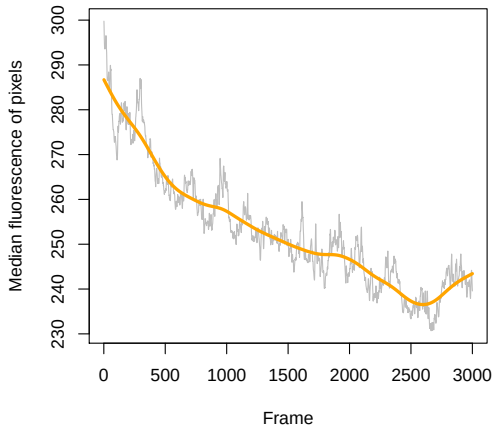

FIG S1. *For the example calcium imaging video, we plot the median fluorescence of the pixels within each frame, along with the smoothing spline fit (▬) that is used to correct for the bleaching effect.*

Supplemental Materials for *SCALPEL: Extracting Neurons from Calcium Imaging Data*

9. Data Pre-Processing. To begin, we perform three pre-processing steps on the raw data. These are briefly described in Step 0 of Section 4. First, we smooth the raw $P \times T$ data matrix spatially and temporally using a Gaussian kernel smoother with a bandwidth of one pixel (Hastie et al., 2009). Second, we adjust for any bleaching effect over time. Specifically, we fit a smoothing spline with 10 degrees of freedom to the median fluorescence for each frame over time, and subtract the frame-specific smoothed median from the corresponding frame. The smoothing spline fit for the example calcium imaging video is shown in Figure S1.

Finally, we apply a slight variation of the often-used $\Delta f/f$ transformation (Ahrens et al., 2013; Grewe et al., 2010; Grienberger and Konnerth, 2012). For the ith pixel in the jth frame, the standardized fluorescence is equal to

$$y_{i,j} \equiv \frac{y^0_{i,j} - \text{median}_{t=1,\ldots,T}(y^0_{i,t})}{\text{median}_{t=1,\ldots,T}(y^0_{i,t}) + \text{quantile}_{10\%}(\boldsymbol{Y}^0)},$$

where $y^0_{i,j}$ is the fluorescence (after smoothing and bleaching correction) of the ith pixel in the jth frame. This differs from the typical $\Delta f/f$ transformation in that (i) we standardize using the median across image frames instead of the mean; and (ii) we add a small number to the denominator. This adjustment in the denominator prevents small fluctuations in the amount of fluorescence at pixels with very little overall fluorescence from resulting in extremely high standardized fluorescences. In Figure S2, we show the resulting images after each stage of pre-processing for a sample frame.

10. Rationale for the Value of ω Used. In Section 4.2.1 of the main text, we suggested using a default value of $\omega = 0.2$ in Step 2 of SCALPEL. Here, we provide a justification for that choice of default value. To do this, we will derive the spatial and temporal dissimilarities for a pair of neurons. First, we make a couple of simplifying assumptions. We consider two neurons of the same size that share a fraction k of their pixels. We assume that the neurons fire in distinct frames and

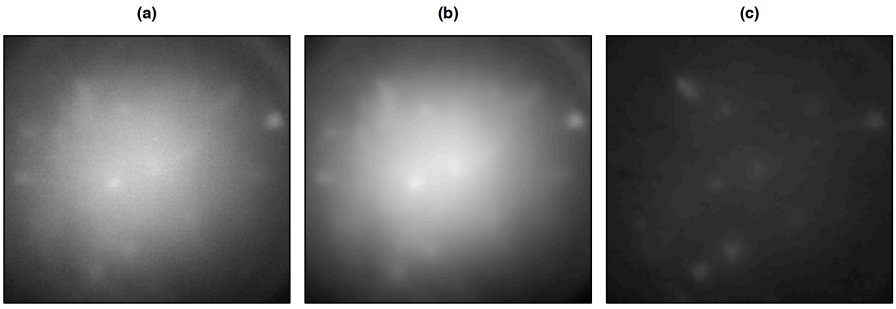

FIG S2. *In (a), we show a sample frame from the raw example calcium imaging video. In (b), we show the same frame after spatial and temporal smoothing has been done and the bleaching effect has been removed. In (c), we show the frame after the $\Delta f/f$ transformation has been performed, which is the final result of the pre-processing in Step 0 of SCALPEL.*

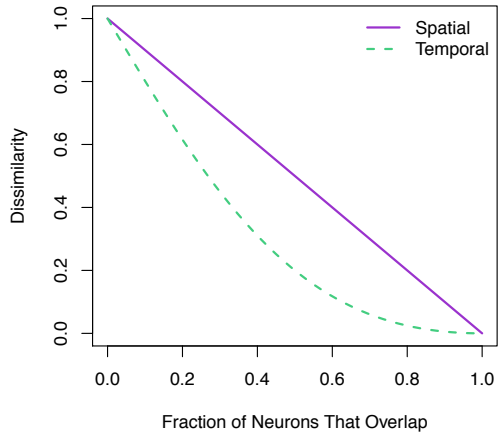

FIG S3. *We compare the spatial and temporal dissimilarities, which are derived in Section 10, for a pair of neurons sharing a certain percent of pixels.*

have the same overall activity (i.e., the ℓ_2 norm of their calcium concentration over time is equal). Under these assumptions, the spatial dissimilarity will be $1-k$ and the temporal dissimilarity will be $1-\frac{2k}{1+k^2}$. These results follow directly from the definitions given in equations (3) and (4). In Figure S3, we plot the spatial and temporal dissimilarities over the range of possible k values. We see that the spatial dissimilarity will always be larger than the temporal dissimilarity. Therefore, to put the spatial and temporal dissimilarities on equal footing, we should use $\omega < 0.5$. While the exact value of ω needed to balance the temporal and spatial information equally depends on the amount of overlap, we chose $\omega = 0.2$ as this corresponds to an intermediate amount of overlap (65%). Indeed, the exact choice of ω is not incredibly important: we show that values of ω between 0.1 and 0.4 perform similarly well for simulated data (Section 7.3) and real data (Section 14).

11. Example of a Cluster in Step 2. To see how the preliminary dictionary element in a cluster relates to the other preliminary dictionary elements assigned to that cluster, we give an example in Figure S4.

12. Filtering Dictionary Elements Prior to Fitting the Sparse Group Lasso. Optionally, at the beginning of Step 3, the elements in the refined dictionary from Step 2 may be filtered

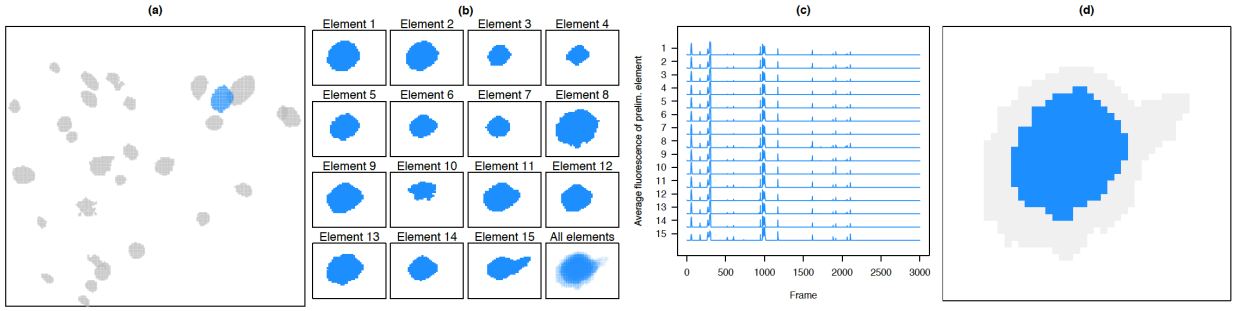

FIG S4. *We focus on a single cluster of preliminary dictionary elements from Step 2. The representative dictionary element for this cluster is highlighted in (a). The spatial maps for a random subset of the 36 preliminary dictionary elements in this cluster are shown in (b), along with a spatial map containing all 36 dictionary elements, in which the hue intensity at a given pixel indicates the number of elements containing that pixel. In (c), we plot the average thresholded fluorescence for each of the dictionary elements from (b). Finally, in (d), we show the representative element for that cluster. The gray coloring indicates the union of all preliminary dictionary elements in the cluster.*

prior to fitting the sparse group lasso in (6). One way to filter the dictionary elements is on the basis of the number of members in the clusters. That is, we can discard any dictionary elements representing clusters containing fewer than some minimum number of members. This number should be chosen based on the goals of the analysis. If we retain all clusters, regardless of size, then we may include some non-neuronal dictionary elements in the sparse group lasso problem. In contrast, if we discard small clusters (e.g., those with fewer than five members), then we may erroneously filter out some true neurons. In Figure S5, we illustrate the sensitivity of the results to the choice of minimum cluster size, on the example video considered in Sections 4 and 6.2.

