Bayesian sparse optimization for large scale dictionary learning

Daniela Calvetti

Case Western Reserve University

Department of Mathematics, Applied Mathematics and Statistics

Based on joint work with Erkki Somersalo, Nathan Waniorek and Alberto Bocchinfuso.

Supported in part by NSF DMS-1951446

AWM Workshop: Women in Inverse Problems, June 1, 2023

Dictionary matching formulation

Consider the system:

$$\mathrm{D}x = b\,, \quad \mathrm{D} \in \mathbb{R}^{m\times n}\,,\, b \in \mathbb{R}^m\,,$$

where

- D is a dictionary of n atoms.
- b is the datum to represent .
- x_j is the weight of the contribution of j-th atom to b.

Sparse coding

Assume b admits a sparse dictionary representation, and we want to find $\hat{x}$

$$\hat{x} = \arg\min_x \|\mathrm{D}x - b\| , \quad \|x\|_0 \ll n$$

To reduce complexity, replace D with a low rank factorization

$$\mathrm{D} = \mathrm{W}\mathrm{H} + \mathrm{E}, \quad \mathrm{W} \in \mathbb{R}^{m\times k}, \ \mathrm{H} \in \mathbb{R}^{k\times n}, \ k \ll n,$$

then solve

$$b = \mathrm{W}h + \epsilon', \quad h \geq 0,$$

where ϵ' accounts for all discrepancy contributions.

A multiphase approach

To reduce complexity and promote sparsity we suggest a four step approach.

1. **Clustering**: group original atoms into subdictionaries $\mathrm{D}^{(j)}$, $1 \leq j \leq K$.
2. **Reduction**: compress each subdictionary into low rank code

$$\mathrm{D}^{(j)} = \mathrm{W}^{(j)}\mathrm{H}^{(j)} + \mathrm{E}^{(j)}.$$

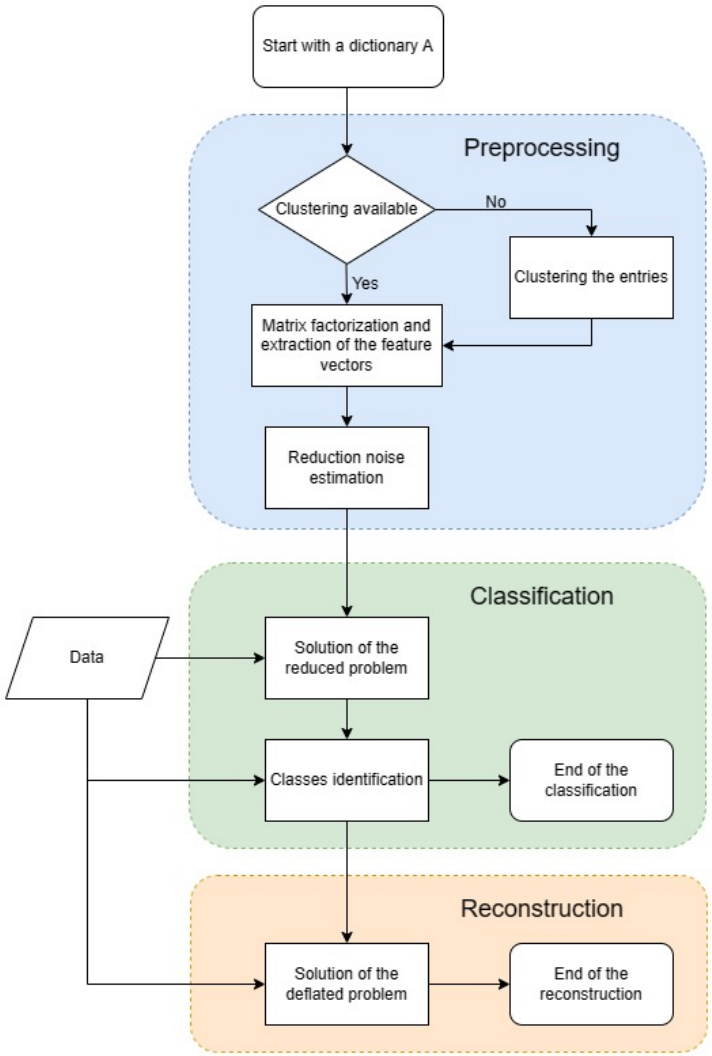
3. **Selection**: Identify clusters needed to explain b,

$$b = \mathrm{W}^{(j_1)}h^{(j_1)} + \ldots + \mathrm{W}^{(j_r)}h^{(j_r)} + \epsilon'.$$

