An algorithm for spike sorting with electrical artifact: Methods

Gonzalo Mena, Lauren Grosberg, Liam Paninski and EJ Chichilnisky

June 19, 2015

Abstract

We developed an algorithm that successfully automatizes spike sorting with electrical artifact. In the following, the method is succinctly described and results are shown for 56 datasets from electrical stimulation experiments

1 Introduction

In a retinal prosthetic context one is ultimately concerned with building probabilistic models of how cells responds to electrical activity. This can be thought in the same terms as in the neural coding problem, which is stated probabilistically : what is the probability of a response across a neural population given a particular natural stimulus (for example, an image)? [**?**]. The difference is that the stimulus is now replaced by an artificial, electrical one. Whichever model is built, data is necessary for it's fitting and for subsequently providing a solution for the inverse decoding problem: how an electrical stimulatus has to be chosen int order to elicit an arbitrary response? Naturally, the most relevant information required for the fitting of any of such models is contained in the set of stimulus-response pairs, and whereas the stimulus is controlled by the experimenter, responses (spikes) have to be identified from electrode recordings. This problem, of distinguishing particular action potentials of neurons from extracellular voltage recordings, is known in the literature as spike sorting. In consequence, any framework for achieving controlled arbitrary responses in neurons via electrical stimulation will rely on spike sorting as a fundamental building block of its computational implementation. Because of the central importance of spike sorting in systems neuroscience, in the past decades many different methodologies have been developed for automatizing the detection of spikes, and significant improvements in accuracy and computational efficiency have been achieved[**?**, **?**, **?**, **?**]. However, the context of electrical stimulation constitutes a departure from the realm where current spike sorting methodologies apply, as the electrode recordings are now corrupted by the transient activity induced by this exogenous stimulation. This corruption or artifact can have an overwhelming impact in the recorded traces, making the spike identification process challenging even for of the human expert (see FIGURE WITH EXAMPLE TRACES). Actually, previous to this work the only available spike sorting method relied heavily on human judgement, and because of the difficulty of telling spikes apart from the artifact, it was extremely time consuming even for datasets of modest size. In this article we provide the first scalable algorithmic approach for spike sorting in the presence of electrical stimulation artifact.

1.1 Electrical artifact

LAUREN HAS THINGS TO SAY ABOUT THIS, WHICH ARE THE SOURCES OF THE ARTIFACT, TRIAL BY TRIAL VARIABILITY, CHANGES IN TIME AND AMPLITUDE (INCLUDING BREAKPOINTS) DISTINCTION BETWEEN HARDWARE/AXON BUNDLE The main problem is the presence of electrical artifact, which is a consequence of the electric field generated by the stimulation, and whose ubiquitous presence hampers the spike identification process. Actually, because of the electrical artifact, whose amplitude can be several times larger than of action potentials, spikes can become almost non-identifiable: suppose for example all trials at condition j have spikes and there is little variability in spiking times. Suppose also the artifact is nearly the same across trials, so the recorded traces, assumed to be the artifact plus action potentials plus noise would look almost identical for all trials, and we may conclude that either there are no spikes at all, or that there are spikes at every trial (in which case, spiking times could be any). A more dramatic example corresponds to the situation where there is only one trial per condition. Then, spike sorting is impossible as there is no way to tell the artifact and action potential (if any) apart. The moral is that if we are completely agnostic about how artifact looks like there is little we can do. However, if we impose structure about how the artifact looks like and how often spikes should show up as a function of condition, then better chances are spikes will be identified better.

2 Methods

As shown in figure WITH EXAMPLE TRACES trial by trial spike identification can be impossible even if templates are available. An appropriate experimental design is crucial to allow the simultaneous inference of spikes and artifact. This design should exploit neurophysiological the overcome the identifiability issues

