EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE
FAKULTÄT

Least Informative Dimensions

Fabian Sinz[†,1,2], Anna Stöckl[*,3], Jan Grewe[†,1], Jan Benda[†,1,2]

[†]{first.last}@uni-tuebingen.de, [*]anna.stockl@biol.lu.se

[1]Institute for Neurobiology, Department for Neuroethology, University Tübingen [2]Bernstein Center for Computational Neuroscience, Tübingen
[3]Department for Functional Zoology, Lund University, Sweden

Abstract

We present a novel non-parametric method for finding a subspace of stimulus features that contains all information about the response of a system. Our method generalizes similar approaches to this problem such as *spike triggered average*, *spike triggered covariance*, or *maximally informative dimensions*. Instead of maximizing the information between features and responses directly, we *minimize* the information between non-informative features and the pair of informative features and system responses. This has the advantage that we can use certain *integral probability metrics* that are computationally much more feasible than mutual information estimation. Estimators of these metrics are (i) easy to compute and (ii) exhibit good theoretical convergence properties which are independent of the dimensionality of the data. For that reason, our method can be easily generalized to populations of neurons or spike patterns. By using a particular expansion of the mutual information, we can show that the informative features must contain all information if we can make the un-informative features independent of the rest.

Decomposition of information

Direct approach: If $Q \in SO(n)$ and $(\mathbf{U},\mathbf{V})^\top = Q\mathbf{X}$ the mutual information between stimuli $\mathbf{X}$ and responses $\mathbf{Y}$ can be decomposed into

$$I[\mathbf{Y}:\mathbf{X}] = I[\mathbf{Y}:\mathbf{V}] + \mathrm{E}_{\mathbf{V}}[I[\mathbf{Y}|\mathbf{V}:\mathbf{U}|\mathbf{V}]].$$

For maximally informative features one could either maximize $I[\mathbf{Y}:\mathbf{V}]$ (hard or infeasible) or minimize $\mathrm{E}_{\mathbf{V}}[I[\mathbf{Y}|\mathbf{V}:\mathbf{U}|\mathbf{V}]]$.

Our approach: Minimize

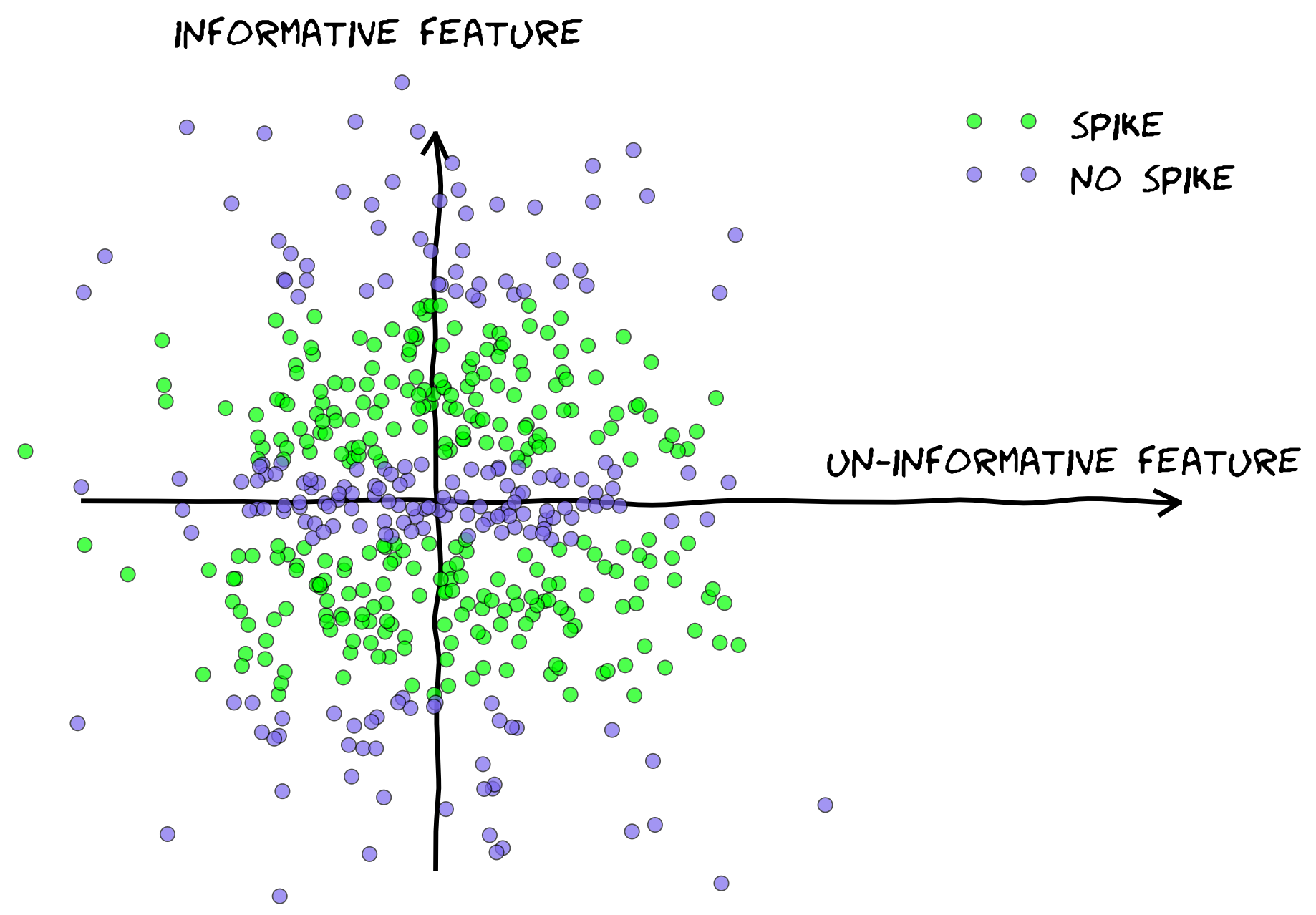
$$I[\mathbf{Y},\mathbf{U}:\mathbf{V}] = I[\mathbf{Y}:\mathbf{X}] + I[\mathbf{U}:\mathbf{V}] - I[\mathbf{Y}:\mathbf{U}]$$

