2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

CONVOLUTIONAL HIGHER ORDER MATCHING PURSUIT

Gergő Bohner, Maneesh Sahani

University College London
Gatsby Computational Neuroscience Unit
25 Howland Street, London W1T 4JG.

ABSTRACT

We introduce a greedy generalised convolutional algorithm to efficiently locate an unknown number of sources in a series of (possibly multidimensional) images, where each source contributes a localised and low-dimensional but otherwise variable signal to its immediate spatial neighbourhood. Our approach extends convolutional matching pursuit in two ways: first, it takes the signal generated by each source to be a variable linear combination of aligned dictionary elements; and second, it executes the pursuit in the domain of high-order multivariate cumulant statistics. The resulting algorithm adapts to varying signal and noise distributions to flexibly recover source signals in a variety of settings.

Index Terms— Matching pursuit, feature decomposition, higher order, multi-sample, convolutional

1. INTRODUCTION

Sparse signal decomposition is a long-standing and well-studied problem, with many applications in audio- and image-processing. The most common case seeks to decompose a multivariate signal $\mathbf{y}$ into a sparse linear combination of basis elements drawn from a known overcomplete dictionary $\{\mathbf{b}_1 \dots \mathbf{b}_K\}$, with weight vector $\mathbf{x} = [x_1 \dots x_K]$, that solves

$$\min_{\mathbf{x}} \left\| \mathbf{y} - \sum_{k=1}^{K} x_k \mathbf{b}_k \right\|_2 + \gamma \|\mathbf{x}\|_0 \qquad (1)$$

for an appropriate sparsity parameter γ. If $\mathbf{y}$ is an image, the coefficients x_k may represent unknown signal amplitudes, with the corresponding basis element $\mathbf{b}_k$ reflecting the spatial influence of the unknown signal. An exact solution requires evaluating all 2^K subsets of the dictionary and so approximations are necessary in large-scale problems. One choice is the convex relaxation, replacing the L_0 norm by its L_1 counterpart, which leads to a family of methods called Basis Pursuit. The other common approach, which we expand upon here, involves the greedy sequential selection of non-zero coefficients. This is known as Matching Pursuit [1].

In the basic formulation of Eq. 1, the support of the basis elements coincides with that of the data vector, and so projection into the basis requires computation that scales with both data and basis size. Often, the sources of the unknown signals are separate objects, which influence the measurements through local, stereotypical features with restricted support. In this case, each element of the dictionary may be constructed by translating a single feature to a new location, and projections may be achieved efficiently by convolution. Greedy search then results in an algorithm called Convolutional Matching Pursuit [2].

Convolutional Block Matching Pursuit [3] allows greater variability in the shape of the sparse unknown object signals, with each described by a linear combination of a limited set of basis functions. Sparsity, in this approach, applies to the number of component signals, but not to the basis functions explored by each signal. This allows the method to identify better localised objects, and more interpretable signals. In particular, in biological imaging applications, the signals may correspond to individual cells of variable appearance [3].

Here, we consider a setting in which multiple data samples or frames are available, each generated by the same set of underlying objects with varying signal amplitudes. Again, the goal is to infer the object locations, often as a prelude to extracting the associated varying signal amplitudes. Examples include sequences of micrographs of the same field of view, repeated radio signal packets, or seismic recordings.

Below we describe how to efficiently exploit information provided by the multiple samples to identify objects by Convolutional Higher-Order Matching Pursuit (CHOMP). Section 2 defines the generative model associated with the problem. Section 3 describes the inference algorithm, evaluated on simulated data in Section 4. Finally, we discuss the results and possible further avenues of research in Section 5.

2. PROBLEM DESCRIPTION

Consider data generated by a set of stationary objects O located within a d-dimensional space. Each object generates a

This work was supported by the Gatsby Charitable Foundation and the Simons Foundation (SCGB 323228, MS).

978-1-5090-0746-2/16/$31.00

measurable signal in a finite d-dimensional patch around its centre, described by a linear combination of known basis elements $\{\mathbf{b}_k\}_{k=1}^K$ centred at the object location. We obtain T noisy sample measurements $\{\mathbf{y}^t\}$, with conserved object locations but possibly variable signals. Our goal is to recover the object locations, which may then be used to infer the time-varying signals $\{\mathbf{x}^{s,t}\}$ for each object s.