2.1 Data description

Suppose spike sorting is required for a set of N neurons. Apart from the templates of these neurons, Data consists of a set of I voltage traces, or trials, measured both across time ($t = 1 \dots T$, corresponding to multiples of the sample rate) and a set of recording electrodes in the array ($e = 1 \dots E$), as a response to electrical stimulation (although in principle there are as many as 512 electrodes, we only consider for subsequent analysis the ones where the spike waveform have a strong enough signal, that should correspond to nearby locations of the somas of neurons) These traces are organized into a design as follows: first, a number of J different stimulus are chosen, defined by the current amplitudes of the pulses that are passed into a subset of electrodes in the array (that may or may not overlap with the recording electrodes). For separate datasets, references to specific currents and stimulating electrodes can be avoided, and replaced by the reference to the $j-th$ stimulus condition, since the stimulating electrodes remain the same and currents into each electrode always belong to the same line and increase monotonically with the index j. Then, for condition j, a number of I_j traces are available. Thus, $I = \sum_j I_j$. Naturally, templates of neurons for which spike sorting are required is done in

2.2 Assumptions

In order to come up with a generative model for which we can do tractable inferences, some assumptions have to be made. Although they are simplifications, taken together they seem to provide reasonable enough account of the data. First, we assume that spikes, artifact and noise interact linearly, and thus the observed traces are the sum of these three components (HOW CAN WE JUSTIFY THIS). Regarding the noise terms, they are assumed gaussian and trial-to-trial independent, and their variances electrode and condition dependent, reflecting different responses of the background neural tissue (which are believed to be responsible of the observed noise) to different stimulus (see FIGURE WITH EXAMPLE TRACES). Also, noise processes are assumed uncorrelated in in time and electrodes. This is certainly arguable: it is known that noise terms exhibit large both spatial and temporal correlations [**?**], but if these correlations were explicitly accounted for by the model then different covariance matrices should be estimated, one per each condition j. This would entail a non negligible additional computational burden, but most importantly, would prescribe the gathering of larger number of trials to allow reliable estimation of such matrices. As results with synthetic data created from the projection of real data on the simplified model shows that results are essentially the same regardless the correlations, and because in the first applications of this method we will deal with rather simple cases where noise correlations should not be a major issue, we prefer to maintain this assumption. Regarding the artifact, we assume it is a function of time and condition j, but remains the same for different trials. This comes from voltage recordings in TTX experiments, where no neural driven activity is expected to occur, and where it is observed that trial-to-trials variations in the traces within the same condition are well explained by the noise. The Artifact is also assumed to possess a regularized structure: both variations in time and condition (see FIGURE THAT SHOWS NATURE OF ARTIFACT FROM TTX EXPERIMENTS) are assumed smooth. Regarding activity of neurons for which we do spike sorting, we assume a set of N spike templates are available (SEE FIGURE WITH TEMPLATES) , each of them showing an action potential as recorded in all the E electrodes that are relevant for the analysis. Rather than an assumption, this is a reality as templates are available from previous experiments using visual stimuli, where no electrical artifact is expected. The actual underlying assumption is that templates faithfully describe the real action potentials, as they are are subject to all the estimation problems related to regular spike sorting(IS THIS THE RIGHT PALCE TO TALK ABOUT HOW TEMPLATES ARE OBTAINED?). For spike timing, it is assumed that spikes can occur only over a set of T' consecutive multiples of the sample rate, denoted $\{t_1, \dots t_{T'}\} \subseteq \{1 \dots T\}$. Again, this is to keep a tractable setup, since a model that allows spiking at arbitrary times, as the one developed in [**?**] would require computations that for now can be avoided. Also, we assume at most one spike per neuron can occur in the recording time, which is a consequence that spikes are sought in a time window that is no bigger than the usual refractory period (cite). Finally, for spiking probabilities, an underlying parametric model for the activation curves is assumed, reflecting the known sigmoideal response saturation phenomenon, which is also observed in the electrical stimulation context. (CITE about activation curves). The need of making this assumption will be clear after the generative model and algorithm for it's fitting are introduced, which is done in the following.

