^{T-l} F_{i+l}N_i\right)\right.\\
&\quad + \sum_{l=-\infty}^{\infty}\left(\sum_{i=\max(1,1-l)}^{\min(T,T-l)} N_{i+l}N_i\right)\left(\sum_{j=-\infty}^{\infty} K[\Delta t\,\mathrm{j}]K[\Delta t(j + l)]\right)\\
&\quad \left. - \sum_{j=T+1}^{\infty}\left(\sum_i K[(j - i + 1)\Delta t]N_i\right)^2\right\}
\end{aligned}
\tag{52}
$$

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 $\sum_{j=T+1}^{\infty}\left(\sum_{i=1}^{T} K[(j - i + 1)\Delta t]N_i\right)^2$

The first three terms can be precomputed in $O(T)$ 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 $O(l_{\max})$, 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 $O(l_{\max})$ and optimized efficiently.

Appendix D. Detailed computations for the point spread function estimation

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

$$
\begin{aligned}
N_i &= \mathrm{a}\,\mathrm{n}\delta_{i,i_0+\delta}\\
\mathcal{L}(n,\delta) &= \frac{1}{2}\sum_i F_i^2 + \frac{a^2}{2}\left\{-2n\sum_i K[\Delta t(i-i_0+1-r)]K[\Delta t(i-i_0+1-\delta)]\right.\\
&\left.+ n^2\sum_i K[\Delta t(i-i_0+1-\delta)]^2\right\} - a\sigma\sum_i K[\Delta t(i-i_0+1-\delta)]\varepsilon_i + \lambda\mathrm{a}\,\mathrm{n}\\
n^\delta &= \underset{n\ge 0}{\mathrm{argmax}}\,\mathcal{L}(n,\delta) = \frac{\min\left\{\sum_i K[\Delta t(i-i_0+1-r)]K[\Delta t(i-i_0+1-\delta)] + \frac{\sigma}{a}\sum_i K[\Delta t(i-i_0+1-\delta)]\varepsilon_i - \frac{\lambda}{a}, 0\right\}}{\sum_i K[\Delta t(i-i_0+1-\delta)]^2}\\
\mathcal{L}(n^\delta,\delta) &= \frac{1}{2}\sum_i F_i^2 + \frac{\min\left\{\sum_i K[\Delta t(i-i_0+1-r)]K[\Delta t(i-i_0+1-\delta)] + \frac{\sigma}{a}\sum_i K[\Delta t(i-i_0+1-\delta)]\varepsilon_i - \frac{\lambda}{a}, 0\right\}}{2\sum_i K[\Delta t(i-i_0+1-\delta)]^2}
\end{aligned}
\tag{53}
$$

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

$$
\begin{aligned}
\tilde{\varepsilon}_\delta &= \sum_i K[\Delta t(i-i_0+1-\delta)]\varepsilon_i\\
\langle\tilde{\varepsilon}_\delta\rangle &= 0\\
\langle\tilde{\varepsilon}_\delta\tilde{\varepsilon}_{\delta'}\rangle &= \sum_i K[\Delta t(i-i_0+1-\delta)]K[\Delta t(i-i_0+1-\delta')]\\
\mathcal{L}(n^\delta,\delta) &= \frac{1}{2}\sum_i F_i^2 + \frac{\min\left\{\rho_{r,\delta} + \frac{\sigma}{a}\tilde{\varepsilon}_\delta - \frac{\lambda}{a}, 0\right\}}{2\|K\|^2}
\end{aligned}
\tag{54}
$$

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 point spread 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 point spread 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.

Appendix E. 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 $F_i = \mathrm{a}\,\mathrm{K}\left[\Delta t(i - i_0) + \frac{\Delta t(2-r_0)}{s}\right] + \sigma\varepsilon_i$.