4. **Deflate and Solve**: Express b in terms of identified subdictionaries,

$$b = \mathrm{D}^{(j_1)}x^{(j_1)} + \ldots + \mathrm{D}^{(j_r)}x^{(j_r)} + \tilde{\epsilon}'.$$

Flowchart summary



Clustering

- Applied unsupervised clustering to dictionary atoms, to exploit latent structure of data.
- Use k-medoids.
- Regroup atoms into block-wise dictionaries.

$$\mathrm{D} = \left[\; \mathrm{D}^{(1)} \; \middle| \; \ldots \; \middle| \; \mathrm{D}^{(k)} \; \right]$$

Reduction

Compute low-rank approximation of each subdictionary

$$D^{(j)} = W^{(j)} H^{(j)} + E^{(j)}, \quad W^{(j)} \in \mathbb{R}^{m \times k_j}, \; H^{(j)} \in \mathbb{R}^{k_j \times n}, \; k_j < n.$$

where

- $H^{(j)} \geq 0$;
- Feature vector = columns of $W^{(j)}$, retain characteristics of atoms in $D^{(j)}$.

Reduction

To find W, H, proceed as in NMF, relaxing nonnegativity for W.

- Given W, columns of H satisfy

$$h^{\ell} = \arg\min_{h} \|d^{(\ell)} - \mathrm{W}h\|_2$$

- Given H, W satisfies

$$\begin{aligned} \mathrm{W} &= \arg\min_{\mathrm{V}} \|\mathrm{D} - \mathrm{V}\mathrm{H}\|_{\mathrm{F}} \\ &= \arg\min_{\mathrm{V}} \|\mathrm{D}^{\mathsf{T}} - \mathrm{H}^{\mathsf{T}}\mathrm{V}^{\mathsf{T}}\|_{\mathrm{F}} \end{aligned}$$

NNLS- NMF algorithm

Given: data matrix $\mathrm{D} \in \mathbb{R}^{m \times n}$, $\mathrm{X} \geq 0$, rank $k > 0$,

Initialize: $\mathrm{W}^0 \in \mathbb{R}^{m \times k}$, $t = 0$.

Iterate until stopping criterion is met:

1. Update H: Set $h^{(j)} = \arg\min_{h \geq 0} \|d^{(j)} - \mathrm{W}^{(t)} h\|_2$ for $1 \leq j \leq n$, and set

$$\mathrm{H}^{(t+1)} = \left[\begin{array}{ccc} h^{(1)} & \dots & h^{(n)} \end{array} \right];$$

2. Scale the rows of $\mathrm{H}^{(t+1)}$,

$$h_{(j)} \to \frac{h_{(j)}}{\|h_{(j)}\|_1}, \quad 1 \leq j \leq k.$$

3. Update W: Set $w_{(j)} = \arg\min_{w \geq 0} \|\left(\mathrm{H}^{(t+1)}\right)^{\mathsf{T}} w - d_{(j)}\|_2$ for $1 \leq j \leq m$, and set

$$\mathrm{W}^{(t+1)} = \left[\begin{array}{c} w_{(1)}^{\mathsf{T}} \\ \vdots \\ w_{(m)}^{\mathsf{T}} \end{array} \right];$$

4. Check the convergence criterion, advance the counter $t \to t + 1$.

Dictionary Reduction

Approximate the dictionary matrix D by:

$$\left[\ \mathrm{D}^{(1)}\ \middle|\ \ldots\ \middle|\ \mathrm{D}^{(k)}\ \right] \approx \left[\ \mathrm{W}^{(1)}\ \middle|\ \ldots\ \middle|\ \mathrm{W}^{(k)}\ \right] \begin{bmatrix} \mathrm{H}^{(1)} & 0 & \ldots & 0 \\ 0 & \mathrm{H}^{(2)} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & \mathrm{H}^{(k)} \end{bmatrix}$$

$$= \left[\ \mathrm{W}^{(1)}\mathrm{H}^{(1)}\ \middle|\ \ldots\ \middle|\ \mathrm{W}^{(k)}\mathrm{H}^{(k)}\ \right] + \mathrm{E}$$

then

$$b = \mathrm{W}\underbrace{\mathrm{H}z}_{=h} = \mathrm{W}\begin{bmatrix} \mathrm{H}^{(1)}z^{(1)} \\ \vdots \\ \mathrm{H}^{(k)}z^{(k)} \end{bmatrix} + \epsilon.$$