2.3 The generative model

Now we provide the mathematical details about the data generating process. The ultimate goal will be to make inferences about the variables that represent the artifact and spikes. Denote $Y_{t,e}^{i,j}$ the observed voltage for trial i of condition j at time t and electrode e. Then the model is

$$Y_{t,e}^{i,j} = A_{t,e}^{j} + \sum_{n}^{N} (K_n s_n^{i,j})_{t,e} + \epsilon_{t,e}^{i,j} \quad \epsilon_{t,e}^{i,j} \sim \mathcal{N}\left(0, \sigma_{e,j}^2\right) \quad \text{i.i.d} \tag{1}$$

Here $A_{t,e}^{j}$ is the artifact at time t, electrode j and condition j, and $s_n^{i,j}$ is abinary vector containing spiking information for neuron n at trial i of condition j: $s_n^{i,j}(l) = 1$ if spike occurs at time t_l. Since at most one spike per neuron occurs for a single trial, $\sum_{l=1}^{T'} s_n^{i,j}(l) \leq 1$. K_n is a $(T \times E, T')$ convolution matrix whose rows contain copies of action potentials for neuron n as recorded in all electrodes, but with spike onset aligned at all different possible spike times. For notational convenience, we also consider a vectorized version of equation (1), in which voltage traces are concatenated across time, electrode, trial and condition, to generate a unique huge vector Y. The same can be done with the artifact, spikes and neurons. This leads to the re-statement of equation (1) as

$$Y = XA + Ks + \epsilon \tag{2}$$

Where $\epsilon \sim \mathcal{N}(0, \Sigma)$, Σ with diagonal given by the $\sigma_{e,j}^2$'s and X is the covariate matrix that indicates which artifact variables are active in the different positions of the vector Y. Equivalent, in terms of the likelihood

$$p(Y|A, s, \Sigma) \propto \exp\left(-\frac{1}{2}(Y - XA - Ks)^t \Sigma^{-1} (Y - XA - Ks)\right)$$

In the following, we use the vectorized notation with the convention that whenever a sub or super script appears it refers to the sub vectors and sub matrices that are obtained when information pertaining to only those indexes is considered. Notice until now nothing has been said about the explicit structure of the artifact in the model, which needs to be constrained to avoid overfitting. We impose regularity constrains that aim to reflect our knowledge of the artifact shape, as discussed in (WHERE?). Hardware breakpoints induce, in each recording electrode, a partition of the set $\{1 \dots J\}$ into $B(e)$ inter-breakpoint ranges, denoted $\{b(j', e)\}_{j'=1 \dots B(e)}$. With this notation, we penalize the following differences:

$$\sum_{t} \sum_{j} ||A_{t+1,e}^{j} - A_{t,e}^{j}||^2, \quad \forall e = 1 \dots E$$

$$\sum_{j \in b(j',e)} \sum_{t} ||A_{t,e}^{j+1} - A_{t,e}^{j}||^2 ; \; e = 1 \dots E, \forall j' = 1 \dots B(e)$$

They are essentially smoothness constraints for the artifact on each electrode, both in conditions within the same inter-breakpoint range and in time. Although for simplicity not explicitly shown in the equations, time regularization is also done in blocks: as artifact oscillates more

around the stimulus onset than at the end of the recordings, it is expected that time smoothness is also time-varying. In our implementations we divide the recording time in two blocks, and penalize time differences within each block separately. Also, for numerical stability we include an overall ridge penalty.
We use squared l_2 regularization instead of l_1 (as in the fused lasso CITE) because we will exploit heavily the gaussian conjugacy that is entailed by the l_2 framework, and because we are not making any assumption about sparsity of artifact variables. The above squared sums can be represented as products $A_e^t D_{e,k} A_e$ for suitable semi positive definite matrices $D_{e,k}$, and to turn this into the probabilistic setting, we introduce the hyperparameters $\lambda_{e,k}$ which control the amount of regularization in each of the features. Thus, the following gaussian prior reflect our knowledge about the artifact smoothness structure.

$$p(A|\lambda) \propto \exp\left(-\frac{1}{2}\sum_{e,k}\lambda_{e,k}A_e^t D_{e,k}A_e\right)$$

For the spike probabilities, we assume a logistic regression prior for each neuron. That is, the variables $r_n^{i,j} = \sum_l s_n^{i,j}(l)$ are conditionally independent given the logistic regression parameters $\alpha_n = (\alpha_n^1 + \alpha_n^2)$, and

$$p(r_n^{i,j} = 1|\alpha_n) = \frac{1}{1+\exp\left(-\alpha_n^0 - j\alpha_n^1\right)}$$

. Regarding spike times, we don't make any explicit assumption (DISCUSS IN FURTHER DIRECTION): we set a uniform prior on the times t_l.
Finally, for the entries of the diagonal of the matrix Σ,, the variances $\sigma_{e,j}^2$ we consider a non-informative prior for the joint:

$$p(\sigma_{e,j}^2, e = 1\ldots E, j = 1\ldots J) \propto \prod_{e,j}\frac{1}{\sigma_{e,j}^2}$$

Again, this choice is made to exploit conjugacies.