13. Further Discussion of Step 3. We now elaborate on issues related to solving the sparse group lasso problem (6) in Step 3. This discussion is somewhat technical, and can be skipped by readers interested in only the practical use of SCALPEL. We discuss the solution to (6) when $K_f = 1$ in Section 13.1, our algorithm for solving (6) for any value of K_f in Section 13.2, the justification for the scaling of $\boldsymbol{A}^f$ in Section 13.3, a result about the tuning parameters α and λ that lead to a sparse solution in Section 13.4, and the ability of the group lasso penalty in (6) to zero out unwanted dictionary elements in Section 13.5.

13.1. *Single Component Problem.* We first consider solving (6) in the setting with a single spatial component ($K_f = 1$). While calcium imaging data will not have only a single neuron, this setting provides intuition, and will prove useful when we later solve (6) for $K_f > 1$ in Section 13.2.

LEMMA 13.1. *The solution to*

$$\underset{\boldsymbol{z}\in\mathbb{R}^T,\boldsymbol{z}\geq\boldsymbol{0}}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{Y}-\tilde{\boldsymbol{a}}^f\boldsymbol{z}^\top\right\|_F^2+\lambda\alpha\left\|\boldsymbol{z}\right\|_1+\lambda(1-\alpha)\left\|\boldsymbol{z}\right\|_2 \tag{S1}$$

is

$$\hat{\boldsymbol{z}}=\left(1-\frac{\lambda(1-\alpha)}{\left\|(\boldsymbol{Y}^\top\tilde{\boldsymbol{a}}^f-\lambda\alpha\boldsymbol{1})_+\right\|_2}\right)_+\left(\frac{\boldsymbol{Y}^\top\tilde{\boldsymbol{a}}^f-\lambda\alpha\boldsymbol{1}}{(\tilde{\boldsymbol{a}}^f)^\top\tilde{\boldsymbol{a}}^f}\right)_+, \tag{S2}$$

where $(a)_+ = \max(0, a)$ *is applied element-wise.*

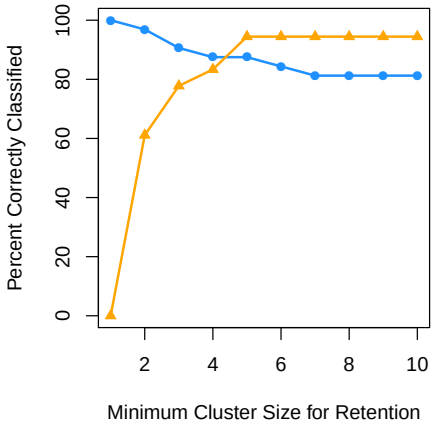

FIG S5. *On the example video considered in Sections 4 and 6.2, we manually inspected each dictionary element resulting from Step 2 of SCALPEL, by examining the frames from which each dictionary element was derived. Based on this manual inspection, we classified each of the 50 dictionary elements as a "neuron" or a "non-neuron". Next, we considered whether simply filtering each dictionary element based on the number of elements in its cluster (as described at the beginning of Step 3 of SCALPEL) would accurately distinguish between "neurons" and "non-neurons". In the figure, the y-axis shows the percentage of "neurons" that would remain after filtering (●), and the percentage of "non-neurons" that would be eliminated via filtering (▲), as a function of the filtering threshold (shown on the x-axis). We find that in this video, a careful manual analysis of each dictionary element yields very similar results to simply filtering each dictionary element based on the number of elements in its cluster.*

The proof of Lemma 13.1 is in Section 15.1. We can inspect the solution (S2) to gain intuition. Recall that $\tilde{\boldsymbol{a}}^f_{\cdot,k} \equiv \boldsymbol{a}^f_{\cdot,k}/\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$, where $\boldsymbol{a}^f_{\cdot,k}$ has binary elements. Therefore, $\boldsymbol{Y}^\top\tilde{\boldsymbol{a}}^f \in \mathbb{R}^T$ is the average fluorescence of pixels in the filtered dictionary element at each of the frames and $\frac{1}{(\tilde{\boldsymbol{a}}^f)^\top\tilde{\boldsymbol{a}}^f}$ equals the number of pixels in the dictionary element. When $\lambda = 0$, $\hat{\boldsymbol{z}} = \left(\frac{\boldsymbol{Y}^\top\tilde{\boldsymbol{a}}^f}{(\tilde{\boldsymbol{a}}^f)^\top\tilde{\boldsymbol{a}}^f}\right)_+$, which is simply the positive part of the total fluorescence at all pixels in the dictionary element over time. We now consider the impact of λ for three different settings of α:

- $\alpha = 1$*:* In this setting, $\hat{\boldsymbol{z}}$ is the positive part of the soft-thresholded total fluorescence. This soft-thresholding encourages elements of $\hat{\boldsymbol{z}}$ to be exactly zero for frames in which the dictionary element has low fluorescence.
- $\alpha = 0$*:* In this setting, $\hat{\boldsymbol{z}}$ is found by scaling all elements of the total fluorescence by the same amount. Thus, individual elements of $\hat{\boldsymbol{z}}$ are not encouraged to be 0, though $\hat{\boldsymbol{z}} = \boldsymbol{0}$ if the dictionary element has a low amount of fluorescence across all frames (i.e., $\|\boldsymbol{Y}^\top\tilde{\boldsymbol{a}}^f\|_2$ small) or λ is very large.
- $\alpha \in (0, 1)$*:* Both soft-thresholding and soft-scaling are performed, which encourages sparsity of individual elements of $\hat{\boldsymbol{z}}$ and the entire vector $\hat{\boldsymbol{z}}$, respectively.

In Figure S6, we illustrate the values of $\hat{\boldsymbol{z}}$ for the three scenarios described above.

13.2. *Algorithm.* We now consider how to solve (6) for $K_f > 1$. While generalized gradient descent (Beck and Teboulle, 2009) can be used to solve concurrently for $\boldsymbol{z}_{1,\cdot}, \ldots, \boldsymbol{z}_{K_f,\cdot}$ in (6), the problem is solved more efficiently by noting that (6) is decomposable into groups of overlapping spatial components.