Dictionary Compression Error

Want to analyze

$$\mathsf{E} = \mathsf{D} - \mathsf{W}\mathsf{H}.$$

Observe that

$$\begin{aligned} d^{(\ell)} = \mathsf{D}e_\ell &= \mathsf{W}h^{(\ell)} + (\mathsf{D}e_\ell - \mathsf{W}h^{(\ell)}) \\ &= \mathsf{W}h^{(\ell)} + m^{(\ell)}, \quad 1 \le \ell \le n. \end{aligned}$$

where $m^{(\ell)}$ is the dictionary compression error (DCE). In Bayesian framework,

model $m^{(\ell)}$ as a random variable. To estimate the probability density of DCE

- Regard atom $d^{(\ell)}$ as a realization from underlying distribution π_d
- π_{DCE} is the push forward of π_d by mapping $d^{(\ell)} \to m^{(\ell)}$.
- Note that $h^{(\ell)} = \arg\min_h \|d^{(\ell)} - \mathsf{W}h\|_2^2$.
- It is not clear that discrepancies $m^{(\ell)}$ are scaled white noise.

Model discrepancy as noise

Dictionary Compression Error

$$m = b - \mathsf{W}\mathsf{H}x$$

To estimate m,

1. Draw s uniformly distributed random realizations, $\tilde{x}^1, \ldots, \tilde{x}^m$, of x,
2. Compute $\mu_{\mathrm{DCE}} = \frac{1}{m}\sum_{j=1}^{m}\left(b - \mathsf{W}\mathsf{H}\tilde{x}^j\right)$.
3. Compute $\mathsf{C}_{\mathrm{DCE}} = \frac{1}{m}\sum_{j=1}^{m}(b - \mathsf{W}\mathsf{H}\tilde{x}^j - \mu)(b - \mathsf{W}\mathsf{H}\tilde{x}^j - \mu)^{\mathsf{T}} + \epsilon\mathsf{I}$.
4. Use the Laplace approximation $\epsilon \sim \mathcal{N}(\mu_{\mathrm{DCE}}, \mathsf{C}_{\mathrm{DCE}})$.

Whitening

$$\mathsf{C}_{\mathrm{DCE}}^{-1/2}\left(b - \mu_{\mathrm{DCE}}\right) = \mathsf{C}_{\mathrm{DCE}}^{-1/2}\mathsf{W}h + w$$

where w is white noise.

Cluster detection

Identify subdictionaries contributing to b.
If only few subdictionaries contribute, solve

$$h = \arg\min \left\{ \left\| \mathsf{C}_{\mathrm{DCE}}^{-1/2}(b - \mu_{\mathrm{DCE}} - \mathsf{W}h) \right\|^2 + \lambda \left\| \left(\|h^{(1)}\|_{(1)}, \ldots, \|h^{(K)}\|_{(K)} \right) \right\|_0 \right\},$$

where the norms $\| \cdot \|_{(j)}$, are constructed by using the full dictionary information.

Structural prior for subcluster coefficients

- The $h^{(j)}$ is expected to have a structural similarity to the column vectors of the corresponding matrices $\mathrm{H}^{(j)}$ arising from the full dictionary $\mathrm{D}^{(j)}$.
- $h^{(j)}$ is not expected to be close to the group mean, as most groups $\mathrm{W}^{(j)}$ contribute little or not at all to explain b.
- $h^{(j)}$ should in the general direction of $\mathrm{H}^{(j)}$, with amplitude controlled by a group sparsity promoting prior.