2.4 Relation to other methods

FILL WITH REFERENCES AND DISCUSSION ON WHAT WE ARE TAKING FROM OTHERS, AND WHAT NO

3 The Algorithm

The algorithm we introduce here is concerned with the inference of the variables for the posterior distribution spawned by the data likelihood and the priors corresponding to the constrains in the parameters. It is important to mention that although all priors are log concave, the log posterior is not concave, because of the variables $s_n^{i,j}$, which are defined in a non concave space. If these variables were fixed to arbitrary values, then the log posterior would become concave, therefore, the number of local optima could grow exponentially with the number of trials and neurons. Many of these local optima will have a small probability, and thus if we

constrain the search to the region of the parameter space where the posterior is high, better chances are that meaningful results will be obtained. We handle that problem by taking initial values provided by a convex relaxation (SEE BELOW?) of the original problem. Notice, however, there is no reason to believe that the maximum of the posterior will provide the right spike sorting solution: we make strong assumptions about the data generating process, which can be reasonable as a first approximation, but there are other phenomena that are not being accounted for by the model (for example, spiking of neurons for which templates are not available), and thus there is not a sharp correspondence between what is told by the model and what happens in reality. For this reason we don't seek for the MAP solution (CITE); instead, we perform bayesian inference, implemented as a Gibbs sampler that exploit the inherent conjugacies of the model. In this sampling setting the question is how to transform the obtained samples in a spike sorting solution. It is the case that the Gibbs sampler rapidly lead to convergence of the variables s (FIGURE MAYBE?) that barely change after some iterations, while the rest have only small fluctuations. Roughly speaking, the Gibbs sampler gets stuck around of the peaks of the posterior. Again, the question that arises is: even if this solution should be more robust that the one obtained via optimization (coordinate ascent), how can we tell whether or not it is a meaningful one. We address this question by looking at a certain aspect of the residuals, so if a huge mismatch is detected between the current fit to the model and data, changes are made to the current solution to allow a 'jump' from the current local solution to an eventually better one. This procedure is of heuristic nature, no convergence is guaranteed but there is little that could be done in this non log concave and misspecified context. The use of these heuristics that aim to detect mismatches is highly motivated by prior experiments in which a search through the parameter space was done using sophisticated but 'blind' sampling methods, as parallel tempering or the Wang-Landau method (cite), devised to facilitate sampling of sharply peaked multimodal distributions. Application of those methods led to a sampler that indeed visited several modes, but the procedure was slow and in many times none of the visited modes corresponded to a solution. Summarizing, our algorithm has three fundamental components: an initialization to obtain a reasonable first guess, a Gibbs sampling stage that iterates until convergence to a spike sorting solution, and a set of heuristics or rules that permit the movement from 'bad' solutions to better ones. In the following we detail the computations performed in each of these stages.

3.1 Initialization

The initialization is divided in two stages. In the first, we solve a quadratic program (cite) that can be deemed as a relaxation of the maximization of the posterior. In the second part, given the initial we set the initial values for all the variables, most importantly, the regularization hyperparameters λ needed for the Gibbs sampler.