2.1. Generative model

We write $\mathbf{y} \in \mathbb{R}^{|\mathcal{I}|\times\mathcal{T}}$ for the entire collection of measurements, where $\mathcal{I} = \mathcal{I}_1 \times \mathcal{I}_2 \times \ldots \times \mathcal{I}_d$ is a set of tuples indexing points in the d-dimensional data space, and T is the number of measurements. The measured value at location $l \in \mathcal{I}$ in measurement t is $y_l^t \in \mathbb{R}$.

The objects in the set $\mathsf{O} = \{O_s\}_{s=1}^S$ each influence a confined cuboidal region around their corresponding centre location l^s, which we call a "patch" (Fig. 1). The size of the patch in the ith dimension is $2m_i + 1$, and the region influenced by object s is defined by the indices $\mathcal{P}^{l^s} = \mathcal{P}_1^{l^s} \times \mathcal{P}_2^{l^s} \times \ldots \times \mathcal{P}_d^{l^s}$, where $\mathcal{P}_i^{l^s} = \{l_i^s - m_i, \ldots, l_i^s + m_i\}$. Let $M = |\mathcal{P}^0|$ be the number of elements within a patch. For notational convenience we treat the indices $\mathcal{P}^l$ as well as the regions indexed by them as column vectors ($\mathcal{P}^l \in \mathbb{Z}^{M\times 1}$ and $\mathbf{y}_{\mathcal{P}^l}^t \in \mathbb{R}^{M\times 1}$).

The set of possible basis elements within a patch is $\mathsf{B} = \{\mathbf{b}^k \in \mathbb{R}^M\}_{k=1}^K$ and the coefficients describing the signal produced by object s in each measurement are represented by $\mathbf{x}^s \in \mathbb{R}^{K\times T}$. Thus, each object and its signal is fully represented by the tuple $O_s = (l^s, \mathbf{x}^s)$, and we will sometimes write $\mathsf{O}_l = \{l^s\}$ and $\mathsf{O}_{\mathbf{x}} = \{\mathbf{x}^s\}$.

The observed data is the sum of the signals generated by the objects, corrupted by additive noise:

$$\forall\, l' \in \mathcal{I}, t \in \{1 \ldots T\} \quad \hat{y}_{l'}^t(\mathsf{O}, \mathsf{B}) = \left(\sum_{s=1}^{S} \sum_{k=1}^{K} b_{l'-l^s}^k x_k^{s,t} \right)$$

$$\mathbf{y}^t = \hat{\mathbf{y}}^t(\mathsf{O}, \mathsf{B}) + \boldsymbol{\epsilon} \tag{2}$$

where $\boldsymbol{\epsilon} \sim D$ is a sample from noise distribution D and indexing outside the size of the basis function returns zero: that is, $b_l^k = 0$ if for any j, $|l_j| > m_j$

3. CHOMP

3.1. Motivation

We consider the problem of inference, where the basis functions B are known and we wish to find an estimate $\hat{\mathsf{O}}$ of the generating latent signals. The general sparse decomposition problem for normally distributed additive noise is

$$\hat{\mathsf{O}} = \underset{\mathsf{O}}{\operatorname{argmin}} \sum_{t=1}^{T} \|\mathbf{y}^t - \hat{\mathbf{y}}^t(\mathsf{O}, \mathsf{B})\|^2 + \gamma|\mathsf{O}| \tag{3}$$

where we minimise the data reconstruction error, while incurring a cost for each object added to the set.

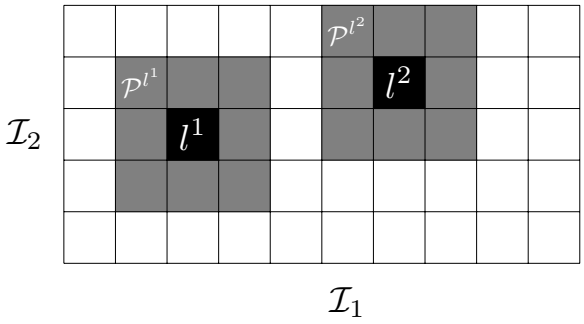


Fig. 1. Object locations and affected patches. An example two-dimensional space with two embedded objects affecting a 3x3 region around their locations