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

$$
\begin{aligned}
n &= \max\left\{\frac{\|K_{r_0}\|}{\|K_{r_0+\delta}\|}\cos\theta_{\Delta,\delta} + \frac{\sigma}{a}\|K_{r_0+\delta}\|(\tilde{\varepsilon}_1 - z)\right\}\\
\mathcal{L}(\hat{\mathbf{N}},\mathbf{K}) &= \frac{1}{2}\|K_{r_0}\|^2(1-\cos^2\theta_{\Delta,\delta}) + \mathrm{z}\,\mathrm{a}\sigma\|K_r\|\cos\theta_{\Delta,\delta}\\
&+ \frac{\sigma^2}{2}\left(\sum_i \varepsilon_i^2 - \tilde{\varepsilon}_1^2 - z^2\right) + \sigma^2 z\tilde{\varepsilon}_1 - a\sigma\|K_r\|\sqrt{2(1-\cos\theta_{\Delta,\delta})}\,\tilde{\varepsilon}_2
\end{aligned}
\tag{55}
$$

where

$$
\begin{aligned}
\|K_r\| &= \sqrt{\sum_l K\left[l\Delta t + \frac{\Delta t(s+1-r)}{s}\right]^2}\\
\cos\theta_{\Delta,\delta} &= \frac{\sum_i K\left[i\Delta t + \frac{\Delta t(s+1-r)}{s}\right]K\left[(i+\Delta)\Delta t + \frac{\Delta t(s+1-r')}{s}\right]}{\|K_r\|\|K_{r'}\|}
\end{aligned}
\tag{56}
$$

and $\tilde{\varepsilon}_1$, $\tilde{\varepsilon}_2$ are gaussian noises of variance unity (see Appendix D). 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 Section A, 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 $\sum_i K_{ij}^2 = 1\ \forall j$.

Appendix F. 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, Eq. (30) becomes:

$$\begin{aligned} A_F(l) - \sigma^2\delta_{l,0} &= a^2\nu\frac{\Delta t}{s}\sum_{k=1}^{\mathrm{Ts}} K\left[\Delta t\,\mathrm{i} - \frac{\Delta t}{s}(k-1)\right]K\left[\Delta t(i+l) + \frac{\Delta t}{s}(k-1)\right] \\ &= a^2\nu\Delta t\sum_{j=1}^{T}\frac{1}{s}\left(\sum_{r=1}^{s}K\left[\Delta t(i-j-1) + \frac{\Delta t(s+1-r)}{s}\right]K\left[\Delta t(i+l-j-1) + \frac{\Delta t(s+1-r)}{s}\right]\right) \\ &\approx a^2\nu\Delta t\sum_{m=-\infty}^{\infty}\frac{1}{s}\left(\sum_{r=1}^{s}K\left[\Delta t\,\mathrm{m} + \frac{\Delta t(s+1-r)}{s}\right]K\left[\Delta t(m+l) + \frac{\Delta t(s+1-r)}{s}\right]\right) \end{aligned} \tag{57}$$

For $s > 1$, this formula is different from Eq. (30). It can be shown (see Appendix G) that the right-hand side has a well-defined limit when $s \to \infty$, ie in the continuous setting.

Similarly, the iterative kernel update Eq. (52) is different:

$$\begin{aligned} &\frac{1}{2}\|F - \mathrm{K\,N}\|^2 \\ &= \frac{1}{2}\left\{\sum_{i=1}^{T}F_i^2 - 2\sum_{l=-\infty}^{\infty}\sum_{r=1}^{s}K\left[\Delta t\,\mathrm{l} + \frac{s+1-r}{s}\right]\left(\sum_{i=1}^{T}F_{i+l}N_{s(i-1)+r}\right)\right. \\ &\qquad + \sum_{l=-\infty}^{\infty}\sum_{r=1}^{s}\sum_{r'=1}^{s}\left(\sum_{m=-\infty}^{\infty}K\left[\Delta t\,\mathrm{m} + \frac{\Delta t(s+1-r)}{s}\right]K\left[\Delta t(m+l)\right.\right. \\ &\qquad \left.\left.\left. + \frac{\Delta t(s+1-r')}{s}\right]\right)\left(\sum_{i=1}^{T}N_{(i+l-1)s+r}N_{(i-1)s+r'}\right)\right\} \end{aligned} \tag{58}$$