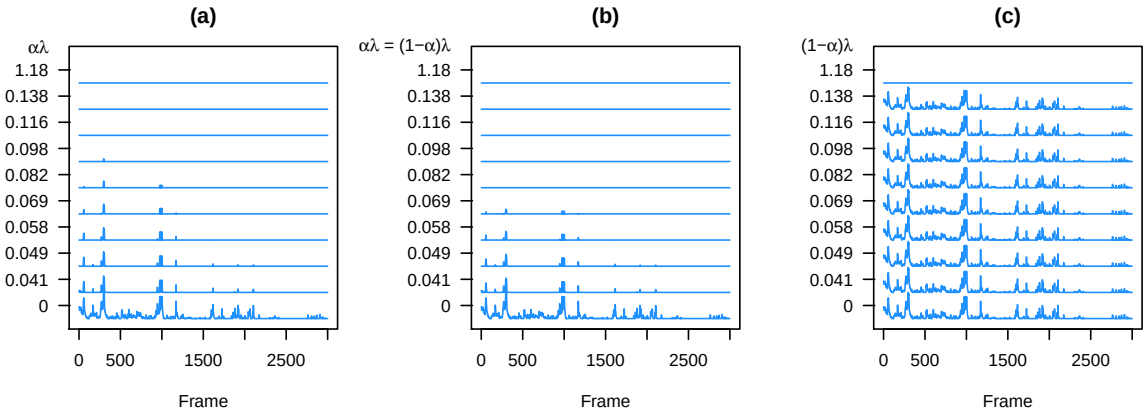

FIG S6. *For a single dictionary element in the example video, we plot the solution* $\hat{\boldsymbol{z}}$*, as given in* (S2)*, for a range of* λ *when (a) only a lasso penalty is used* ($\alpha = 1$)*, (b) a mixture of penalties is used* ($\alpha = 0.5$)*, and (c) only a group lasso penalty is used* ($\alpha = 0$)*.*

Let $\mathcal{N}_1, \ldots, \mathcal{N}_S$ denote a partition of the K_f elements of the filtered dictionary, such that $\mathcal{N}_s \cap \mathcal{N}_{s'} = \emptyset$ for $s \neq s'$, and $\cup_{s=1}^{S} \mathcal{N}_s = \{1, \ldots, K_f\}$. Define the mapping

$$\mathcal{M}(\mathcal{N}_s) = \{p \in (1, \ldots, P) : (\tilde{\boldsymbol{A}}^f{}_{p,\mathcal{N}_s})^\top \mathbf{1} > 0\}. \tag{S3}$$

That is, $\mathcal{M}(\mathcal{N}_s)$ indexes the set of pixels that are active in that subset of neurons.

LEMMA 13.2. *Suppose that* $\mathcal{M}(\mathcal{N}_s) \cap \mathcal{M}(\mathcal{N}_{s'}) = \emptyset$ *for all* $s \neq s'$*, so that there is no spatial overlap between the sets of filtered dictionary elements* $\mathcal{N}_1, \ldots, \mathcal{N}_S$*. Then solving* (6) *gives the same solution as solving*

$$\underset{\boldsymbol{Z}_{\mathcal{N}_s,\cdot} \in \mathbb{R}_+^{|\mathcal{N}_s| \times T}}{\text{minimize}} \quad \frac{1}{2} \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \boldsymbol{Z}_{\mathcal{N}_s,\cdot} \right\|_F^2 + \lambda\alpha \sum_{n \in \mathcal{N}_s} \|\boldsymbol{z}_{n,\cdot}\|_1 + \lambda(1-\alpha) \sum_{n \in \mathcal{N}_s} \|\boldsymbol{z}_{n,\cdot}\|_2 , \tag{S4}$$

for $s = 1, 2, \ldots, S$*.*

The proof of Lemma 13.2 is in Section 15.2.

Our approach to solving (S4) depends on the size of $\mathcal{N}_s$. For $|\mathcal{N}_s| = 1$, we can simply use the closed-form solution for $\boldsymbol{z}_{\mathcal{N}_s,\cdot}$ given by Lemma 13.1. This is advantageous as the calcium imaging data sets that we have analyzed often have some dictionary elements that do not overlap with any others. For $|\mathcal{N}_s| > 1$, we use generalized gradient descent to solve for the global optimum of (S4) (Beck and Teboulle, 2009).

In light of Lemma 13.2, in order to solve (6), we first partition the filtered dictionary elements into S sets, $\mathcal{N}_1, \ldots, \mathcal{N}_S$, such that there is no overlap between the pixels in the S sets, and so that no set can be partitioned further. This can be done quickly, as outlined in Step 1 of Algorithm 1. Then, we solve (S4) for $s = 1, \ldots, S$. Details are provided in Algorithm 1.

We typically solve (6) for a sequence of exponentially decreasing λ values. To improve computational performance, Step 2(b) of Algorithm 1 can be implemented using warm starts, in which $\boldsymbol{Z}^{(0)}_{\mathcal{N}_s,\cdot}$ is initialized as the solution for $\boldsymbol{Z}_{\mathcal{N}_s,\cdot}$ for the previous value of λ. Additional details regarding the derivation of the generalized gradient descent algorithm used in Step 2(b) of Algorithm 1 are given in Section 15.3.

Algorithm 1 — Algorithm for Solving Equation (6)

1. Construct the adjacency matrix $\boldsymbol{N} \in \mathbb{R}^{K_f \times K_f}$ with $n_{i,j} = \begin{cases} 1 & \text{if } (\boldsymbol{a}^f_{\cdot,i})^\top \boldsymbol{a}^f_{\cdot,j} > 0 \\ 0 & \text{if } (\boldsymbol{a}^f_{\cdot,i})^\top \boldsymbol{a}^f_{\cdot,j} = 0 \end{cases}$. Let $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_S$ denote the connected components of the graph corresponding to $\boldsymbol{N}$. That is, $\mathcal{N}_s$ indexes the filtered dictionary elements in the sth connected component. Define the mapping $\mathcal{M}(\mathcal{N}_s) = \{p \in (1, \ldots, P) : (\tilde{\boldsymbol{A}}^f_{p,\mathcal{N}_s})^\top \mathbf{1} > 0\}$.

2. For $s = 1, 2, \ldots, S$, solve

$$\text{(S5)} \qquad \underset{\boldsymbol{Z}_{\mathcal{N}_s,\cdot} \geq 0}{\text{minimize}} \quad \frac{1}{2} \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \boldsymbol{Z}_{\mathcal{N}_s,\cdot} \right\|_F^2 + \lambda\alpha \sum_{n \in \mathcal{N}_s} \|\boldsymbol{z}_{n,\cdot}\|_1 + \lambda(1-\alpha) \sum_{n \in \mathcal{N}_s} \|\boldsymbol{z}_{n,\cdot}\|_2 ,$$

   using one of the two following approaches:

   (a) By Lemma 13.1, if $|\mathcal{N}_s| = 1$, the closed-form solution for $\boldsymbol{z}_{\mathcal{N}_s,\cdot}$ in (S5) is

$$\text{(S6)} \qquad \hat{\boldsymbol{z}}_{\mathcal{N}_s,\cdot} = \left(1 - \frac{\lambda(1-\alpha)}{\left\| \left((\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot})^\top \tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} - \lambda\alpha \mathbf{1} \right)_+ \right\|_2} \right)_+ \left(\frac{(\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot})^\top \tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} - \lambda\alpha \mathbf{1}}{(\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top \tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}} \right)_+ .$$

   (b) If $|\mathcal{N}_s| > 1$, use generalized gradient descent to solve (S5) for $\boldsymbol{Z}_{\mathcal{N}_s,\cdot}$:

      i. Let $f(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}) = \frac{1}{2} \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \boldsymbol{Z}_{\mathcal{N}_s,\cdot} \right\|_F^2 + \lambda\alpha \mathbf{1}^\top \boldsymbol{Z}_{\mathcal{N}_s,\cdot} \mathbf{1}$. Initialize $\boldsymbol{Z}^{(0)}_{\mathcal{N}_s,\cdot} := \mathbf{0}$ and let $t := (\max_{n \in \mathcal{N}_s} \sum_{j \in \mathcal{N}_s} (\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),j})^\top \tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),n})^{-1}$.