The interdependence of the elements of the solution set, l^s and $\mathbf{x}^s$, makes identification of the true optimum in Eq. 3 intractable for large T. One practical approach is to first locate the objects using the mean of the sample frames $\langle\mathbf{y}\rangle = \frac{1}{T}\sum_t \mathbf{y}^t$, and only then reconstruct the time-courses of the signals at those locations:

$$\hat{\mathsf{O}}_l = \underset{\mathsf{O}_l}{\operatorname{argmin}} \min_{\mathsf{O}_{\mathbf{X}}} \|\langle\mathbf{y}\rangle - \hat{\mathbf{y}}(\mathsf{O}_l, \mathsf{O}_{\mathbf{X}}, \mathsf{B})\|^2 + \gamma|\mathsf{O}| \tag{4}$$

$$\hat{\mathsf{O}}_{l,\mathbf{x}^t} = \underset{\mathsf{O}_{\mathbf{x}^t}}{\operatorname{argmin}} \|\mathbf{y}_{\mathcal{P}^l}^t - \hat{\mathbf{y}}_{\mathcal{P}^l}^t(l, \mathsf{O}_{\mathbf{x}^t}, \mathsf{B})\|^2 \tag{5}$$

where $\mathbf{X}$ represents the coefficients needed to reconstruct the mean data vector. Note that $\mathbf{X}^s \in \mathbb{R}^K$ whereas the original coefficients $\mathbf{x}^s \in \mathbb{R}^{K\times T}$, making the optimum of Eq. 4 far easier to find.

However, in many contexts the mean (or DC) signals associated with each object may be too weak to support robust localisation. Thus, we propose and formulate a new approach called *Convolutional Higher Order Matching Pursuit* (CHOMP), which locates objects using higher-order summary statistics of the data. CHOMP finds an efficient compromise between the intractable simultaneous decomposition of all samples (Eq. 3) and the potential paucity of surviving signal in the mean (Eq. 4), providing a mechanism to trade off the full exploitation of the available data, against the demands for computational and storage resources.

3.2. Cost function

CHOMP extends the domain of signal pursuit from the mean alone to all empirical cumulant tensors of orders $r \in \{1 \ldots R\}$ estimated using unbiased, minimum-variance, multivariate K-statistics [4][1].

Define the unnormalised non-central vector moments of the data $\mathbf{S}_r = \sum_{t=1}^T (\mathbf{y}^t)^{\otimes r}$, where $(\cdot)^{\otimes r}$ is the rth generalised (tensor) outer product $(\cdot) \otimes (\cdot) \otimes \ldots \otimes (\cdot)$. Then the

[1] In principle, it is possible to use any multilinear function of the data patches; we use the cumulant tensors as they provide convenient interpretation and additivity, which simplifies matching pursuit

first three K-statistics are (see [5] for higher-order expressions, and [6] for a general discussion):

$$\mathbf{Y}^{(1)} = \frac{\mathbf{S}_1}{T}$$
$$\mathbf{Y}^{(2)} = \frac{T\mathbf{S}_2 - \mathbf{S}_1^{\otimes 2}}{T(T-1)}$$
$$\mathbf{Y}^{(3)} = \frac{T^2\mathbf{S}_3 - 3T\mathbf{S}_2 \otimes \mathbf{S}_1 + 2\mathbf{S}_1^{\otimes 3}}{T(T-1)(T-2)} \quad (6)$$

From these K-statistics of the entire data, we extract the rth patch-cumulant around location l, $\mathbf{Y}^{l\cdot(r)} \in \mathbb{R}^{M^r}$, by selecting entries $\mathbf{Y}^{(r)}_{\mathbf{l}}$ for which the index r-tuples $\mathbf{l} = [l_1, l_2 \dots l_r]$ fall within $\mathcal{P}^{l\cdot(r)} \equiv \mathcal{P}^l \times \mathcal{P}^l \times \dots \times \mathcal{P}^l$. Note that the cumulants for spatially overlapping patches contain shared elements, and that only those elements of $\mathbf{Y}^{(r)}$ that fall within some $\mathcal{P}^{l\cdot(r)}$ need be computed, reducing computational and storage demands. Using these multilinear features of the data we can define a general cost function $\mathcal{C}$ that represents a discrepancy measure f between the original and the reconstructed feature cumulants $\mathbf{Y}^{(r)}$ and $\widehat{\mathbf{Y}}^{(r)}$. For the squared-difference discrepancy, weighting cumulant orders using hyperparameters $\{\sigma_r^2\}_{r=1}^R$:

$$\hat{\mathsf{O}}_l = \operatorname*{argmin}_{\mathsf{O}_l} \min_{\mathsf{O}_\mathbf{X}} \mathcal{C} \quad (7)$$

$$\mathcal{C} = \sum_{r=1}^{R} \frac{1}{\sigma_r^2} f(\mathbf{Y}^{(r)}, \widehat{\mathbf{Y}}^{(r)}(\mathsf{O}_l, \mathsf{O}_\mathbf{X}, \mathsf{B})) + \gamma|\mathsf{O}| \quad (8)$$