The sparsity penalty becomes:

$$\lambda^T N = \sum_{j=1}^{\mathrm{sT}}\lambda_j N_j = z\sigma\sum_{r=1}^{s}\sqrt{\sum_{m=-\infty}^{\infty}K\left[m\Delta t + \frac{(s+1-r)\Delta t}{s}\right]}\left(\sum_{i=1}^{T}N_{(i-1)s+r}\right) \tag{59}$$

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

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

Appendix G. 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:

$$\begin{aligned} K(t) &= \frac{1}{M(\tau_r, \tau_d)}\left[e^{-\frac{t}{\tau_d}} - e^{-\frac{t}{\tau_r}}\right]1_{t\geq 0} \\ M(\tau_r, \tau_d) &= \left(\frac{\tau_r}{\tau_d}\right)^{-\frac{\tau_r}{\tau_d-\tau_r}} - \left(\frac{\tau_r}{\tau_d}\right)^{\frac{\tau_d}{\tau_d-\tau_r}} \end{aligned} \tag{60}$$

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

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

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:

$$\|K_r\| = \frac{1}{M(\tau_r,\tau_d)}\sqrt{\sum_i K\left[\Delta i + \frac{\Delta t(s+1-r)}{s}\right]^2} = \sqrt{\frac{\lambda_d^{2\frac{s+1-r}{s}}}{1-\lambda_d^2} - 2\frac{(\lambda_d\lambda_r)^{\frac{s+1-r}{s}}}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{2\frac{s+1-r}{s}}}{1-\lambda_r^2}}$$
$$\|K_r\|_1 = \frac{1}{M(\tau_r,\tau_d)}\frac{\lambda_d^{\frac{s+1-r}{s}}}{1-\lambda_d} + \frac{\lambda_r^{\frac{s+1-r}{s}}}{1-\lambda_r} \tag{62}$$

Kernel overlaps. Useful for assessing temporal uncertainty and for kernel estimation

$$\cos\theta(\delta\Delta t) \equiv \frac{\sum_{i=-\infty}^{+\infty} K[i\Delta t]K[(i+\delta)\Delta t]}{\|K\|^2} = \frac{\frac{\lambda_d^{2+\delta}}{1-\lambda_d^2} - \frac{(\lambda_d^\delta+\lambda_r^\delta)\lambda_d\lambda_r}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{2+\delta}}{1-\lambda_r^2}}{\frac{\lambda_d^2}{1-\lambda_d^2} - \frac{2\lambda_d\lambda_r}{1-\lambda_d\lambda_r} + \frac{\lambda_r^2}{1-\lambda_r^2}} \tag{63}$$

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

$$\sum_{j=T+1}^{\infty}\left(\sum_{i=1}^{T} K[(j-i+1)\Delta t]N_i\right)^2 = \frac{1}{M(\tau_r,\tau_d)^2}\sum_{j=T+1}^{\infty}\left(\lambda_d^{j-T}\left[\sum_i \lambda_d^{T-i+1}N_i\right] - \lambda_r^{j-T}\left[\sum_i \lambda_r^{T-i+1}N_i\right]\right)^2$$
$$= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^4\left(\sum_i \lambda_d^{T-i}N_i\right)^2}{1-\lambda_d^2} - \frac{2(\lambda_d\lambda_r)^2\left(\sum_i \lambda_r^{T-i}N_i\right)\left(\sum_i \lambda_d^{T-i}N_i\right)}{1-\lambda_d\lambda_r} + \frac{\lambda_r^4\left(\sum_i \lambda_r^{T-i}N_i\right)^2}{1-\lambda_r^2}\right) \tag{64}$$