Structural prior

Let

$$\mathsf{H}^{(j)} = \mathsf{U}^{(j)}\Sigma^{(j)}\left[\mathsf{V}^{(j)}\right]^{\mathsf{T}}, \quad \Sigma^{(j)} = \mathrm{diag}\left([\sigma_1^{(j)}, \ldots, \sigma_{r_j}^{(j)}, 0, \ldots, 0]\right),$$

and define

$$\begin{aligned} \mathsf{C}^{(j)} &= \left(\frac{1}{\sigma_1^{(j)}}\right)^2 \mathsf{U}^{(j)}\Sigma^{(j)}\left[\Sigma^{(j)}\right]^{\mathsf{T}}\left[\mathsf{U}^{(j)}\right]^{\mathsf{T}} + \epsilon\,\mathsf{I} \\ &= u_1^{(j)}\left[u_1^{(j)}\right]^{\mathsf{T}} + \sum_{\ell=2}^{r_j}\left(\frac{\sigma_\ell^{(j)}}{\sigma_1^{(j)}}\right)^2 u_\ell^{(j)}\left[u_\ell^{(j)}\right]^{\mathsf{T}} + \epsilon\,\mathsf{I}, \end{aligned}$$

The *direction vector* $\widehat{h}^{(j)}$ of $h^{(j)}$, follows the angular central Gaussian measure with covariance $\mathsf{C}^{(j)}$, defined as the pullback of the Gaussian measure $\mathscr{N}(0, \mathsf{C}^{(j)})$ with respect to the radial projection $\mathbb{R}^{k_j} \to \mathbb{S}^{k_j-1}$.

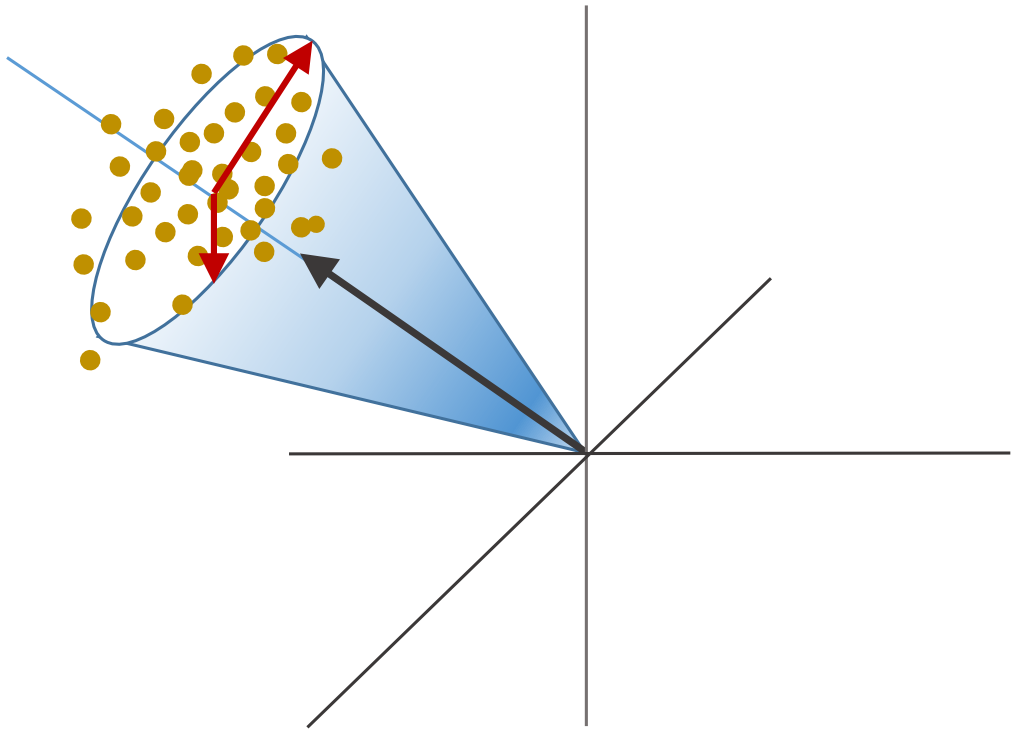
Alternatively, we may express this prior in terms of the conditional distribution, given the length of the vector, as

$$\pi^{(j)}(\widehat{h}^{(j)} \mid \|h^{(j)}\|) \propto \exp\big(-\|h^{(j)}\|^2_{(j)}\big),$$

where the structural norm is defined as

$$\|h^{(j)}\|^2_{(j)} = \big[h^{(j)}\big]^{\mathsf{T}} \big[\mathsf{C}^{(j)}\big]^{-1} h^{(j)}. \tag{1}$$

This expresses belief that $h^{(j)}$ could have an arbitrarily large component along $u_1^{(j)}$, but it should remain within the cone defined by the columns of $\mathsf{H}^{(j)}$.