$$f(\cdot,\cdot) = \sum_{\mathbf{l} \in \bigcup_{l\in\mathcal{I}} \mathcal{P}^{l\cdot(r)}} \|\mathbf{Y}^{(r)}_{\mathbf{l}} - \widehat{\mathbf{Y}}^{(r)}_{\mathbf{l}}(\mathsf{O}_l, \mathsf{O}_\mathbf{X}, \mathsf{B}))\|^2 \quad (9)$$

$$\widehat{\mathbf{Y}}^{(r)}_{\mathbf{l}'} = \sum_{s\in\mathsf{O}} \sum_{\mathbf{k}\in\{1\dots K\}^r} X^{s,r}_{\mathbf{k}} b^{k_1}_{l'_1-l^s} b^{k_2}_{l'_2-l^s} \dots b^{k_r}_{l'_r-l^s} \quad (10)$$

where the union over $\mathcal{P}^{l\cdot(r)}$ in Eq. 9 ensures that all elements that appear in a patch-cumulant are counted exactly once.

The reconstruction of Eq. 10 resembles that of Eq. 2, but the elements of the $\mathbf{X}^r \in \mathbb{R}^{K^r}$ tensors are now the reconstruction weights for the estimated r-order cumulants in the linear space over the basis functions B. The hyperparameters $\{\sigma_r^2\}_{r=1}^R$ correspond to the expected error variances in the various cumulant reconstructions. These may be estimated by approximating the noise level in the data – using the lower quantiles of the absolute pixel values, which are unlikely to represent signals — and then applying Isserlis' theorem [7][8] to estimate the expected residual in each order, taking into account the reconstruction over the basis function space. Note that for $R = 1$ and taking into account the support of the basis functions $\mathcal{P}^0$, we obtain an equivalent problem to Eq. 4, whereas for large R we represent the full sample distribution more completely and thus approximate the localisation of the original matching pursuit problem (Eq. 3).

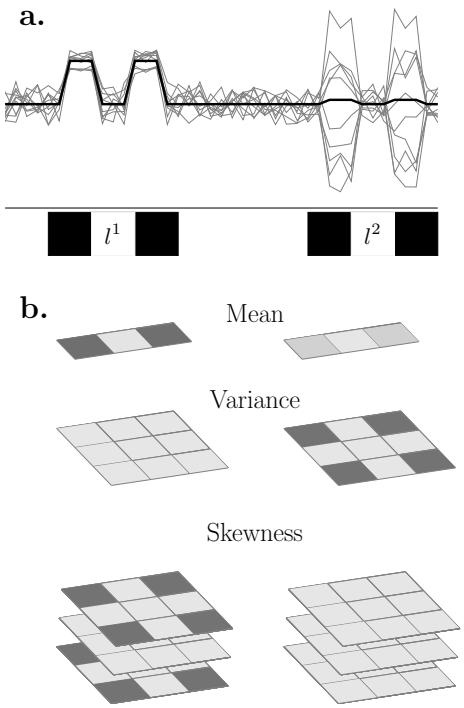


Fig. 2. Feature extraction **a.** Two identical sources in a 1D space, l^1 with high mean and skewness, l^2 with high variance. The thick line is the true mean, the thin ones are individual samples. **b.** The resulting cumulants at the source locations

3.3. Solution

The problem defined by Eqs. 7–10, although easier that the complete temporal decomposition of Eq. 3, is still generally intractable owing to the large search space $\{0,1\}^{|\mathcal{I}|}$. Matching pursuit offers a greedy solution to the problem, and comes with several computational advantages. The object locations are identified one-by-one as follows (Fig. 3):

1. Initialise $\hat{\mathsf{O}} = \{\}$ and $s = 1$.
2. For all locations $l \in \mathcal{I}$ compute the change in cost that would follow from assuming a new signal centred at l with optimal signal parameters $\mathbf{X}$.
3. Set l^s to the source location with the largest decrease in cost, add it to the solution set $\hat{\mathsf{O}}_l$, and store the corresponding optimal signal parameters $\mathbf{X}^s$.
4. Update the cost changes associated with each l so that they reflect incremental changes from the solution that now includes l^s. This update affects only those locations for which the associated patch overlaps with $\mathcal{P}^{l^s}$. Let $s = s + 1$ and return to step 3,
5. If no change decreases the cost, accept the locations $\hat{\mathsf{O}}_l$.
6. For all s and t, solve for $\mathbf{x}^{s,t}$ (Eq. 5), and add to the final source solution set $\hat{\mathsf{O}}$.

The core computations of steps 2 and 4 are elaborated below.