3.1.1 Convex relaxation via quadratic programing (QP)

Instead of considering the discrete space for the spiking binary vectors $s_n^{i,j}$, we allow them to belong to their convex hull (CITE), the $(T'-1)$-probability simplex: $\{s \in [0,1]^{T'} : \sum_{l=1}^{T'} s(l) \leq 1\}$. In this new setting we no longer look for spikes but rather 'generalized spikes': presence or absence of spikes at certain times is replaced by a probability distribu-

tion. The new parameter space is convex, and thus the (log negative) posterior is convex. Unfortunately, this relaxation requires intensive computations, and depends on unknown regularization hyperparameters. Our approach is to re-state the maximum a posteriori convexified program as a QP (CITE), since in this framework there are a number of efficient solvers available. This re-statement won't correspond exactly to the original problem, and the question is how to faithfully capture the structure imposed by the original model in the QP setting. We take the following approach: to avoid artifact regularization hyperparameters, we consider instead a simple polynomial model; for each time artifact is assumed a polynomial function of condition, for conditions in the same inter-breakpoint range. This leads to the replacement of artifact XA (equation VECTORIZED) by $X'A'$, with A' the new artifact variables and X' representing covariates in this polynomial representation. The logistic regression prior is replaced by the (linear) constraint that spiking probabilities increase with condition, for each neuron. That is, for $j = 1 \dots J-1$ and $n = 1 \dots N$ we set $\sum_{i,l} s_n^{i,j}(l) \leq \sum_{i,l} s_n^{i,j+1}(l)$. Finally, the non-informative prior for the variances is dropped.
In the QP we aim to minimize the residual sum of squares (RSS) that comes from the likelihood, and is a quadratic function of the variables: indeed, notice that artifact and spike variables can be concatenated, by defining $z = (A', s)$. Moreover, the artifact and (generalized) spike activity terms can be expressed as a unique matrix product between some matrix M and z. With this notation the RSS is $||Y - Mz||^2$ and the problem can be stated as follows

$$\min_{z=(A',s),\rho} \quad ||Y - Mz||^2 \tag{3}$$

$$\begin{aligned}
\text{s.t.} \qquad & \forall i,j,n \;\; s_n^{i,j} \in (T'-1) - \text{probability simplex} \qquad \text{(generalized spikes)} \\
& \forall n, j = 1 \dots J-1 \; \textstyle\sum_{i,l} s_n^{i,j}(l) \leq \sum_{i,l} s_n^{i,j+1}(l) \qquad \text{increasing spike probabilities} \\
& ||Y - Mz||^2 \leq \rho \qquad \text{Explaining data}
\end{aligned}$$

There is no a priori rule to select the degree of the polynomials apart from privileging low degrees (e.g., not greater thatn 3) to avoid overfitting. However, ultimate decisions about what degree to choose are made by the context (see results section). (THIS IS FURTHER RESEARCH?) There is a possibility of including another term in the objective to enhance sparseness of solutions, via the l_1 sum of the spike variables. However, this would imply the tuning of another parameter that modulates the trade-off between minimizing the RSS and maximizing sparseness. This could be done via cross validation (CITE), which could be computationally tractable only if an efficient warm start method is available for this problem.

3.1.2 Initial values for Σ, α, A and λ

Initial values for the residual variances can be obtained by breaking down the resulting RSS: $\sigma^2_{j,e_0} = \frac{1}{TI}||Y_{j,e} - M_{j,e}z||^2$ where $Y_{j,e}$ and $M_{j,e}$ are obtained by considering only the rows corresponding to condition j and electrode e. [1] For α_0, the initial parameters of the logistic regressions, we consider a least square fit with the obtained activation curves (DISCUSS ABOUT THIS WITH LIAM). Regarding the artifact, an initial solution is obtained by just taking $A_0 \equiv X'A'$. This initial artifact is used to compute the values of the hyperparameters λ,

[1] Notice the above estimates are biased and should be corrected by the number of degrees of freedom, however, this has little impact in results and thus this further step is avoided.

which are found via the maximization of $p(A_0|\lambda)$ with respect to λ. This leads to the following problems (one for each electrode)

$$\lambda_0 = \arg\min_{\lambda_e} \tfrac{1}{2} A_{0,e}^t \left(\sum_k \lambda_{e,k} D_{e,k}\right) A_{0,e} - \log \left| \sum_k \lambda_{e,k} D_{e,k} \right|$$
$$\text{s.t.} \quad \lambda_e \geq 0$$
$$\textstyle\sum_k \lambda_{e,k} D_{e,k} \succ 0$$

The above (convex) problem is known in the literature as *MAXDET* (cite), and we solve it by gradient descent. Once the values of λ are set, they are not subjected to any further change. Essentially what we do is to drop the artifact polynomial structure and replace it by a more flexible one, that will allow to account for departures of the initial assumption. Here we are taking our artifact initial estimate as a 'point of truth', which is certainly arguable as this initial solution doesn't have to be the actual one. However, we take this 'point of truth' in the absence of more information, and because in the worst case at least gives a rough estimate of the orders at which the λ's fluctuate.
Notice the only parameter we have not initialized is the first estimate of the spikes s_0. However, this is not necessary, the Gibbs sampler can sample from them in the first place provided that all the rest of the variables have been set.