Thresholding

- After computing $h^{(j)}$ with group sparsity, cannot expect most $\|h^{(j)}\|^2_{(j)}$ to be zero.
- We expect many $h^{(j)}$ to be negligible.
- The only significant $h^{(j)}$ are those above std of DCE, that is,

$$h^{(j)} \text{ is significant if and only if } \|\mathsf{W}^{(j)} h^{(j)}\|^2 > \mathrm{trace}\big(\mathsf{C}^{(j)}_{\mathrm{DCE}}\big).$$

- Let $J = [j_1, \ldots, j_r\} \subset \{1, 2, \ldots, K\}$ be the indices of relevant subdictionaries.

Deflation

Once identified the relevant classes:

- Deflate original dictionary D removing non contributing classes and solve

$$\left[\; \mathrm{D}^{(i_1)} \; \middle| \; \ldots \; \middle| \; \mathrm{D}^{(i_\ell)} \; \right] \begin{bmatrix} x^{(i_1)} \\ \vdots \\ x^{(i_\ell)} \end{bmatrix} = \hat{\mathrm{D}}\hat{x} = b \, .$$

- If all the relevant classes have been identified, then

$$\mathrm{D}x = \hat{\mathrm{D}}\hat{x} = b \, .$$

- b admits sparse representation if

$$\|\hat{x}\|_0 \ll n \, .$$

Next we solve the deflated problem

$$\hat{\mathrm{D}}\hat{x} = b.$$

IAS algorithm for sparsity promotion

Given: matrix $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, β, ϑ_j, $1 \le j \le n$,

Initialize: $\theta_j^{(0)} = \vartheta_j$, $1 \le j \le n$, set the counter $k = 0$.

Iterate until stopping criterion is met:

1. Update x: Set $\theta_j = \theta_j^{(k)}$, $1 \le j \le n$ and set

$$x^{(k+1)} = \arg\min \left\{ \|b - Ax\|_2^2 + \|D_{\theta^{(k)}}^{-\frac{1}{2}} x\|_2^2 \right\}.$$

2. Update θ_j, $1 \le j \le n$ by solving

$$\frac{\partial \mathscr{E}}{\partial \theta_j}(x^{(k+1)}, \theta) = 0,$$

yielding

$$\theta_j^{(k+1)} = \frac{\vartheta_j}{2}\left(\eta + \sqrt{\eta^2 + \frac{[x_j^{(k+1)}]^2}{4\vartheta_j}}\right), \quad \eta = \beta - \frac{3}{2}.$$

3. Advance the counter $k \to k+1$ and check convergence.

Properties of IAS

When using gamma distribution

- The functional minimized by the IAS algorithm is convex.
- The IAS iterates converge to the unique minimizer.
- In the limit as $\eta = \beta - 3/2 \to 0^+$, the solution computed with the IAS algorithm converges to a scaled ℓ_1-regularized solution,

$$x_{\ell_1} = \text{argmin} \left\{ \frac{1}{2} \|b - Ax\|^2 + \sqrt{2} \sum_{j=1}^{n} \frac{|x_j|}{\sqrt{\vartheta_j}} \right\},$$

- The value of the parameter β can selected depending on how sparse the solution is believed to be.

IAS with Group sparsity and Structural Prior for Class Detection

Want to find an approximate representation

$$b = \mathrm{W}h + \varepsilon',$$

where

$$\mathrm{W} = \left[\begin{array}{c|c|c} \mathrm{W}^{(1)} & \dots & \mathrm{W}^{(k)} \end{array} \right], \quad h = \begin{bmatrix} h^{(1)} \\ \vdots \\ h^{(k)} \end{bmatrix}, \quad h^{(j)} \in \mathbb{R}^{n_j},$$

with the prior belief that most of the vectors $h^{(j)}$ are insignificant.