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

$$A_K(l,r_1,r_2) = \sum_i K\left[(i+l+\frac{s+1-r_1}{s})\Delta t\right]K\left[(i+\frac{s+1-r_2}{s})\Delta t\right]$$
$$= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^{l+\frac{2(s+1)-r_1-r_2}{s}}}{1-\lambda_d^2} - \frac{\lambda_d^{l+\frac{s+1-r_1}{s}}\lambda_r^{\frac{s+1-r_2}{s}} + \lambda_r^{l+\frac{s+1-r_1}{s}}\lambda_d^{\frac{s+1-r_2}{s}}}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{l+\frac{2(s+1)-r_1-r_2}{s}}}{1-\lambda_r^2}\right) \tag{65}$$

In particular:

$$\frac{1}{s}\sum_{r=1}^{s} A_K(0,r,r) = \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\phi_s(\lambda_d^2)}{1-\lambda_d^2} - \frac{\phi_s(\lambda_d\lambda_r)}{1-\lambda_d\lambda_r} + \frac{\phi_s(\lambda_r^2)}{1-\lambda_r^2}\right)$$
$$\phi_s(x) = \frac{1-x}{s(-1+x^{-\frac{1}{s}})} \tag{66}$$

Boundary-term for super-resolution

$$\sum_{i=T+1}^{\infty}\left(\sum_{j=1}^{sT} K\left[i\Delta t - \frac{(j-1)\Delta t}{s}\right]N_j\right)^2$$
$$= \frac{1}{M(\tau_r,\tau_d)^2}$$
$$\sum_{i=T+1}^{\infty}\left(\lambda_d^{i-T}\left[\sum_i \lambda_d^{\frac{Ts-j+1}{s}}N_j\right] - \lambda_r^{i-T}\left[\sum_i \lambda_r^{\frac{Ts-j+1}{s}}N_j\right]\right)^2$$
$$= \frac{1}{M(\tau_r,\tau_d)^2}\left(\frac{\lambda_d^{\frac{4}{s}}\left(\sum_j \lambda_d^{\frac{Ts-j}{s}}N_j\right)^2}{1-\lambda_d^2} - \frac{2(\lambda_d\lambda_r)^{\frac{2}{s}}\left(\sum_j \lambda_r^{\frac{Ts-j}{s}}N_j\right)\left(\sum_j \lambda_d^{\frac{Ts-j}{s}}N_j\right)}{1-\lambda_d\lambda_r} + \frac{\lambda_r^{\frac{4}{s}}\left(\sum_j \lambda_r^{\frac{Ts-j}{s}}N_j\right)^2}{1-\lambda_r^2}\right) \tag{67}$$

Appendix H. 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: for instance neurons in the spinal chord tend to have longer rising time constants than neurons in the hindbrain, and neuropil regions have longer decay time. 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.

Appendix I. Drawback of approximating Poisson prior to exponential prior in MAP

In order to see why approximating a Bernoulli or Poisson distribution with an exponential approximation with same mean, as is done in oopsi can be problematic for signal reconstruction, we consider the following single-variable inference problem:

$$y = \mathrm{aN} + \varepsilon \tag{68}$$

where $\varepsilon \sim 0, \sigma^2, N \sim \text{Bernoulli}(\nu)$. Given y, we wish to estimate N by MAP using either the exact or approximate prior. In the first case, $P(N|y)$ writes:

$$p(N=1|y) = \frac{e^{-\frac{(y-a)^2}{2\sigma^2}\nu}}{e^{-\frac{(y-a)^2}{2\sigma^2}\nu} + e^{-\frac{y^2}{2\sigma^2}(1-\nu)}} \tag{69}$$

such that:

$$N^\star = \arg\max p(N|y) = \begin{cases} 1 & \text{if } y > \frac{a}{2} - \frac{\sigma^2}{a}\log\frac{\nu}{1-\nu} \\ 0 & \text{Otherwise} \end{cases} \tag{70}$$

In the second case, $P(N|y)$ writes:

$$P(N|y) \propto e^{-\frac{(y-\mathrm{aN})^2}{2\sigma^2} - \nu \mathrm{a}\,\mathrm{N}} \tag{71}$$

such that:

$$N^\star = \arg\max p(N|y) = \frac{1}{a}\max(y - \frac{\sigma^2}{\nu}, 0) \tag{72}$$

For typical values such as $a = 1$, $\sigma = 0.2$, $\nu = 0.01$, we get thresholds at respectively 0.32 and 4. Clearly, all spikes ($N = 1$) would be missed using the MAP with approximate exponential prior. On the other hand, the MAP with an exact prior is an unbiased estimate, i.e. such that $< N^\star = \nu$, and correctly reproduces the average spike rate.