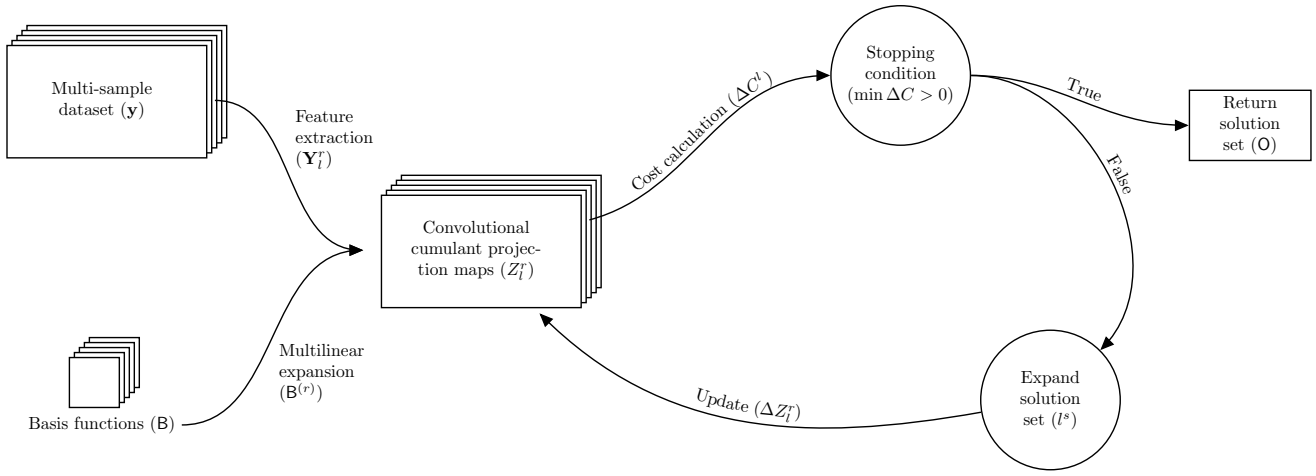


Fig. 3. Algorithm flowchart. The iterative updates only affect the projection maps locally, so we only need to compute the expensive convolutional projection once.

3.3.1. Initial computation

The decrease in cost (Eqs. 8 and 9) obtained by assuming an object at a location l depends on the value of the signal parameters $\mathbf{X}_l^r$, and so pursuit of the greatest decrease requires that we compute the optimal parameter $\mathbf{X}_l^r$ at each location. This is achieved by least-squares regression[2] from all possible r-th order tensor combinations of basis functions to the r-th order cumulant at location l:

$$\begin{aligned} \hat{\mathbf{X}}_l^r &= \operatorname*{argmin}_{\mathbf{X}} \|\mathbf{Y}^{l\cdot(r)} - \mathsf{B}^{(r)}\mathbf{X}\|^2 \\ \hat{\mathbf{X}}_l^r &= (\mathsf{B}^{(r)\top}\mathsf{B}^{(r)})^{-1} Z_l^r \\ Z_l^r &= \mathsf{B}^{(r)\top}\mathbf{Y}^{l\cdot(r)} \end{aligned} \tag{11}$$

where $\mathsf{B}^{(r)} \in \mathbb{R}^{M^r \times K^r}$ is a matrix that contains as columns all vectorised $\mathbf{b}^{k_1} \otimes \ldots \otimes \mathbf{b}^{k_r}$, $k_i \in \{1 \ldots K\}$ and $Z_l^r \in \mathbb{R}^{K^r}$ denotes the vectorised projection of the r-th order patch cumulant tensor onto each of the basis function tensors. Then the change in cost associated with an object located at l is:

$$\begin{aligned} \Delta C^l &= \sum_{r=1}^{R} \frac{1}{\sigma_r^2}(\|\mathbf{Y}^{l\cdot(r)} - \mathsf{B}^{(r)}\hat{\mathbf{X}}_l^r\|^2 - \|\mathbf{Y}^{l\cdot(r)}\|^2) + \gamma \\ &= \sum_{r=1}^{R} \frac{1}{\sigma_r^2}(2\hat{\mathbf{X}}_l^{r\top} Z_l^r - \hat{\mathbf{X}}_l^{r\top}\mathsf{B}^{(r)\top}\mathsf{B}^{(r)}\hat{\mathbf{X}}_l^r) + \gamma \\ &= \sum_{r=1}^{R} \frac{1}{\sigma_r^2} Z_l^{r\top}(\mathsf{B}^{(r)\top}\mathsf{B}^{(r)})^{-1} Z_l^r + \gamma \end{aligned} \tag{12}$$

We proceed in a greedy fashion choosing $l^1 = \operatorname{argmin}_l \Delta C^l$ to be the first element in the solution set $\hat{\mathsf{O}}_l$ as in step 3.