Structural Prior

Recalling structural covariance matrices $\mathsf{C}^{(j)} \in \mathbb{R}^{n_j \times n_j}$, we define a group sparsity promoting structural prior for h,

$$\begin{aligned}
\pi_{H|\Theta}(h \mid \theta) &\propto \prod_{j=1}^{K} \left(\frac{1}{\det(\theta_j \mathsf{C}^{(j)})} \right)^{1/2} \exp\left(-\sum_{j=1}^{K} \frac{[h^{(j)}]^{\mathsf{T}} [\mathsf{C}^{(j)}]^{-1} h^{(j)}}{2\theta_j} \right) \\
&\propto \exp\left(-\frac{1}{2} \sum_{j=1}^{K} n_j \log \theta_j - \frac{1}{2} \sum_{j=1}^{K} \frac{\|h^{(j)}\|^2_{(j)}}{\theta_j} \right) \\
&\propto \exp\left(-\frac{1}{2} \sum_{j=1}^{K} n_j \log \theta_j - \frac{1}{2} \|\mathsf{D}_\theta^{-1/2} h\|_2^2 \right),
\end{aligned}$$

where D_θ is the block diagonal matrix

$$\mathsf{D}_\theta = \begin{bmatrix} \theta_1 \mathsf{C}_1 & 0 & \dots & 0 \\ 0 & \theta_2 \mathsf{C}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \theta_k \mathsf{C}_k \end{bmatrix}.$$

IAS with Structural Prior and Group Sparsity

- To promote group sparsity, most θ_j should be close to zero, with a few large outliers.
- Assume that the variables θ_j are independent and distributed according to a gamma distribution.
- Combining the likelihood, enhanced by the dictionary compression error, we obtain a posterior density of the form

$$\pi(h, \theta \mid b) \propto \exp\Big(-\frac{1}{2}\|\mathsf{C}_{\mathrm{DCE}}^{-12}(b - \mu_{\mathrm{DCE}} - \mathsf{W}h)\|^2 - \frac{1}{2}\sum_{j=1}^{k}\frac{\|h^{(j)}\|_{(j)}^2}{\theta_j}$$

$$+2\sum_{j=1}^{k}\eta_j \log\theta_j - \sum_{j=1}^{K}\frac{\theta_j}{\vartheta_j}\Big),$$

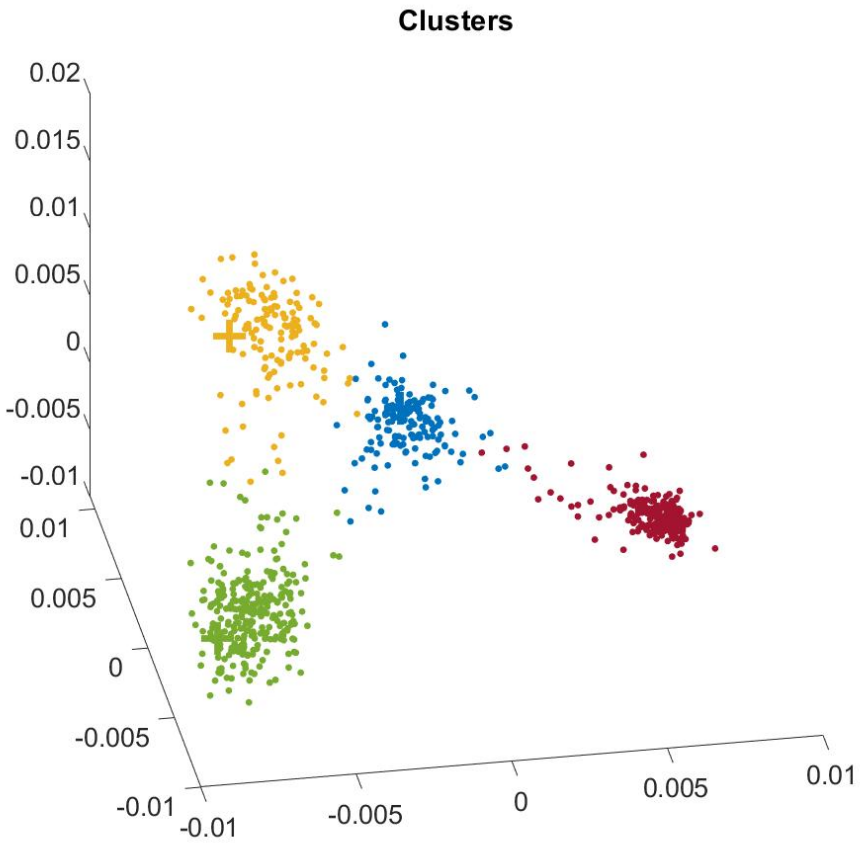
$$\eta_j = \beta - \frac{n_j + 2}{2}, \quad 1 \le j \le K.$$

Compute MAP with Group Sparsity IAS (GS-IAS), a version of the IAS algorithm that promotes group sparsity.