      ii. For $b = 1, 2, \ldots$, until convergence, iterate:

$$\begin{aligned} &\nabla f\left(\boldsymbol{Z}^{(b-1)}_{\mathcal{N}_s,\cdot}\right) := -(\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top \left(\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \boldsymbol{Z}^{(b-1)}_{\mathcal{N}_s,\cdot} \right) + \lambda\alpha \mathbf{1}\mathbf{1}^\top, \\ &\tilde{\boldsymbol{Y}}_{\mathcal{N}_s,\cdot} := \boldsymbol{Z}^{(b-1)}_{\mathcal{N}_s,\cdot} - t \nabla f\left(\boldsymbol{Z}^{(b-1)}_{\mathcal{N}_s,\cdot}\right), \text{ and} \\ &\boldsymbol{z}^{(b)}_{n,\cdot} := \left(1 - \frac{\lambda(1-\alpha)t}{\left\| (\tilde{\boldsymbol{y}}_{n,\cdot})_+ \right\|_2} \right)_+ (\tilde{\boldsymbol{y}}_{n,\cdot})_+ \text{ for } n \in \mathcal{N}_s. \end{aligned}$$

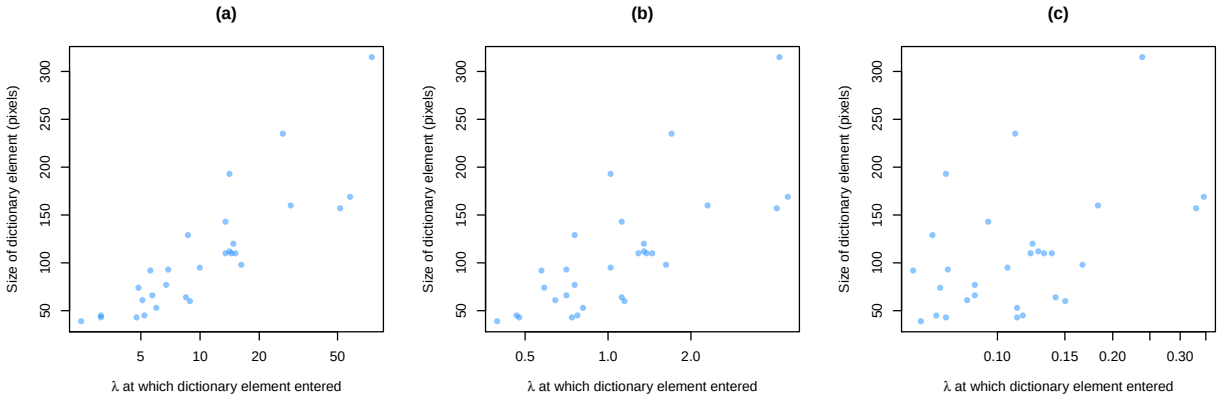

FIG S7. *We solve* (6) *for different scalings of* $\tilde{\boldsymbol{A}}^f$. *For each spatial component, we note the value of* λ *at which the spatial component enters the model (i.e., the largest* λ *for which* $\hat{\boldsymbol{z}}_{k,\cdot} \neq \boldsymbol{0}$*). We plot the size of each spatial component versus the value of* λ *at which the spatial component enters for (a)* $\tilde{\boldsymbol{a}}^f_{\cdot,k} = \boldsymbol{a}^f_{\cdot,k}$*, (b)* $\tilde{\boldsymbol{a}}^f_{\cdot,k} = \boldsymbol{a}^f_{\cdot,k}/\|\boldsymbol{a}^f_{\cdot,k}\|_2$*, and (c)* $\tilde{\boldsymbol{a}}^f_{\cdot,k} = \boldsymbol{a}^f_{\cdot,k}/\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$*. There is a high correlation in the scatterplots in panels (a) and (b), but little correlation in (c). This lack of correlation motivates us to use the scaling* $\tilde{\boldsymbol{a}}^f_{\cdot,k} = \boldsymbol{a}^f_{\cdot,k}/\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$ *in Step 3, so that dictionary elements receive a fair shot of selection by the sparse group lasso* (6)*, regardless of their size.*

13.3. *Scaling of* $\boldsymbol{A}^f$. In (6), the kth column of the matrix $\tilde{\boldsymbol{A}}^f$ encodes the spatial mapping of the kth filtered dictionary element, after scaling. To obtain $\tilde{\boldsymbol{A}}^f$, we divide the kth column of $\boldsymbol{A}^f$ by $\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$, the number of pixels in the kth filtered dictionary element. This scaling is performed so that the sizes of the dictionary elements do not impact when the components enter the model. That is, we would like $\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$ to be independent of the largest value of λ for which $\hat{\boldsymbol{z}}_{k,\cdot} \neq \boldsymbol{0}$. The following lemma supports this particular scaling of the columns of $\boldsymbol{A}^f$.

LEMMA 13.3. *Suppose* $\boldsymbol{Y} = \boldsymbol{A}^f\boldsymbol{Z}^*$ *where the following conditions hold:*

(i) $\boldsymbol{A}^f \in \mathbb{R}^{P\times K_f}$ *with* $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,2} = 0$, $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ *for* $k = 3,\ldots,K_f$, *and* $(\boldsymbol{a}^f_{\cdot,2})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ *for* $k = 3,\ldots,K_f$ *and*

(ii) $\boldsymbol{Z}^* \in \mathbb{R}^{K_f\times T}$ *with* $\boldsymbol{z}^*_{1,\cdot} = \boldsymbol{P}\boldsymbol{z}^*_{2,\cdot}$ *for some* $T \times T$ *permutation matrix* $\boldsymbol{P}$.

If we solve (6) *for* $\boldsymbol{Z}$ *with* $\tilde{\boldsymbol{A}}^f$ *such that* $\tilde{\boldsymbol{a}}^f_{\cdot,k} = \boldsymbol{a}^f_{\cdot,k}/\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$, *then* $\hat{\boldsymbol{z}}_{1,\cdot} = \boldsymbol{0}$ *if and only if* $\hat{\boldsymbol{z}}_{2,\cdot} = \boldsymbol{0}$.

The proof of Lemma 13.3 is in Section 15.4. Lemma 13.3 indicates that two non-overlapping spatial components, possibly of different sizes, whose temporal components are identical up to a permutation, will enter the model at the same value of λ. In Figure S7, we provide empirical evidence for the chosen scaling of $\boldsymbol{A}^f$.

13.4. *Sparsity of the Solution.* We now consider the range of λ for which the solution to (6) is completely sparse (i.e., $\hat{\boldsymbol{Z}} = \boldsymbol{0}$) for a fixed value of α.

LEMMA 13.4. *For any* $\alpha \in [0,1]$, *the solution to* (6) *is completely sparse if and only if*

$$\lambda(1-\alpha) \geq \left\| \left(\left[(\tilde{\boldsymbol{A}}^f)^\top \boldsymbol{Y} \right]_{k,\cdot} - \lambda\alpha\boldsymbol{1} \right)_+ \right\|_2 \tag{S7}$$

for $k = 1,\ldots,K_f$.

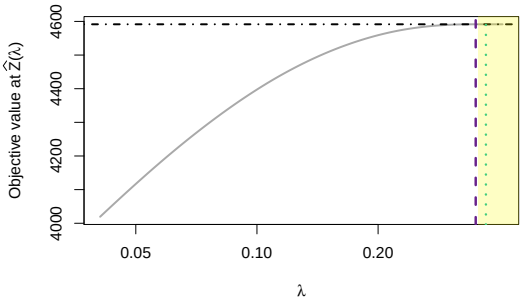

FIG S8. *We plot the value of the objective of* (6) *at* $\hat{\boldsymbol{Z}}(\lambda)$*, the minimizer of* (6) *at* λ*, for a replicate of data as* λ *varies. We compare two ways of finding a* λ *large enough such that* $\hat{\boldsymbol{Z}}(\lambda) = \mathbf{0}$*, which results in the objective shown as* **- · -**. *We take* λ *that satisfies Lemma 13.4 (- -) or* λ *as defined in Corollary 13.5 (· · ·). The former (- -) gives the smallest* λ *such that* $\hat{\boldsymbol{Z}}(\lambda) = \mathbf{0}$*. We can see this from the fact that the line (- -) is on the boundary of the shaded box, which indicates the range of* λ *for which the objective value at* $\hat{\boldsymbol{Z}}(\lambda)$ *equals the objective value at* $\mathbf{0}$*, i.e., the range of* λ *for which* $\hat{\boldsymbol{Z}}(\lambda) = \mathbf{0}$.