because

$$I[\mathbf{Y},\mathbf{U}:\mathbf{V}] = 0 \quad \text{implies} \quad I[\mathbf{Y}:\mathbf{X}] = I[\mathbf{Y}:\mathbf{U}]$$

and $\mathbf{U}$ carries all information about $\mathbf{Y}$.

General goal



Goal: Find orthogonal matrix Q that decomposes the stimulus $\mathbf{X}$ into features

$$(\mathbf{U},\mathbf{V})^\top = Q\mathbf{X}$$

that are informative and un-informative about the responses $\mathbf{Y}$, respectively.

Approach: *Minimize* the information between un-informative features and the responses. Thus, the informative features should carry all the information.

Generalization: The responses $\mathbf{Y}$ are not limited to single spike responses. Therefore, our approach can be sensible to correlations between successive spikes or population responses and generalizes spike triggered techniques and *maximally informative dimensions*.

Why un-informative features?

Direct estimation infeasible: Finding informative features $\mathbf{U} = Q_{\mathbf{U}}\mathbf{X}$ via direct estimation of the mutual information

$$I[\mathbf{U}:\mathbf{Y}] = D_{KL}[p(\mathbf{U},\mathbf{Y})\|p(\mathbf{U})p(\mathbf{Y})]$$

between $\mathbf{U}$ and the responses $\mathbf{Y}$ is computationally expensive or infeasible because every new choice of Q_U requires the re-estimation of $I[\mathbf{U}:\mathbf{Y}]$.

Alternative measures: *Maximum mean discrepancy (MMD)* metrics based on characteristic reproducing kernels k with associated RKHS $\mathcal{H}$ provide a different class of divergences between two distributions