[2]If $K^r > M^r$ the problem is overdetermined; it is not meaningful for the local feature dictionary to be overcomplete as there is no sparsity penalty on local features, only on object locations.

3.3.2. Update

At step 4, the changes in cost associated with each possible object location must be revised to take into account the addition of a new source $\{l^s, \hat{\mathbf{X}}_{l^s}\}$ to the current solution set. Given an established set of sources, the incremental change in cost associated with a new assumed source depends on its ability to fit the *residual* from the current model. Conceptually, the update thus requires the subtraction of the contribution of the reconstructed object ($\mathsf{B}^{(r)}\hat{\mathbf{X}}_{l^s}^r$) from the data, and then the computation of $\hat{\mathbf{X}}_l$ and thus ΔC^l for any location l such that $\mathcal{P}^{l\cdot(r)}$ overlaps $\mathcal{P}^{l^s\cdot(r)}$. However, as the data only appear in Eq. 12 through the basis projections Z_l^r, it is possible to compute the change in projection at l based on the new source $\{l^s, \hat{\mathbf{X}}_{l^s}\}$ implicitly:

$$\begin{aligned} \Delta_s Z_l^r &= -\mathbf{G}_{l-l^s}^r \hat{\mathbf{X}}_{l^s}^r \\ \mathbf{G}_{l-l^s}^r &= \mathsf{B}^{(r)\top}\mathcal{S}_{l-l^s}^r \mathsf{B}^{(r)} \end{aligned}$$

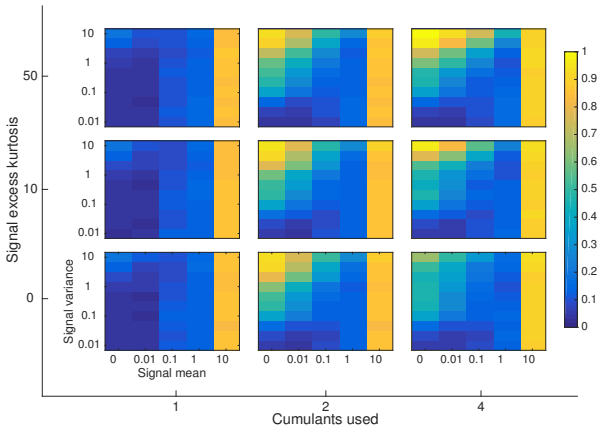
where $\mathcal{S}_{l-l^s}^r \in \{0,1\}^{M^r \times M^r}$ is a sparse binary shift tensor indicating which elements of the respective r-th order tensors interact for shifts of $l - l^s$ in the original space[3]. The explicit forms of the updates for projections of the first two cumulant orders at location l are:

$$\Delta_s Z_{l,k}^1 = \sum_{l'} b_{l'-l}^k \left(\sum_{k'} b_{l'-l^s}^{k'} \hat{\mathbf{X}}_{l^s,k'}^1 \right)$$

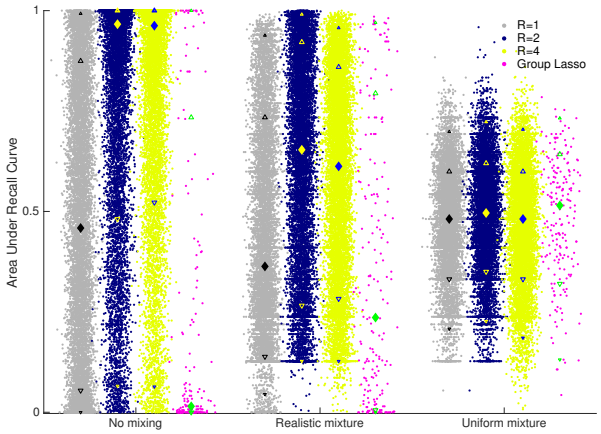
$$\Delta_s Z_{l,k_1k_2}^2 = \sum_{l'_1,l'_2} b_{l'_1-l}^{k_1} b_{l'_2-l}^{k_2} \left(\sum_{k'_1,k'_2} b_{l'_1-l^s}^{k'_1} b_{l'_2-l^s}^{k'_2} \hat{\mathbf{X}}_{l^s,k'_1k'_2}^2 \right)$$