Unfortunately, when $\alpha \in (0,1)$, λ is on both sides of the inequality in (S7). Though we can solve for λ in (S7) using a root finder when $\alpha \in (0,1)$, the following corollary provides a simple alternative.

COROLLARY 13.5. *For any* $\alpha \in (0,1)$*, if*

$$\lambda \geq \max_{k=1,\ldots,K_f} \left[\min \left(\max_{l=1,\ldots,T} \frac{\left(\left[(\tilde{\boldsymbol{A}}^f)^\top \boldsymbol{Y} \right]_{k,l} \right)_+}{\alpha}, \frac{\left\| \left(\left[(\tilde{\boldsymbol{A}}^f)^\top \boldsymbol{Y} \right]_{k,\cdot} \right)_+ \right\|_2}{1-\alpha} \right) \right],$$

then the solution to (6) *is completely sparse.*

The condition in Corollary 13.5 is sufficient, but not necessary. Proofs of Lemma 13.4 and Corollary 13.5 can be found in Sections 15.5 and 15.6, respectively. An illustration of Lemma 13.4 and Corollary 13.5 is provided in Figure S8.

13.5. *Zeroing Out of Double Neurons.* Some of the elements in the preliminary dictionary obtained in Step 1 may be *double neurons*, i.e., elements that are a combination of two separate neurons. This occurs when two neighboring neurons are active during the same frame. In Step 2 of SCALPEL, these double neurons are unlikely to cluster with elements representing either of the individual neurons they combine, and thus there may be double neurons that remain in the filtered set of dictionary elements, $\boldsymbol{A}^f$, used in Step 3 of SCALPEL. Fortunately, as detailed in the following lemma, the group lasso penalty in (6) filters out these double neurons by estimating their temporal components to be the zero vector.

LEMMA 13.6. *Suppose that the following conditions hold on* $\boldsymbol{A}^f \in \mathbb{R}^{P \times K_f}$*:*

(i) $\boldsymbol{a}^f_{\cdot,3} = \boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2}$,
(ii) $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,2} = 0$,

(iii) $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ *for* $k = 4, \ldots, K_f$, *and*
(iv) $(\boldsymbol{a}^f_{\cdot,2})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ *for* $k = 4, \ldots, K_f$.

Then, define $\tilde{\boldsymbol{a}}^f_{\cdot,k} \equiv \boldsymbol{a}^f_{\cdot,k} / \|\boldsymbol{a}^f_{\cdot,k}\|_2$, *and consider solving* (6) *for* $\boldsymbol{Z}$ *with* $\alpha < 1$. *Then,* $\hat{\boldsymbol{z}}_{3,\cdot} = \boldsymbol{0}$.

The proof of Lemma 13.6 is in Section 15.7. Note that Lemma 13.6 assumes that the individual elements for the neighboring neurons, $\boldsymbol{a}^f_{\cdot,1}$ and $\boldsymbol{a}^f_{\cdot,2}$, do not overlap at all. The group lasso penalty can also be effective at zeroing out double neurons resulting from overlapping neurons, though this depends on the amount of overlap, among other factors.

13.6. *Selecting* λ *in* (6) *in Step 3.* To choose λ for (6) via a validation set approach, we perform the following steps:

1. Obtain $\tilde{\boldsymbol{A}}^f \in \mathbb{R}^{P \times K_f}$ by dividing the kth column of $\boldsymbol{A}^f$ by $\|\boldsymbol{a}^f_{\cdot,k}\|_2^2$, which ensures that the sizes of the dictionary elements do not impact when the components enter the model.
2. Construct a training set $\mathcal{T}$ by sampling 60% of the pixels in each overlapping group of neurons. That is, we sample 60% of the elements in $\mathcal{M}(\mathcal{N}_1), \mathcal{M}(\mathcal{N}_2), \ldots, \mathcal{M}(\mathcal{N}_S)$, which were defined in (S3). Assign the remaining pixels to the validation set, $\mathcal{V} = \{v \in (1, \ldots, P) : v \notin \mathcal{T}\}$.
3. Using Algorithm 1, solve (6) on the training set of pixels for a decreasing sequence of 20 λ values, $\lambda_1, \ldots, \lambda_{20}$:

$$\hat{\boldsymbol{Z}}(\lambda_i) = \underset{\boldsymbol{Z} \in \mathbb{R}^{K_f \times T}, \boldsymbol{Z} \geq 0}{\operatorname{argmin}} \quad \frac{1}{2} \left\| \boldsymbol{Y}_{\mathcal{T},\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{T},\cdot} \boldsymbol{Z} \right\|_F^2 + \lambda_i \alpha \sum_{k=1}^{K_f} \|\boldsymbol{z}_{k,\cdot}\|_1 + \lambda_i (1-\alpha) \sum_{k=1}^{K_f} \|\boldsymbol{z}_{k,\cdot}\|_2 .$$

4. For each λ_i, calculate the validation error,

$$err_{\mathcal{V}}(\lambda_i) = \frac{1}{|\mathcal{V}|} \left\| \boldsymbol{Y}^B_{\mathcal{V},\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{V},\cdot} \hat{\boldsymbol{Z}}(\lambda_i) \right\|_F^2 ,$$

where $\left[\boldsymbol{Y}^B_{\mathcal{V},\cdot}\right]_{j,k} = \begin{cases} [\boldsymbol{Y}_{\mathcal{V},\cdot}]_{j,k} & \text{if } [\boldsymbol{Y}_{\mathcal{V},\cdot}]_{j,k} > -\text{quantile}_{0.1\%}(\boldsymbol{Y}) \\ 0 & \text{otherwise} \end{cases}$. We use a thresholded version of $\boldsymbol{Y}$ when calculating the validation error, as we only care about the reconstruction error on the brightest parts of the video. Select the optimal value of λ as

$$\lambda^* = \underset{\lambda_i}{\operatorname{argmax}} \left\{ \lambda_i : \frac{err_{\mathcal{V}}(\lambda_i) - \min_{\lambda_j} err_{\mathcal{V}}(\lambda_j)}{\min_{\lambda_j} err_{\mathcal{V}}(\lambda_j)} \leq 0.05 \right\};$$

this is the largest value of λ that results in a validation error within 5% of the minimum validation error achieved by any value of λ considered.
5. Solve (6) on all pixels:

$$\underset{\boldsymbol{Z} \in \mathbb{R}^{K_f \times T}, \boldsymbol{Z} \geq 0}{\text{minimize}} \quad \frac{1}{2} \left\| \boldsymbol{Y} - \tilde{\boldsymbol{A}}^f \boldsymbol{Z} \right\|_F^2 + \frac{\lambda^*}{|\mathcal{T}|/P} \alpha \sum_{k=1}^{K_f} \|\boldsymbol{z}_{k,\cdot}\|_1 + \frac{\lambda^*}{|\mathcal{T}|/P} (1-\alpha) \sum_{k=1}^{K_f} \|\boldsymbol{z}_{k,\cdot}\|_2 ,$$

where we have scaled the tuning parameter by the percent of pixels in the training set, to account for the fact that the sum of squared errors in the loss function is not scaled by the number of pixels.

This process can be done separately for each group of overlapping neurons $\mathcal{N}_1, \ldots, \mathcal{N}_S$ to select a different value of λ for each group, or for all groups at once to select a single value of λ. By following steps similar to those described above, λ can alternatively be selected via cross-validation.