3.2 Gibbs sampler

Given our initial solution we would like to sample from regions of high posterior probability and eventually converge to a spike sorting solution s for which no changes are obtained if further iterations are made. This can be easily implemented in a Gibbs sampler (CITE) that samples from the joint distribution of blocks of variables (for example, all the s) conditional on the data and the rest of the variables.
Now we detail the several stages of the Gibbs sampler

3.2.1 Sampling spikes

The conditional spike probabilities for a neuron at a given trial and condition, given all the rest of the variables and data is multinomial:

$$p(s_n^{i,j}(l) = 1|Y, A, \Sigma, \lambda, \alpha, s\backslash_n^{i,j}) \propto \frac{p(r_n^{i,j}(l) = 1|\alpha_n)}{T'} \exp\left(-\frac{1}{2}\sum_e \frac{||Y_e^{i,j} - \sum_n (K_n s_n^{i,j})_e - A_e^j||^2}{\sigma_{e,j}^2}\right)$$

Notice posterior spike probabilities of a given trial don't depend on other trials or conditions, but spiking in one neuron does depend on spiking from others. A swipe of this block sample goes sequentially though all neurons and for within each neuron sequentially through all trials.

3.2.2 Sampling artifact

Denoting $\Lambda_e = \left(\sum_k \lambda_{e,k} D_{e,k}\right)^{-1}$ we have the following expression for the distribution of the artifact, conditional on the rest of the variables and data.

$$A_e|Y, s, \Lambda, \Sigma, s, \alpha, A\backslash_e \sim \mathcal{N}(\mu_e, \Sigma_e') \quad \Sigma_e' = \left(X_e^t \Sigma_e^{-1} X_e + \Lambda_e^{-1}\right)^{-1} \quad \mu_e = \Sigma_e' X_e^t \Sigma_e^{-1} (Y_e - K_e s),$$

Again, a swipe consists in the sequential sampling of the artifact on the different electrodes.

3.2.3 Sampling from Σ

Sampling is straightforward once noticing that the conditional distribution of $\sigma^2_{e,j}$ conditional on data and the rest of variables follows an inverse gamma distribution.

$$\sigma^2_{e,j}|Y, s, A, \alpha, \lambda, \sigma^2\backslash_{e,j} \sim \text{ Inv-Gamma}\left(\frac{TI_j - 1}{2}, \frac{1}{2}||Y^j_e - X^j_e - K^j_e s||^2\right)$$

3.2.4 Esimation of α

For the logistic regression parameters α_n we don't follow a sampling approach, as there are no conjugacies to exploit, and because results don't sensibly change if sampling is replaced by an optimization procedure [2]: instead, we aim to minimize for each neuron

$$-\log p(\alpha_n|s, A, Y, \Sigma, \lambda, \alpha\backslash_n) \propto \sum_{j,i} \log\left(1 + exp\left(-(2r^{i,j}_n - 1)(\alpha^0_n + j\alpha^1_n)\right)\right)$$

However, some regularization is needed: first, there is no solution for logistic regression in the separable case, which in this context corresponds to the situation where there is no spiking until condition j, and from condition j on there are spikes at all trials. Also, we want to rule out solutions where the spike probabilities are decreasing in j, so we explicitly look for solutions such that $\alpha^2_n \geq 0$. These requirements can be easily met if we consider a constrained logistic regression with a small ridge penalty w:

$$\alpha_n = \arg\min_{\alpha^0_n, \alpha^1_n \geq 0} \sum_{j,i} \log\left(1 + exp\left(-(2r^{i,j}_n - 1)(\alpha^0_n + j\alpha^1_n)\right)\right) + w\sum_{m=0}^{1}(\alpha^m_n)^2$$