Once these updates have been computed, the projections,

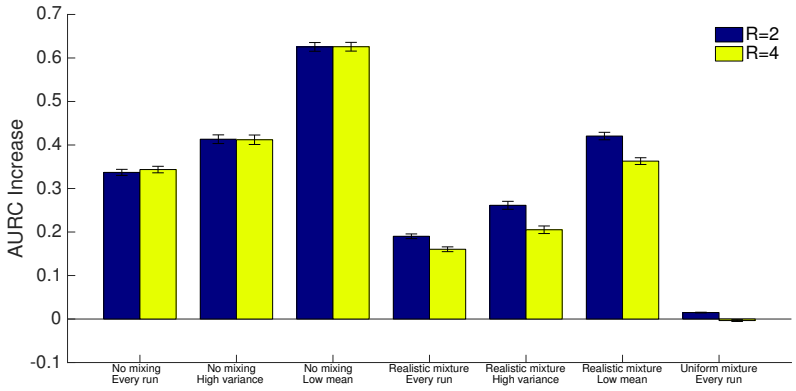
[3]Note that if the original space is $d > 1$ dimensional then the l indices are themselves d-dimensional vectors. In our formulation, however, this extra complication only enters into the structure of $\mathcal{S}$ during the algorithm.



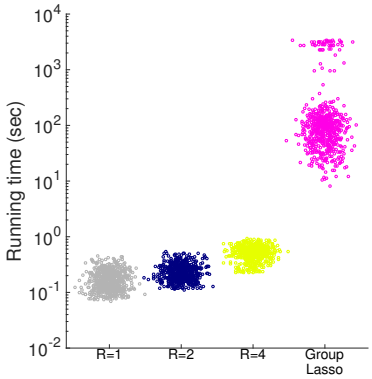
(a) Localisation performance for varying signal parameters



(b) Performance levels of source signal distribution mixtures within a single field of view



(c) The use of higher order features substantially increases localisation performance



(d) Substantial gain in running time compared to matrix factorisation

Fig. 4. **a.** CHOMP incorporating higher order cumulants offers substantial gain in localisation performance when the corresponding structure is present in the signal distribution. Each cell shows the fraction of sources correctly localised in $n = 10000$ runs for different signal parameters (exact parameters were distributed normally with a standard deviation of one order of magnitude around the specified value). **b.** Localisation performance expressed as AURC for all runs ($n = 3 \times 10000 + 600$) with varying mixtures of sources within a single run. *No mixing*: All sources share the same signal distribution. *Realistic mixture*: as in (**a**). *Uniform mixture*: Signal distributions may vary up to 4 orders of magnitude within the same field of view. Means, 1σ and 2σ quantiles indicated. **c.** Assessing the improvement within a single field of view gained by incorporating higher order cumulants. Bars are the mean gains in AURC over the first order method, runs selected by signal distribution criterion ($n <= 10000$). *High variance*$>= 1$, *Low mean*$<= 0.1$. Error bars are SEM. **d.** Comparison of running times ($n = 4 \times 600$).

optimal signal parameters, and cost increments are straightforward to update for the next iteration:

$$
\begin{aligned}
Z_l^r &\leftarrow Z_l^r + \Delta_s Z_l^r \\
\hat{\mathbf{X}}_l^r &\leftarrow (\mathsf{B}^{(r)\top}\mathsf{B}^{(r)})^{-1} Z_l^r \\
\Delta C^l &\leftarrow \sum_{r=1}^{R} \frac{1}{\sigma_r^2} Z_l^{r\top} (\mathsf{B}^{(r)\top}\mathsf{B}^{(r)})^{-1} Z_l^r + \gamma \qquad (13)
\end{aligned}
$$

If $\forall l: \ \Delta C^l > 0$, inference is complete. If not, we proceed to identify the next source.

4. EVALUATION

We evaluated the impact of incorporating higher order cumulants in pursuit under a broad range of signal distributions. Data were simulated from the generative model (Eq. 2) in a one dimensional space ($I = 512$), with a source density of 0.05. Each simulation was based on a new random set of basis functions ($M = 11$, $K = 2$). Signal distributions were modelled as mixtures of Gaussians, with parameters selected by non-linear least-squares to match an intended set of moments. We explored symmetric distributions with means

and variances spanning multiple orders of magnitudes and a number of kurtosis values. Coefficients $x^s_{k,t}$ were sampled iid ($T = 1000$), combined with the basis functions, and the products summed along with zero-mean, unit-variance, additive Gaussian noise to yield the generated data.