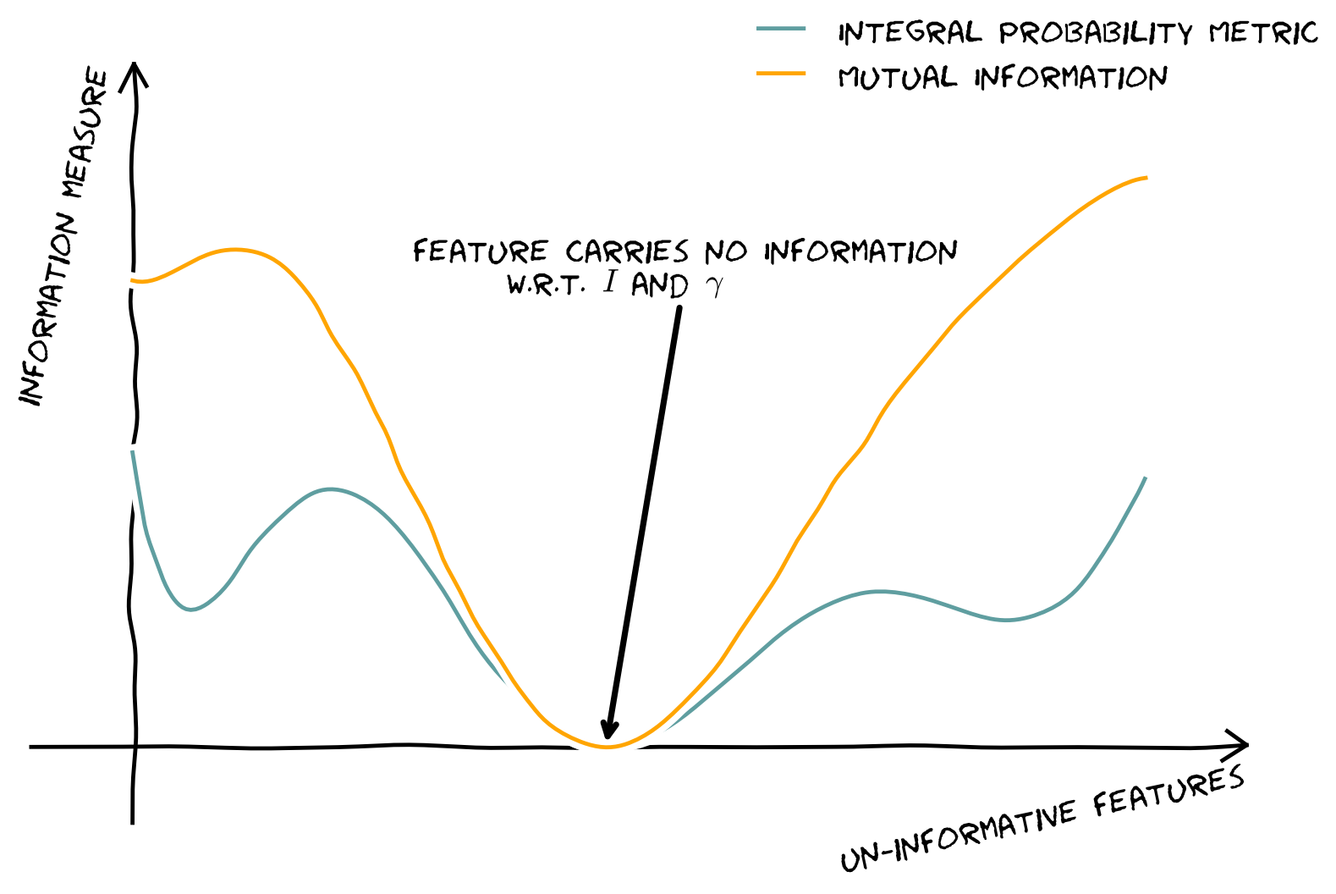
$$\gamma^2_{\mathcal{H}}(p(\mathbf{Z}_1), p(\mathbf{Z}_2)) = \|\mathrm{E}[k(\cdot,\mathbf{Z}_1)] - \mathrm{E}[k(\cdot,\mathbf{Z}_2)]\|^2_{\mathcal{H}}$$

can be computed easily by averaging over kernel matrices. They are a special case of *integral probability metrics (IPMs)*

$$\gamma_{\mathcal{F}}[\mathbf{Z}_1:\mathbf{Z}_2] = \sup_{f\in\mathcal{F}} |\mathrm{E}[f(\mathbf{Z}_1)] - \mathrm{E}[f(\mathbf{Z}_2)]|.$$

The empirical estimation of MMD can be shown to converge in $\mathcal{O}\left(1/\sqrt{m}\right)$ (independent of dimensions, m number of data points).

IPMs and mutual information coincide only at the minimum: IPMs are different from ϕ-divergences like the Kullback-Leibler divergence D_{KL} which the mutual information I is a special case of. This means that maximization of $\gamma^2_{\mathcal{H}}[p(\mathbf{Y},\mathbf{U}), p(\mathbf{Y})p(\mathbf{U})]$ potentially leads to different results than maximization of $I[\mathbf{U}:\mathbf{Y}]$.