Appendix J. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jneumeth.2020.108763.

References

Ahrens, M.B., Orger, M.B., Robson, D.N., Li, J.M., Keller, P.J., 2013. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat. Methods 10 (5), 413–420.

Akerboom, J., Chen, T.-W., Wardill, T.J., Tian, L., Marvin, J.S., Mutlu, S., Calderón, N.C., Esposti, F., Borghuis, B.G., Sun, X.R., et al., 2012. Optimization of a gcamp calcium indicator for neural activity imaging. J. Neurosci. 32 (40), 13819–13840.

Berens, P., Freeman, J., Deneux, T., Chenkov, N., McColgan, T., Speiser, A., Macke, J.H., Turaga, S.C., Mineault, P., Rupprecht, P., et al., 2018. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. PLoS Comput. Biol. 14 (5), e1006157.

Chen, T.-W., Wardill, T.J., Sun, Y., Pulver, S.R., Renninger, S.L., Baohan, A., Schreiter, E.R., Kerr, R.A., Orger, M.B., Jayaraman, V., et al., 2013. Ultrasensitive fluorescent proteins for imaging neuronal activity. Nature 499 (7458), 295–300.

Cocco, S., Monasson, R., Posani, L., Tavoni, G., 2017. Functional networks from inverse modeling of neural population activity. Curr. Opin. Syst. Biol. 3, 103–110.

Cocco, S., Monasson, R., Posani, L., Rosay, S., Tubiana, J., 2018. Statistical physics and representations in real and artificial neural networks. Physica A: Stat. Mech. Appl. 504, 45–76.

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

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

Fernández-Suárez, M., Ting, A.Y., 2008. Fluorescent probes for super-resolution imaging in living cells. Nat. Rev. Mol. Cell Biol. 9 (12), 929–943.

Freeman, W., Durand, F., Weiss, Y., Levin, A., 2009. Understanding and Evaluating Blind Deconvolution Algorithms.

Friedrich, J., Zhou, P., Paninski, L., 2017. Fast online deconvolution of calcium imaging data. PLoS Comput. Biol. 13 (3), e1005423.

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

Giovannucci, A., Friedrich, J., Gunn, P., Kalfon, J., Brown, B.L., Koay, S.A., Taxidis, J., Najafi, F., Gauthier, J.L., Zhou, P., et al., 2019. Caiman an open source tool for scalable calcium imaging data analysis. Elife 8, e38173.

Grewe, B.F., Langer, D., Kasper, H., Kampa, B.M., Helmchen, F., 2010. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. Nat. Methods 7 (5), 399–405.

Heilemann, M., Van De Linde, S., Schüttpelz, M., Kasper, R., Seefeldt, B., Mukherjee, A., Tinnefeld, P., Sauer, M., 2008. Subdiffraction-resolution fluorescence imaging with conventional fluorescent probes. Angew. Chem. Int. Ed. 47 (33), 6172–6176.

Hell, S.W., 2009. Microscopy and its focal switch. Nat. Methods 6 (1), 24–32.

Holekamp, T.F., Turaga, D., Holy, T.E., 2008. Fast three-dimensional fluorescence imaging of activity in neural populations by objective-coupled planar illumination microscopy. Neuron 57 (5), 661–672.

Huang, B., Wang, W., Bates, M., Zhuang, X., 2008. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science 319 (5864), 810–813.