Hopefully the solution s obtained by convergence of the above scheme will correspond to the right spike sorting, and comparison with ground truth shows that is the case sometimes, but in general one still may obtain nonsensical spike trains: in the most typical of such situations spike sorting is correct for some conditions, but there are others in which no spikes are found, where in reality all trials have spikes. Also, it is possible to infer spikes in all trials while in reality spikes are present in some or none of them. In other words, either the artifact or spikes can overfit the data. These situations are largely a consequence of sudden changes in the transient dynamics recorded in the electrodes driven by changes in the stimulus. In the following we describe in detail these pathological situations and introduce heuristic methods for detecting potential nonsensical solutions and subsequently correcting them.

3.3 Post processing

There are several ways to assess the plausibility of solutions provided by the Gibbs sampler: the one that at first seems the most obvious is to look at the residual variances at each condition, as in conditions where the fit is poor these residuals should be higher. However, the problem of just looking at these residuals is that there are other reasons that could drive an increase in the RSS: as the response of the neural tissue changes from one condition to the next, there is no way to tell if an increase in the RSS is due to a poor fit or to a change in the underlying

[2] Notice by doing thus we no longer have a *bona fide* Gibbs sampler, but instead, a Gibbs-EM hybrid

noise process. Conversely, residual variances can be low but it can be the case that either artifact or spikes are overfitting the data. A more amenable alternative comes from assessing differences between empirical spiking probabilities at each condition, $p_n(j) = \sum_i r_n^{i,j}/I(j)$ and probabilities of spiking provided by logistic regression, $\hat{p}_n(j) = p(r_n^{i,j} = 1|\alpha_n)$. To rely on this diagnostic criterion, we need to make the assumption that logistic regression provides a faithful description of the activation curves phenomenon, which seems to be the case (CITE). With this assumption, lack of fit in this aspect of the model diagnoses the attained solution is not explaining data in a crucial sense. However, there are still, rather pathological cases in which even if this fit is good, it can be shadowing more obscure problems: first, activation curves may look nice and have an almost perfect fit to logistic regression, but it could be the result of spikes overfitting data. This can happen, for example, following a breakpoint or at the onset of axonal bundle activation: in those cases it is difficult to tell if changes in data are due to either onset of steady spiking (low variability in latencies), the sudden activation of the bundle or a completely new shape of the electrical artifact. The diagnostic in that situation should not be based on goodness of fit but on a sharp increase in spiking probabilities. Also, if for all conditions in an inter-breakpoint range spiking is steady, then the artifact could overfit data and no spikes will be detected, giving a rather noisy activation curve with 'islands' of high activity followed by lack of activity, and as the logistic regression fit will correspond to a roughly flat curve, no lack of fit will be detected. The diagnostic in this case will be based on detecting sudden drops in the activation curve.

Regardless of the specific diagnostic, something is needed to 'jump' from the current local optima to a better one. We do so using heuristic procedures that aim to locally perturbe the current state of the sampler so that chances are that a better solution will be attained. They all share that changes are done either to the current spike sorting solution s or to the artifact. For the former, changes take the form of a controlled deletion or addition of spikes. For the latter, changes are made to the artifact: at certain suspicious conditions artifact is re-sampled from it's prior gaussian distribution, conditional on current values at other 'non suspicious' conditions. In detail, if E_1 is a subset of the E electrodes, J_1 is the set of conditions for which we require changes, and J_2 is a non-overlapping set of conditions, we re-sample the artifact from $p(A_{E_1}^{J_1}|\lambda, A_{E_1}\backslash^{J_2})$. If J_2 is the complement of J_1 we interpolate, and if $J_2 = [1, \ldots, \max J_1]J_1$ we extrapolate. The electrodes in E_1 are called 'preferred electrodes' of a neuron, they are chosen beforehand and are the ones (or the one) where the action potentials have the strongest signal (recall changes are proposed from lack of fit with information specific to neurons, so if we want to produce only local changes then a correspondence between neurons and electrode has to be stated). In spirit, we borrow statistical strength from conditions where we are more confident the fit is good. After doing this local changes (either to artifact, spikes or both), to attain a new solution the Gibbs sampler is executed until convergence to a new s, that may or may not be different from the original one.