INTEGRAL PROBABILITY METRIC
MUTUAL INFORMATION
INFORMATION MEASURE
FEATURE CARRIES NO INFORMATION W.R.T. I AND γ
UN-INFORMATIVE FEATURES

However, they share the same minimum:

$$\gamma_{\mathcal{F}}(\mathbf{Z}_1:\mathbf{Z}_2) = 0 \Leftrightarrow D_{KL}[\mathbf{Z}_1:\mathbf{Z}_2] = 0.$$

This means that only completely uninformative features in $\gamma^2_{\mathcal{H}}$ are also uninformative in I since $I = \gamma = 0$ links them. For that reason we *minimize* the information in $\gamma^2_{\mathcal{H}}$ (see Box *Decomposition of Information*)

Implementation

Objective: Use *Hilbert-Schmidt Independence Criterion (HSIC)*

$$\text{minimize}_Q \hat{\gamma}^2_{hs} = \text{minimize}_Q \frac{\mathrm{tr}(K_1 H K_2 H)}{(m-1)^2}$$

where K_1 and K_2 denote the matrices of pairwise kernel values between the data sets $\{(\mathbf{u}_i,\mathbf{y}_i)\}_{i=1}^m$ and $\{\mathbf{v}_i\}_{i=1}^m$, respectively, and $H_{ij} = \delta_{ij} - m^{-1}$.

Kernel: Use RBF kernels

$$k(\mathbf{z}_1,\mathbf{z}_2) = \exp\left(-\frac{\|\mathbf{z}_1-\mathbf{z}_2\|_2^2}{\lambda^2}\right),$$

where $\mathbf{z}_i$ denotes either $\mathbf{v}_i$ or $(\mathbf{u}_i \otimes \mathbf{y}_i) = \mathbf{u}_i\mathbf{y}_i^\top$.

Efficiency: Use incomplete Cholesky decomposition to compute low-rank approximations of K_1 and K_2 and evaluate γ^2_{hs} efficiently.

Related work

Spike triggered covariance takes eigenvectors with largest absolute values of

$$C_{X|\text{spike}} - C_X,$$

where $C_{X|\text{spike}}$ and C_X are the covariance of the spike-triggered ensemble and the stimulus, respectively. It cannot find a subspace for the toy example in *General Goal* since $C_{X|\text{spike}} - C_X = 0$ in that case.
Maximally informative dimensions directly maximizes

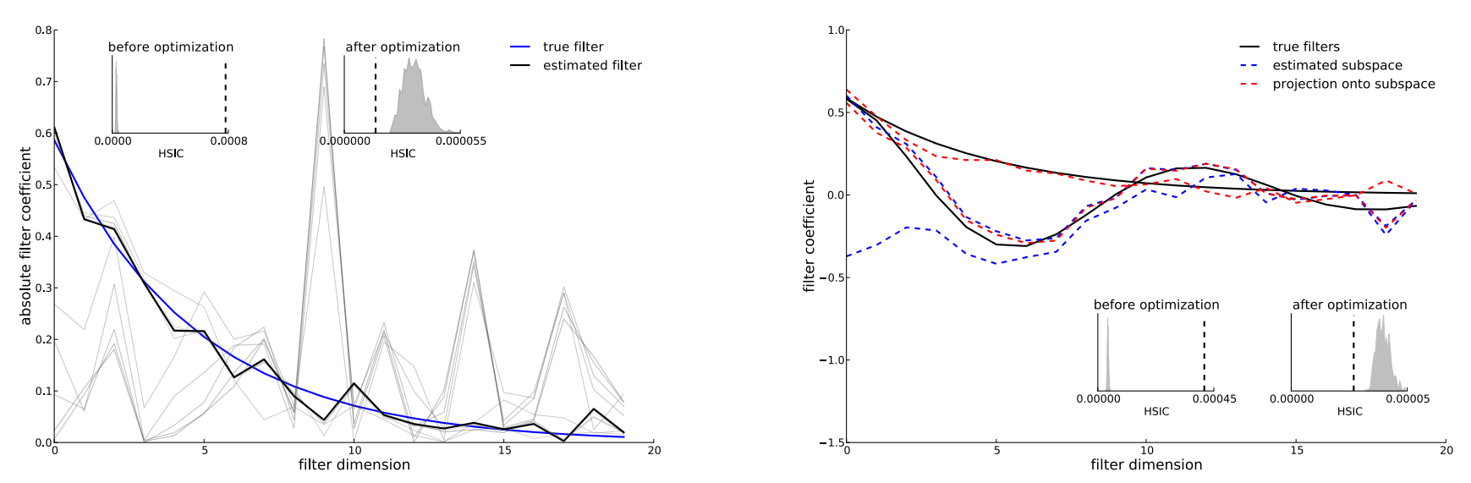
$$I_{\text{spike}} = D_{KL}\left[p\left(\mathbf{v}^\top\mathbf{s}|\text{spike}\right)\|p\left(\mathbf{v}^\top\mathbf{s}\right)\right]$$