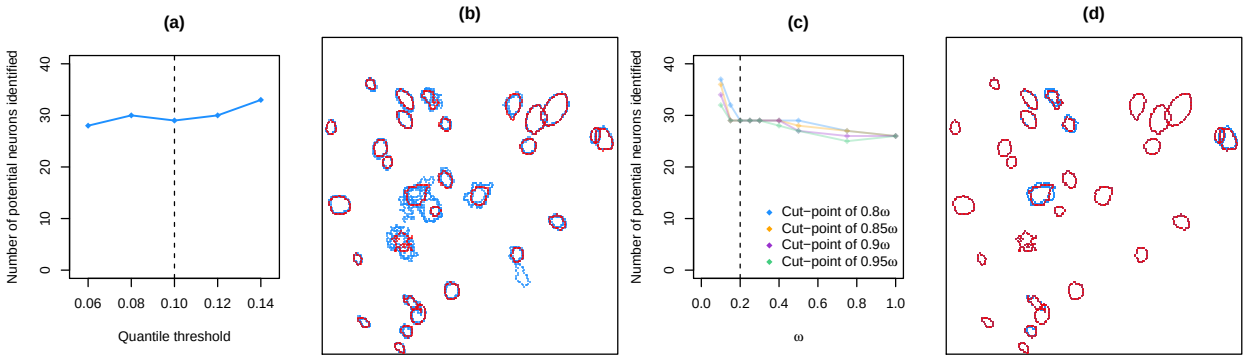

FIG S9. *We present the results of analyzing the one-photon data from Section 6.2 using non-default values of the tuning parameters. In (a), we plot the number of potential neurons identified as a function of the quantile threshold in Step 1. The dashed line indicates the default value for the quantile threshold. In (b), we plot the outlines of the neurons identified using the default value (—) and the non-default values (—) for the quantile threshold. In (c), we plot the number of potential neurons identified as a function of the dendrogram cut-point and ω in Step 2. The dashed line indicates the default value for ω. In (d), we plot the outlines of the neurons identified using the default values (—) and the non-default values (—) for the dendrogram cut-point and ω values that produced the same number of neurons as the default values.*

14. Sensitivity of Results to Changes in the Tuning Parameters. In analyzing the one-photon data in Section 6.2, we utilized default values for all of the tuning parameters. We now consider what impact varying these default values has on the results of our analysis. In particular, we consider the effect of modifying the quantile threshold in Step 1, the dissimilarity weight ω in Step 2, and the dendrogram cut-point in Step 2. In Figure S9(a), we see that varying the quantile threshold used for producing the preliminary dictionary in Step 1 results in a small variation in the final number of neurons identified, producing between 28 and 33 neurons, compared to the 29 neurons identified using the default value. Additionally, the shapes of the neurons identified using different quantile thresholds are quite similar (Figure S9(b)). In Figure S9(c), we see that a large range of values of ω and the dendrogram cut-point produce the exact same number of neurons as the default values of these parameters. Indeed, in Figure S9(d), we see that there is very little change in the neurons identified. These results illustrate the performance of SCALPEL does not diminish with modest variations in the values of the tuning parameters.

Similar analyses for the simulated calcium imaging data are provided in Section 7.3.

15. Technical Proofs Related to Section 13.

15.1. *Proof of Lemma 13.1.* We first prove a result that we will use later.

LEMMA 15.1. *The following holds:* $\underset{\boldsymbol{\beta}\geq\mathbf{0}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2 = \left(1-\frac{\lambda}{\|(\boldsymbol{y})_+\|_2}\right)_+(\boldsymbol{y})_+.$

PROOF. Let $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\geq\mathbf{0}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2$ and $\mathcal{C}=\{i: y_i\geq 0\}$. First, we show $\hat{\boldsymbol{\beta}}_{-\mathcal{C}}=\mathbf{0}$. In anticipation of contradiction, assume there exists j such that $j\notin\mathcal{C}$ and $\hat{\beta}_j>0$. Define $\tilde{\boldsymbol{\beta}}$ as $\tilde{\beta}_i=\begin{cases}\hat{\beta}_i & \text{if } i\neq j\\ 0 & \text{if } i=j\end{cases}$. Then

$$\frac{1}{2}\left\|\boldsymbol{y}-\tilde{\boldsymbol{\beta}}\right\|_2^2+\lambda\left\|\tilde{\boldsymbol{\beta}}\right\|_2<\frac{1}{2}\left\|\boldsymbol{y}-\hat{\boldsymbol{\beta}}\right\|_2^2+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_2.$$

This is a contradiction, so we conclude that $\hat{\beta}_i = 0$ for all $i \notin \mathcal{C}$. It remains to solve

$$\underset{\boldsymbol{\beta}_{\mathcal{C}} \geq \mathbf{0}}{\text{minimize}} \quad \frac{1}{2} \left\| \boldsymbol{y}_{\mathcal{C}} - \boldsymbol{\beta}_{\mathcal{C}} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta}_{\mathcal{C}} \right\|_2 . \tag{S8}$$

By a result in Section 3.1 of Simon et al. (2013), the solution to (S8) without the non-negativity constraint on $\boldsymbol{\beta}_{\mathcal{C}}$ is $\left(1 - \frac{\lambda}{\|\boldsymbol{y}_{\mathcal{C}}\|_2}\right)_+ \boldsymbol{y}_{\mathcal{C}}$, which has all non-negative elements. Therefore, it is also the solution to (S8). □

We now proceed to prove Lemma 13.1.

Proof. Our goal is to solve

$$\underset{\boldsymbol{z} \in \mathbb{R}^T, \boldsymbol{z} \geq \mathbf{0}}{\text{minimize}} \quad \frac{1}{2} \left\| \boldsymbol{Y} - \tilde{\boldsymbol{a}}^f \boldsymbol{z}^\top \right\|_F^2 + \lambda\alpha \mathbf{1}^\top \boldsymbol{z} + \lambda(1-\alpha) \left\| \boldsymbol{z} \right\|_2 . \tag{S9}$$

Note that solving (S9) is equivalent to solving (S1), as $\|\boldsymbol{z}\|_1 = \mathbf{1}^\top \boldsymbol{z}$ when $\boldsymbol{z} \geq \mathbf{0}$. By algebraic manipulation, we can show that

$$\left\| \boldsymbol{Y} - \tilde{\boldsymbol{a}}^f \boldsymbol{z}^\top \right\|_F^2 + \lambda\alpha \mathbf{1}^\top \boldsymbol{z} = \left\| \frac{\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f - \lambda\alpha \mathbf{1}}{\sqrt{(\tilde{\boldsymbol{a}}^f)^\top \tilde{\boldsymbol{a}}^f}} - \sqrt{(\tilde{\boldsymbol{a}}^f)^\top \tilde{\boldsymbol{a}}^f} \boldsymbol{z} \right\|_2^2 + C,$$

where C is a constant that does not depend on $\boldsymbol{z}$. Therefore, the solution to (S9) is the same as the solution to

$$\underset{\boldsymbol{z} \geq \mathbf{0}}{\text{minimize}} \quad \frac{1}{2} \left\| \frac{\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f - \lambda\alpha \mathbf{1}}{(\tilde{\boldsymbol{a}}^f)^\top \tilde{\boldsymbol{a}}^f} - \boldsymbol{z} \right\|_2^2 + \frac{\lambda(1-\alpha)}{(\tilde{\boldsymbol{a}}^f)^\top \tilde{\boldsymbol{a}}^f} \left\| \boldsymbol{z} \right\|_2 . \tag{S10}$$

We solve (S10) by applying Lemma 15.1. □

15.2. *Proof of Lemma 13.2.*

Proof. Recall the definition $\mathcal{M}(\mathcal{N}_s)$ of in (S3). Then, the result follows simply from observing that

$$\begin{aligned}
\left\| \boldsymbol{Y} - \tilde{\boldsymbol{A}}^f \boldsymbol{Z} \right\|_F^2 &= \sum_{s=1}^S \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f{}_{\mathcal{M}(\mathcal{N}_s),\cdot} \boldsymbol{Z} \right\|_F^2 \\
&= \sum_{s=1}^S \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \sum_{s'=1}^S \tilde{\boldsymbol{A}}^f{}_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_{s'}} \boldsymbol{Z}_{\mathcal{N}_{s'},\cdot} \right\|_F^2 \\
&= \sum_{s=1}^S \left\| \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f{}_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \boldsymbol{Z}_{\mathcal{N}_s,\cdot} \right\|_F^2 .
\end{aligned}$$