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

Lütcke, H., Gerhard, F., Zenke, F., Gerstner, W., Helmchen, F., 2013 jan. Inference of neuronal network spike dynamics and topology from calcium imaging data. Front. Neural Circuits 7, 201.

Lustig, M., Donoho, D.L., Santos, J.M., Pauly, J.M., 2008. Compressed sensing mri. IEEE Signal Process. Mag. 25 (2), 72–82.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009. Online dictionary learning for sparse coding. Proceedings of the 26th Annual International Conference on Machine Learning 689–696.

Meshulam, L., Gauthier, J.L., Brody, C.D., Tank, D.W., Bialek, W., 2017. Collective behavior of place and non-place neurons in the hippocampal network. Neuron 96 (5), 1178–1191.

Mishchencko, Y., Vogelstein, J.T., Paninski, L., 2011. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Ann. Appl. Stat. 1229–1261.

Pachitariu, M., Stringer, C., Dipoppa, M., Schröder, S., Rossi, L.F., Dalgleish, H., Carandini, M., Harris, K.D., 2017. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. bioRxiv 61507.

Pachitariu, M., Stringer, C., Harris, K.D., 2018. Robustness of spike deconvolution for neuronal calcium imaging. J. Neurosci. 38 (37), 7976–7985.

Panier, T., Romano, S.A., Olive, R., Pietri, T., Sumbre, G., Candelier, R., Debrégeas, G., 2013. Fast functional imaging of multiple brain regions in intact zebrafish larvae using selective plane illumination microscopy. Front. Neural Circuits 7.

Picardo, M.A., Merel, J., Katlowitz, K.A., Vallentin, D., Okobi, D.E., Benezra, S.E., Clary, R.C., Pnevmatikakis, E.A., Paninski, L., Long, M.A., 2016. Population-level representation of a temporal sequence underlying song production in the zebra finch. Neuron 90 (4), 866–876.

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

Pnevmatikakis, E.A., Soudry, D., Gao, Y., Machado, T.A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., et al., 2016. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. Neuron 89 (2), 285–299.

Posani, L., Cocco, S., Monasson, R., 2018. Integration and multiplexing of positional and contextual information by the hippocampal network. PLoS Comput. Biol. 14 (8), e1006320.
Rust, M.J., Bates, M., Zhuang, X., 2006. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). Nat. Methods 3 (10), 793–796.
Sasaki, T., Takahashi, N., Matsuki, N., Ikegaya, Y., 2008. Fast and accurate detection of action potentials from somatic calcium fluctuations. J. Neurophysiol. 100 (3), 1668–1676.
Selesnick, I., 2012. Sparse deconvolution (an MM algorithm). Connexions.
Theis, L., Berens, P., Froudarakis, E., Reimer, J., Rosón, M.R., Baden, T., Euler, T., Tolias, A.S., Bethge, M., 2016. Benchmarking spike rate inference in population calcium imaging. Neuron 90 (3), 471–482.
Vogelstein, J.T., Watson, B.O., Packer, A.M., Yuste, R., Jedynak, B., Paninski, L., 2009. Spike inference from calcium imaging using sequential Monte Carlo methods. Biophys. J. 97 (2), 636–655.
Vogelstein, J.T., Packer, A.M., Machado, T.A., Sippy, T., Babadi, B., Yuste, R., Paninski, L., 2010. Fast nonnegative deconvolution for spike train inference from population calcium imaging. J. Neurophysiol. 104 (6), 3691–3704.
Wilt, B.A., Fitzgerald, J.E., Schnitzer, M.J., 2013. Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing. Biophys. J. 104 (1), 51–62.
Wolf, S., Supatto, W., Debrégeas, G., Mahou, P., Kruglik, S.G., Sintes, J.-M., Beaurepaire, E., Candelier, R., 2015. Whole-brain functional imaging with two-photon light-sheet microscopy. Nat. Methods 12 (5), 379–380.
Yaksi, E., Friedrich, R.W., 2006. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. Nat. Methods 3 (5), 377–383.