Localisation was carried out as described (Fig. 3), with a stopping condition corresponding to the true source density. The values of σ were set as described above, resulting in $\sigma_r^2 \approx (M^r - K^r) * (\sigma^2_{\text{noise}}/T)$ for large M and T, by the central limit theorem. For comparability we used the true value for $\sigma^2_{\text{noise}} = 1$, instead of estimating it from the data.

A natural evaluation metric is the frequency with which the algorithm correctly locates the sources, applied using increasing orders of cumulants. We find that as long as the signal distributions contain significant higher order structure, it is indeed feasible to attempt to reconstruct those tensors, in spite of the vast dimensionality increase involved (Fig. 4(a)). A further feature of greedy algorithms in general, including the current one, is that they provide a natural ordering of the sources found. We thus define the Area Under the Recall Curve as

$$\text{AURC}(\mathsf{O}, \hat{\mathsf{O}}) = \frac{1}{S}\sum_{s=1}^{S} \frac{\text{NumCorrect}[\hat{\mathsf{O}}_{1:s}] - \text{Chance}_s}{s - \text{Chance}_s}$$

$$\text{Chance}_s = \frac{1}{\binom{|\mathcal{I}|}{s}} \sum_{s'=1}^{s} s' \binom{s}{s'} \binom{|\mathcal{I}| - s}{s - s'}$$

and estimate the performance using this metric. This is of interest especially in the case when the field of view contains signals from multiple distributions (Fig. 4(b)). Finally we looked at how much higher order features offer on a case-by-case basis (Fig. 4(c)) and we found that for all practical cases, higher order estimators are substantially beneficial[4].

We compared[5] our proposed method to a group lasso implementation [9], corresponding to the L_1-relaxation of Eq. 3, for which we provided the $I \times KI$ spatial design matrix and grouped time courses belonging to the same locations. Estimated time courses were sorted by their norms to obtain $\hat{\mathsf{O}}$ and the AURC evaluated as above. CHOMP outperformed the group lasso Fig. 4(b) while being over two orders of magnitude faster Fig. 4(d).

5. DISCUSSION

We have described a flexible generalisation of convolutional matching pursuit, that extends the classical formulation using a multilinear expansion of the basis functions to reconstruct higher order features of data distributions. *Convolutional Higher Order Matching Pursuit* (CHOMP) is able to efficiently use multiple samples of the same region to pinpoint the location of generating sources. As each and every step in the algorithm is fully linear and additive, it is easy to extend to arbitrary input dimensionalities and numbers of basis functions, and lends itself to parallel implementation.

We evaluated the localisation performance of the algorithm either using just the mean, or including higher order features. We found that the latter provided substantial improvements, even under challenging circumstances where signal distributions may differ by multiple orders of magnitude.

The general idea of reconstructing higher order tensors of mixed signals of interacting sources may well be applicable outside the multi-sample localisation problem solved here, with the proviso that the nature of the interaction itself may be unknown, whereas in localisation it is well described as a function of spatial location.

[4] Uniform mixture performance can also be improved by up to 0.3 AURC gain by using an adaptive scheme for setting $\{\sigma_r^2\}_{r=1}^R$

[5] Note that we could evaluate the group lasso method only on smaller sample of $n = 600$ due to prohibitively slow running times.

6. REFERENCES

[1] SG Mallat and Z Zhang, "Matcing pursuite with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 11, pp. 3397–3415, 1993.

[2] A Szlam, K Kavukcuoglu, and Y LeCun, "Convolutional Matching Pursuit and Dictionary Training," *Faces*, pp. 1–7, 2010.

[3] M Pachitariu, AM Packer, N Pettit, H Dalgleish, M Hausser, and M Sahani, "Extracting regions of interest from biological images with convolutional sparse block coding," in *Adv. Neural Inf. Process. Syst.*, 2013, vol. 1, pp. 1745–1753.

[4] PR Halmos, "The Theory of Unbiased Estimation," *Ann. Math. Stat.*, vol. 17, no. 1, pp. 34–43, 1946.

[5] A Stuart and JK Ord, *Kendall's advanced theory of statistics*, vol. 1, 1987.

[6] E Di Nardo, G Guarino, and D Senato, "A unifying framework for k -statistics, polykays and their multivariate generalizations," *Bernoulli*, vol. 14, no. 2, pp. 440–468, 2008.

[7] L Isserlis, "On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables," *Biometrika*, vol. 12, no. 1-2, pp. 134–139, 1918.

[8] C. Vignat, "A generalized Isserlis theorem for location mixtures of Gaussian random vectors," *Stat. Probab. Lett.*, vol. 82, no. 1, pp. 67–71, 2011.

[9] S Boyd, N Parikh, E Chu, B Peleato, and J Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.