The last equality follows from the condition of the lemma, which guarantees that $\tilde{\boldsymbol{A}}^f{}_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_{s'}} = 0$ for all $s \neq s'$. □

15.3. *Details of Step 2(b) of Algorithm 1.* Note that minimizing the objective in (S5) subject to $\boldsymbol{Z}_{\mathcal{N}_s,\cdot} \geq \boldsymbol{0}$ is equivalent to minimizing

$$\frac{1}{2}\left\|\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right\|_F^2 + \lambda\alpha\boldsymbol{1}^\top\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\boldsymbol{1} + \lambda(1-\alpha)\sum_{n\in\mathcal{N}_s}\|\boldsymbol{z}_{n,\cdot}\|_2 \tag{S11}$$

subject to $\boldsymbol{Z}_{\mathcal{N}_s,\cdot} \geq \boldsymbol{0}$, since $\boldsymbol{1}^\top\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\boldsymbol{1} = \sum_{n\in\mathcal{N}_s}\|\boldsymbol{z}_{n,\cdot}\|_1$ when $\boldsymbol{Z}_{\mathcal{N}_s,\cdot} \geq \boldsymbol{0}$.

Let $f\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right) = \frac{1}{2}\left\|\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right\|_F^2 + \lambda\alpha\boldsymbol{1}^\top\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\boldsymbol{1}$, which is the differentiable part of (S11), and let $g\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right) = \lambda(1-\alpha)\sum_{n\in\mathcal{N}_s}\|\boldsymbol{z}_{n,\cdot}\|_2$, the non-differentiable part.

Generalized gradient descent (Beck and Teboulle, 2009; Parikh and Boyd, 2014) is a majorization-minimization scheme. First, we find a quadratic approximation to $f\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right)$ centered at our previous estimate for $\boldsymbol{Z}_{\mathcal{N}_s,\cdot}$, $\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}$, that majorizes $f\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right)$. That is,

$$f\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right) \leq f\left(\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right) + \operatorname{Tr}\left[\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot} - \boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right)^\top\nabla f\left(\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right)\right] + \frac{1}{2t}\left\|\boldsymbol{Z}_{\mathcal{N}_s,\cdot} - \boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right\|_F^2,$$

where t is the step size such that $\nabla^2 f(\cdot) \preceq \frac{1}{t}\boldsymbol{I}$. After completing the square, we can see that minimizing the quadratic approximation to $f\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right)$ gives the same solution as solving

$$\underset{\boldsymbol{Z}_{\mathcal{N}_s,\cdot}}{\text{minimize}} \quad \frac{1}{2t}\left\|\boldsymbol{Z}_{\mathcal{N}_s,\cdot} - \left(\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot} - t\nabla f\left(\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right)\right)\right\|_F^2.$$

Thus we perform this minimization with $g\left(\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\right)$ added to the objective function, which gives the proximal problem

$$\underset{\boldsymbol{Z}_{\mathcal{N}_s,\cdot}\geq\boldsymbol{0}}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{Z}_{\mathcal{N}_s,\cdot} - \tilde{\boldsymbol{Y}}_{\mathcal{N}_s,\cdot}\right\|_F^2 + \lambda(1-\alpha)t\sum_{n\in\mathcal{N}_s}\|\boldsymbol{z}_{n,\cdot}\|_2, \tag{S12}$$

where $\tilde{\boldsymbol{Y}}_{\mathcal{N}_s,\cdot} = \boldsymbol{Z}^0_{\mathcal{N}_s,\cdot} - t\left(-(\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top\left(\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - \tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}\boldsymbol{Z}^0_{\mathcal{N}_s,\cdot}\right) + \lambda\alpha\boldsymbol{1}\boldsymbol{1}^\top\right)$. The minimization in (S12) is separable in $\boldsymbol{z}_{n,\cdot}$, so for $n \in \mathcal{N}_s$, we solve

$$\underset{\boldsymbol{z}_{n,\cdot}\geq\boldsymbol{0}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{z}_{n,\cdot} - \tilde{\boldsymbol{y}}_{n,\cdot}\|_2^2 + \lambda(1-\alpha)t\|\boldsymbol{z}_{n,\cdot}\|_2. \tag{S13}$$

By Lemma 15.1 in Section 15.1, the solution to (S13) is $\hat{\boldsymbol{z}}_{n,\cdot} = \left(1 - \frac{\lambda(1-\alpha)t}{\|(\tilde{\boldsymbol{y}}_{n,\cdot})_+\|_2}\right)_+(\tilde{\boldsymbol{y}}_{n,\cdot})_+$.

It only remains to derive a suitable step size t so that $\nabla^2 f(\cdot) = (\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} \preceq \frac{1}{t}\boldsymbol{I}$. A sufficient condition for $\frac{1}{t}\boldsymbol{I} - (\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}$ to be positive semi-definite is that $\frac{1}{t}\boldsymbol{I} - (\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s}$ be diagonally dominant. That is,

$$\frac{1}{t} - (\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),n})^\top\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),n} \geq \sum_{j\in\mathcal{N}_s, j\neq n}(\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),j})^\top\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),n}$$

for all $n \in \mathcal{N}_s$. Thus we choose $t = (\max_{n\in\mathcal{N}_s}\sum_{j\in\mathcal{N}_s}(\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),j})^\top\tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),n})^{-1}$.

15.4. *Proof of Lemma 13.3.*

PROOF. Since $\boldsymbol{a}^f_{\cdot,1}$ does not overlap any other spatial components (i.e., $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ for $k = 2, \ldots, K_f$), we know by the results in Lemmas 13.1 and 13.2 that

$$\hat{\boldsymbol{z}}_{1,\cdot} = \left(1 - \frac{\lambda(1-\alpha)}{\|(\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f_{\cdot,1} - \lambda\alpha\mathbf{1})_+\|_2}\right)_+ \left(\frac{\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f_{\cdot,1} - \lambda\alpha\mathbf{1}}{(\tilde{\boldsymbol{a}}^f_{\cdot,1})^\top \tilde{\boldsymbol{a}}^f_{\cdot,1}}\right)_+ .$$

Note that $\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f_{\cdot,1} = (\boldsymbol{Z}^*)^\top (\boldsymbol{A}^f)^\top \boldsymbol{a}^f_{\cdot,1} / \|\boldsymbol{a}^f_{\cdot,1}\|_2^2$, so $\boldsymbol{Y}^\top \tilde{\boldsymbol{a}}^f_{\cdot,1} = \boldsymbol{z}^*_{1,\cdot}$ since $(\boldsymbol{a}^f_{\cdot,1})^\top \boldsymbol{a}^f_{\cdot,k} = 0$ for $k = 2, \ldots, K_f$. Thus $\hat{\boldsymbol{z}}_{1,\cdot} = \mathbf{0}$ if and only if $\lambda(1-\alpha) \geq \|(\boldsymbol{z}^*_{1,\cdot} - \lambda\alpha\mathbf{1})_+\|_2$. Similarly, $\hat{\boldsymbol{z}}_{2,\cdot} = \mathbf{0}$ if and only if $\lambda(1-\alpha) \geq \|(\boldsymbol{z}^*_{2,\cdot} - \lambda\alpha\mathbf{1})_+\|_2$. Thus, it remains to show that $\|(\boldsymbol{z}^*_{1,\cdot} - \lambda\alpha\mathbf{1})_+\|_2 = \|(\boldsymbol{z}^*_{2,\cdot} - \lambda\alpha\mathbf{1})_+\|_2$. Using the properties of permutation matrices that $\mathbf{1} = \boldsymbol{P}\mathbf{1}$ and $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}$, we see

$$\begin{aligned}
\|(\boldsymbol{z}^*_{1,\cdot} - \lambda\alpha\mathbf{1})_+\|_2 &= \|(\boldsymbol{P}\boldsymbol{z}^*_{2,\cdot} - \lambda\alpha\boldsymbol{P}\mathbf{1})_+\|_2 \\
&= \|\boldsymbol{P}(\boldsymbol{z}^*_{2,\cdot} - \lambda\alpha\mathbf{1})_+\|_2 \\
&= \|(\boldsymbol{z}^*_{2,\cdot} - \lambda\alpha\mathbf{1})_+\|_2,
\end{aligned}$$

and therefore, $\hat{\boldsymbol{z}}_{1,\cdot} = \mathbf{0}$ if and only if $\hat{\boldsymbol{z}}_{2,\cdot} = \mathbf{0}$. □