The generalization to spike patterns or population responses $\varpi_1,\ldots,\varpi_\ell$ would be $I\left[\mathbf{v}^\top\mathbf{s}:\varpi\right] = \sum_i p(\varpi_i)\cdot I_{\varpi_i}$.
Kernel dimension reduction in regression minimizes $\mathrm{E}_{\mathbf{U}}[I[\mathbf{Y}|\mathbf{U}:\mathbf{V}|\mathbf{U}]]$ (via IPMs) with kernels. It makes less restrictive assumptions, but it needs to invert a large kernel matrix and cannot easily generate a null distribution via shuffling.

Results

Insets show Null-distributions of $\hat{\gamma}^2_{hs}$ obtained via shuffling $(\mathbf{u}_i,\mathbf{y}_i)$ pairs vs. $\mathbf{v}_i$ across trials. Stimulus distributions are white Gaussian noise (all toy examples) or band-pass filtered Gaussian noise (P-Unit).

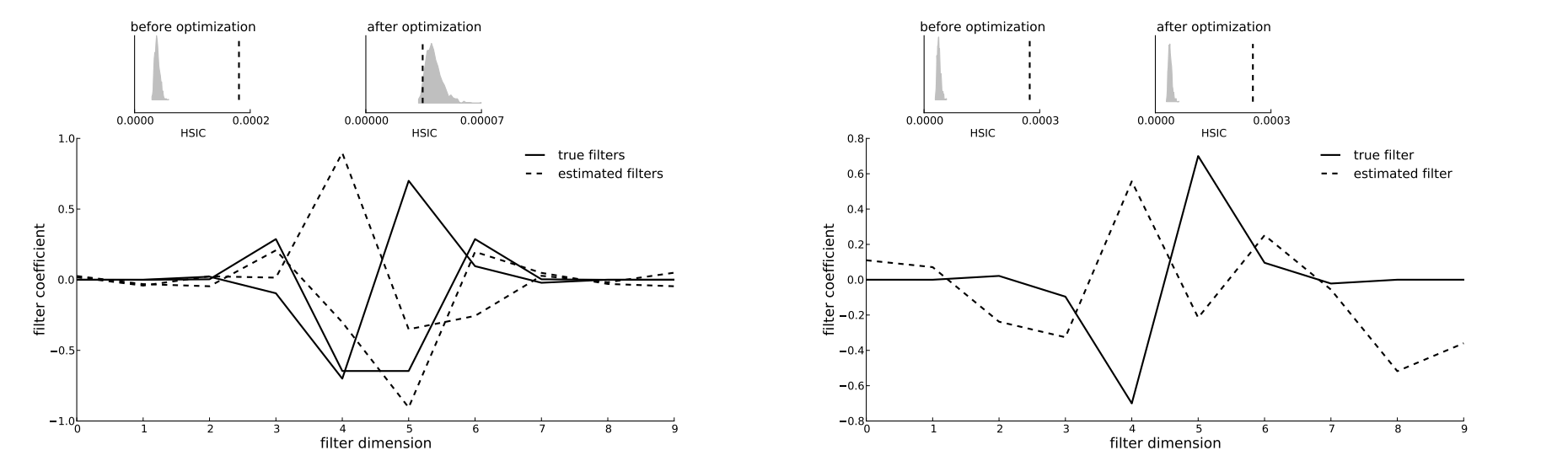
LNP neuron and 2-state neuron



Left: Informative dimension of LID trained on simple linear nonlinear Poisson (LNP) neuron $y_i \sim \text{Poisson}\left(\lfloor\langle\mathbf{w},\mathbf{x}_i\rangle - \theta\rfloor_+\right)$ with an exponentially decaying filter and a rectifying non-linearity.