In the following we give implementation details of these heuristics. None of them gives the ultimate solution to all the described pathological cases, but taken together, and implemented sequentially (in the same order as presented) they seem to actually lead to solutions that in the vast majority of analyzed cases correspond, or are close to the ground truth (see results section for details).

3.3.1 Interpolation at bad fit conditions

For the empirical activation curves $p_n(j)$, χ^2 goodness of fit tests [**?**] are performed to detect departures from the logistic regression null model. These tests are based on the deviance statistic

$$D_n = \sum_{j=1}^{J} D_n(j), \quad D_n(j) = 2I_j\Big(p_n(j)\log\left(\frac{p_n(j)}{\hat{p}_n(j)}\right) + (1-p_n(j))\log\left(\frac{1-p_n(j)}{1-\hat{p}_n(j)}\right)\Big)$$

If lack of fit is diagnosed, then interpolation is made in the condition where $D_n(j)$ is highest and then Gibbs sampling is done until convergence to another solution. If the new solution s is equal to the previous, then mode conditions with high $D_n(j)$ are included for interpolation (one by one), until changes are achieved. The procedure continues until no lack of fit is detected, or until the size of J_1 is greater than a pre-especified number (e.g. 3).

3.3.2 Extrapolation if sudden drops in activity are observed

In the case where a sudden drop of activity follows a the onset of activation (conditions with activation are detected as the ones where $p_n(j) \geq 0.5$ extrapolation is done in the conditions with lack of activity, based on the information until the last condition with detected activation. Again, this is done starting with only one condition, and adding more to the set J_1 if no changes are obtained. Termination occurs either when the situation was corrected, or when J_1 is big enough.

3.3.3 Extrapolation for the first condition of high activity

To correct possibly sudden increases in activation curves not driven by actual spikes, if the first condition with activation has also a high activation (defined by $p_n(j) \geq 0.8$) we extrapolate from that condition based in all previous ones. The procedure continues until no such conditions are found, or until no changes in s are produced by Gibbs sampling.

3.3.4 Deletion of spikes followed by extrapolation

This is the more aggressive heuristic, as an arbitrary deletion of spikes can lead to a lost of valuable information of spikes that were actually there. For this reason we do it in a controlled way, and keeping the solution previous to this heuristic at hand to go back to it if it is found that deletion leads to a worse situation. For deletion we start with a condition defined in the same way as the previous heuristic, delete all spikes and execute the Gibbs sampler until convergence. If the residual variance at that condition increases too much (by a factor of 1.5) we come back to the solution with no deletion, otherwise we continue deleting spikes at the following condition and Gibbs sampling, until the residual variances increase or until we delete spikes at condition J. Afterwards, the heuristic swipes across the set J_{ex} of consecutive conditions between the first for which there is activation and the last for which spikes were deleted. Iterations of the swipe increasingly take a single condition $j \in J_{ex}$, and in each of them a extrapolation of the artifact with $J_1 = j$ and $J_2 = 1\ldots, j-1$ is made, followed by Gibbs sampling until convergence. Extrapolation is made to correct eventual undesired deletion of spikes at conditions with steady spiking. In those conditions, if a deletion is made

the artifact could end up overfitting data, which won't be diagnosed by an increase in the residual variances. We choose the first condition of activation to start with because at the onset of spikes (CITE?) there is more variability in the latencies, preventing the artifact, that doesn't depend on particular trials, from overfitting. Thus, this condition can be used as a reliable pivot for obtaining good estimates of the artifact at the following conditions.

3.3.5 Addition of spikes

This heuristic addresses the cases where no spiking is observed in a certain inter-breakpoint range, but where there is activation in the last condition previous to that range. In this situation there is artifact overfit because of the low variability in spike latencies (if there was enough variability it would have been detected). Thus, even if spikes cannot be distinguished, there is evidence that they are there, which prescribes the addition of spikes. They are added at the same latency as the median of the latencies of the condition with activation, and after placement, Gibbs sampling is executed until a new convergence is achieved, with the aim of updating the artifact estimate and detecting possibly misplaced spikes.