15.5. *Proof of Lemma 13.4.*

PROOF. Recall that solving (6) gives the same solution as solving (S4). Thus we focus on deriving a condition on λ that guarantees that $\hat{\boldsymbol{Z}}_{\mathcal{N}_s,\cdot}$, the solution to (S4), equals zero for $s = 1, \ldots, S$. If $|\mathcal{N}_s| = 1$, we see from (S6) that $\hat{\boldsymbol{z}}_{\mathcal{N}_s,\cdot} = \mathbf{0}$ if and only if

$$\lambda(1-\alpha) \geq \left\| \left((\boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot})^\top \tilde{\boldsymbol{a}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s} - \lambda\alpha\mathbf{1} \right)_+ \right\|_2 . \tag{S14}$$

Recall that if $|\mathcal{N}_s| > 1$, we iteratively solve for $\hat{\boldsymbol{Z}}_{\mathcal{N}_s,\cdot}$ using Step 2(b) of Algorithm 1. We initialize at the sparse solution $\boldsymbol{Z}^{(0)}_{\mathcal{N}_s,\cdot} = \mathbf{0}$ and thus for $n \in \mathcal{N}_s$

$$\boldsymbol{z}^{(1)}_{n,\cdot} = \left(1 - \frac{\lambda(1-\alpha)t}{\left\|(\tilde{\boldsymbol{y}}_{n,\cdot})_+\right\|_2}\right)_+ (\tilde{\boldsymbol{y}}_{n,\cdot})_+ ,$$

where $\tilde{\boldsymbol{Y}}_{\mathcal{N}_s,\cdot} = t(\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} - t\lambda\alpha\mathbf{1}\mathbf{1}^\top$. We will have $\hat{\boldsymbol{Z}}_{\mathcal{N}_s,\cdot} = \mathbf{0}$ if $\boldsymbol{z}^{(1)}_{n,\cdot} = \mathbf{0}$ for all $n \in \mathcal{N}_s$. Note that $\boldsymbol{z}^{(1)}_{n,\cdot} = \mathbf{0}$ if

$$\lambda(1-\alpha)t \geq \left\| \left(t \left[(\tilde{\boldsymbol{A}}^f_{\mathcal{M}(\mathcal{N}_s),\mathcal{N}_s})^\top \boldsymbol{Y}_{\mathcal{M}(\mathcal{N}_s),\cdot} \right]_{n,\cdot} - t\lambda\alpha\mathbf{1} \right)_+ \right\|_2 . \tag{S15}$$

By algebraic manipulation, the sparsity conditions given in (S14) and (S15) can be shown to be equivalent to the condition given in Lemma 13.4. Alternatively, this lemma's result also follows from inspection of the optimality condition for (6). □

15.6. *Proof of Corollary 13.5.*

PROOF. The sufficient condition given in Corollary 13.5 follows from noting that (S15) is satisfied if $\lambda(1-\alpha) \geq \left\| \left(\left[(\tilde{\boldsymbol{A}}^f)^\top \boldsymbol{Y} \right]_{k,\cdot} \right)_+ \right\|_2$ or if $\lambda\alpha \geq \left(\left[(\tilde{\boldsymbol{A}}^f)^\top \boldsymbol{Y} \right]_{k,l} \right)_+$ for $l = 1, \ldots, T$. Thus, when at least one of these two conditions is satisfied for all $k = 1, \ldots, K_f$, then the solution to (6) will be sparse. □

15.7. *Proof of Lemma 13.6.*

PROOF. Let $\hat{\boldsymbol{Z}}$ be the solution to (6). In anticipation of contradiction, assume there exists $j \in \{1, \ldots, T\}$ such that $\hat{\boldsymbol{z}}_{3,j} > 0$. Define $\tilde{\boldsymbol{Z}}$ as $\tilde{\boldsymbol{z}}_{1,\cdot} = \hat{\boldsymbol{z}}_{1,\cdot} + \left(\frac{\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,1}}{\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2})} \right) \hat{\boldsymbol{z}}_{3,\cdot}$, $\tilde{\boldsymbol{z}}_{2,\cdot} = \hat{\boldsymbol{z}}_{2,\cdot} + \left(\frac{\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,2}}{\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2})} \right) \hat{\boldsymbol{z}}_{3,\cdot}$, $\tilde{\boldsymbol{z}}_{3,\cdot} = \mathbf{0}$, and $\tilde{\boldsymbol{z}}_{k,\cdot} = \hat{\boldsymbol{z}}_{k,\cdot}$ for $k = 4, \ldots, K^f$. Let $\text{obj}(\boldsymbol{Z})$ be the value of the objective function of (6) at $\boldsymbol{Z}$ for some fixed λ and α. We have

$$
\begin{aligned}
\text{obj}(\tilde{\boldsymbol{Z}}) - \text{obj}(\hat{\boldsymbol{Z}}) &= \lambda(1-\alpha) \sum_{k=1}^{3} (\|\tilde{\boldsymbol{z}}_{k,\cdot}\|_2 - \|\hat{\boldsymbol{z}}_{k,\cdot}\|_2) \\
&= \lambda(1-\alpha)[\|\hat{\boldsymbol{z}}_{1,\cdot} + (\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,1})/(\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2}))\hat{\boldsymbol{z}}_{3,\cdot}\|_2 \\
&\qquad + \|\hat{\boldsymbol{z}}_{2,\cdot} + (\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,2})/(\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2}))\hat{\boldsymbol{z}}_{3,\cdot}\|_2 - (\|\hat{\boldsymbol{z}}_{1,\cdot}\|_2 + \|\hat{\boldsymbol{z}}_{2,\cdot}\|_2 + \|\hat{\boldsymbol{z}}_{3,\cdot}\|_2)] \\
&< \lambda(1-\alpha)[\|\hat{\boldsymbol{z}}_{1,\cdot}\|_2 + (\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,1})/(\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2}))\|\hat{\boldsymbol{z}}_{3,\cdot}\|_2 + \|\hat{\boldsymbol{z}}_{2,\cdot}\|_2 \\
&\qquad + (\mathbf{1}^\top \boldsymbol{a}^f_{\cdot,2})/(\mathbf{1}^\top (\boldsymbol{a}^f_{\cdot,1} + \boldsymbol{a}^f_{\cdot,2}))\|\hat{\boldsymbol{z}}_{3,\cdot}\|_2 - (\|\hat{\boldsymbol{z}}_{1,\cdot}\|_2 + \|\hat{\boldsymbol{z}}_{2,\cdot}\|_2 + \|\hat{\boldsymbol{z}}_{3,\cdot}\|_2)] \\
&= 0.
\end{aligned}
$$

This is a contradiction, so we conclude $\hat{\boldsymbol{z}}_{3,\cdot} = \mathbf{0}$. □

DIVISION OF BIOSTATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455
E-MAIL: pete6459@umn.edu

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WA 98195
E-MAIL: nrsimon@uw.edu

DEPARTMENTS OF BIOSTATISTICS AND STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WA 98195
E-MAIL: dwitten@uw.edu