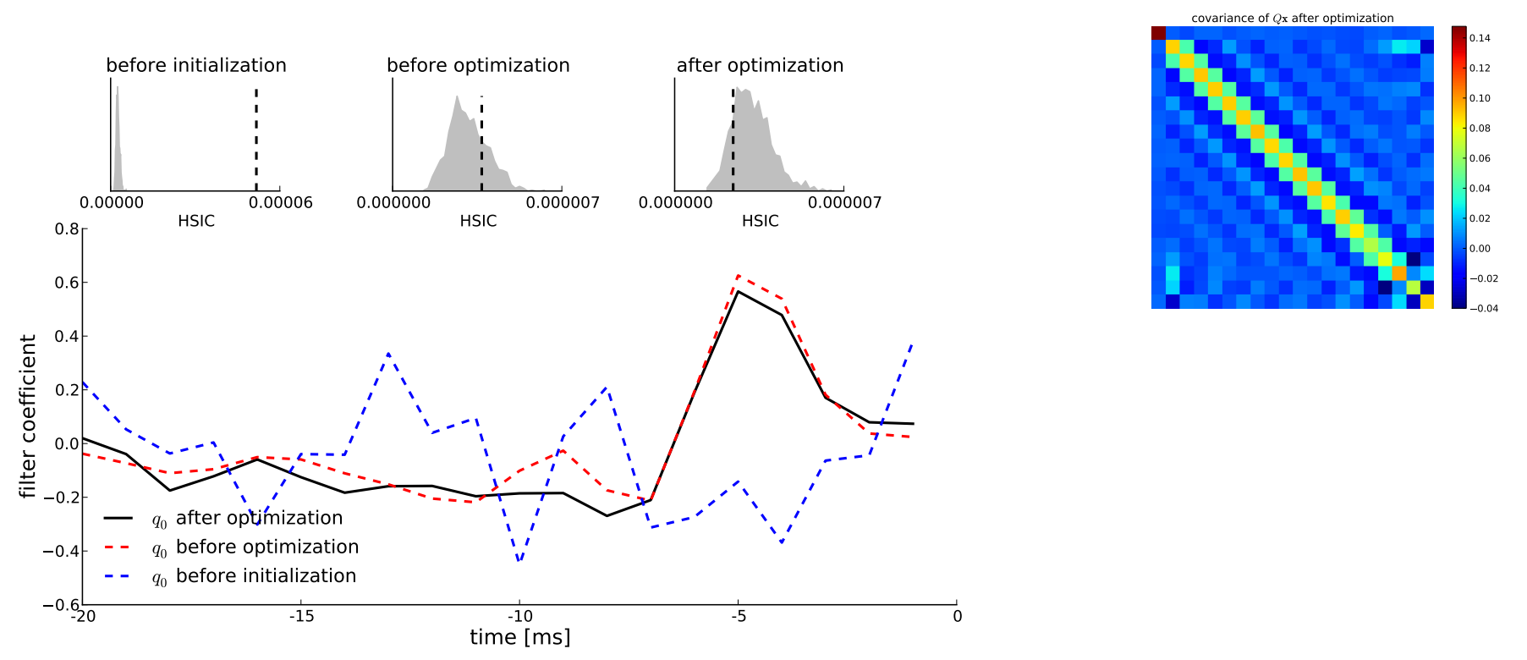
Right: Informative dimensions of LID trained on a simulated neuron with two states that were both attained in 50% of the trials. Stationary state: four output bins are drawn from an LNP neuron with exponentially decaying filter. Burst state: first two bins are drawn from Poisson distribution with a fixed base rate independent of the stimulus and the second two bins are drawn from an LNP neuron with a modulated exponential filter and higher gain

Artificial complex cell



Informative dimensions of LID trained on the response of an artificial complex cell simulated by sampling responses from a Poisson distribution with its rate given by $\lambda_i = \langle\mathbf{w}_1,\mathbf{x}_i\rangle^2 + \langle\mathbf{w}_2,\mathbf{x}_i\rangle^2$, where $\mathbf{w}_1$ and $\mathbf{w}_2$ are quadrature pair filters.

P-Unit recordings from weakly electric fish



A random filter (blue trace) exhibits $\hat{\gamma}^2_{hs}$ values that are clearly outside the domain of the null distribution (left inset). Red trace depicts the spike triggered average, the black trace is the feature found by LID.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

WERNER REICHARDT CENTRUM
FÜR INTEGRATIVE
NEUROWISSENSCHAFTEN

bccn
tübingen