Computed example

The dictionary A consists of $n = 834$ 16×16 digits $0, 1, 3, 9$ from the MNIST database, clustered into $k = 4$ classes using the K-medoid algorithm.

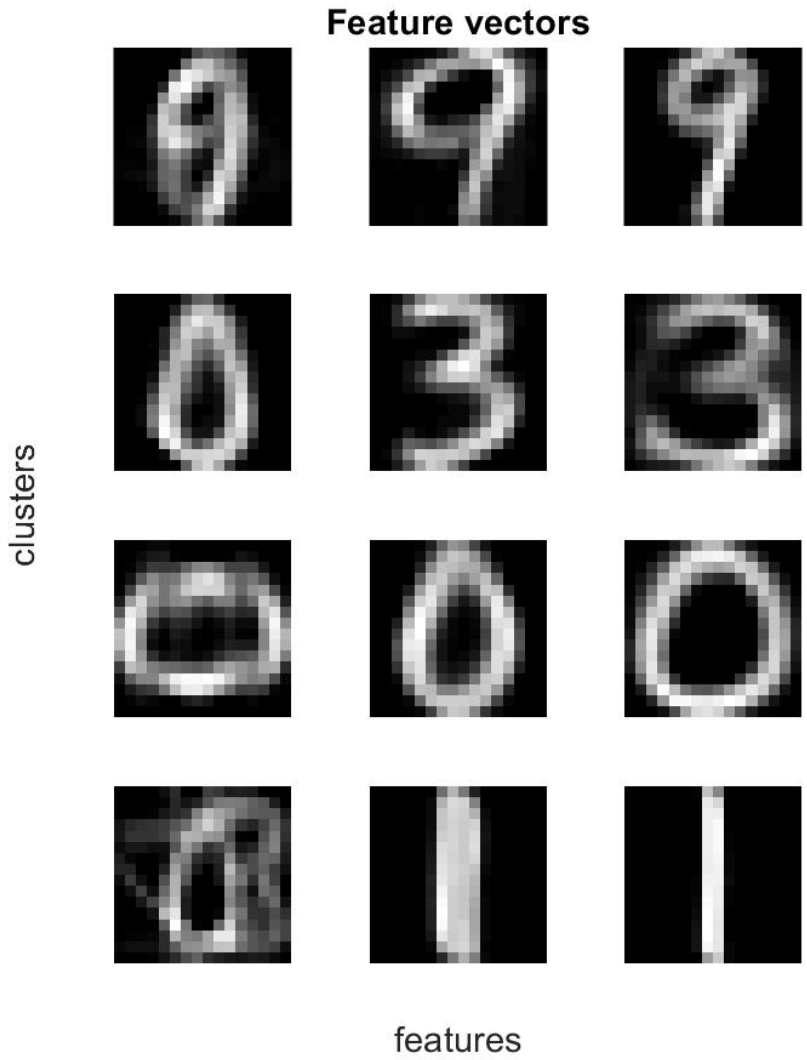
In each test b is a convex combination of $1 \leq l \leq 3$ atoms, with at most one atom per cluster.



Feature extraction and dictionary reduction

Extract $p = 3$ feature vectors per cluster and assemble the reduced dictionary $W \in \mathbb{R}^{256\times 12}$.

The feature vectors capture the features of two different digits in certain classes where the clustering is not perfect.



Evaluation: residual and support retrieval

- 4000 tests problems, data convex combinations of 1 2, or 3 digits from different classes.
- Define the thresholded supports of the vectors,

$$S_\delta = \{j \mid x_j > \delta\}, \quad \widehat{S}_\delta = \{j \mid \widehat{x}_j > \delta\}, \quad \delta > 0.$$

- Define the dissimilarity index as the cardinality of the symmetric difference of them,

$$I_\delta(x, \widehat{x}) = \mathrm{card}(S_\delta \triangle \widehat{S}_\delta) = \mathrm{card}((S_\delta \cup S_\delta) \setminus (S_\delta \cap S_\delta)\ ).$$

- When $I_\delta(x, \widehat{x}) = 0$, the algorithm has correctly identified the atoms that constitute the data.
-

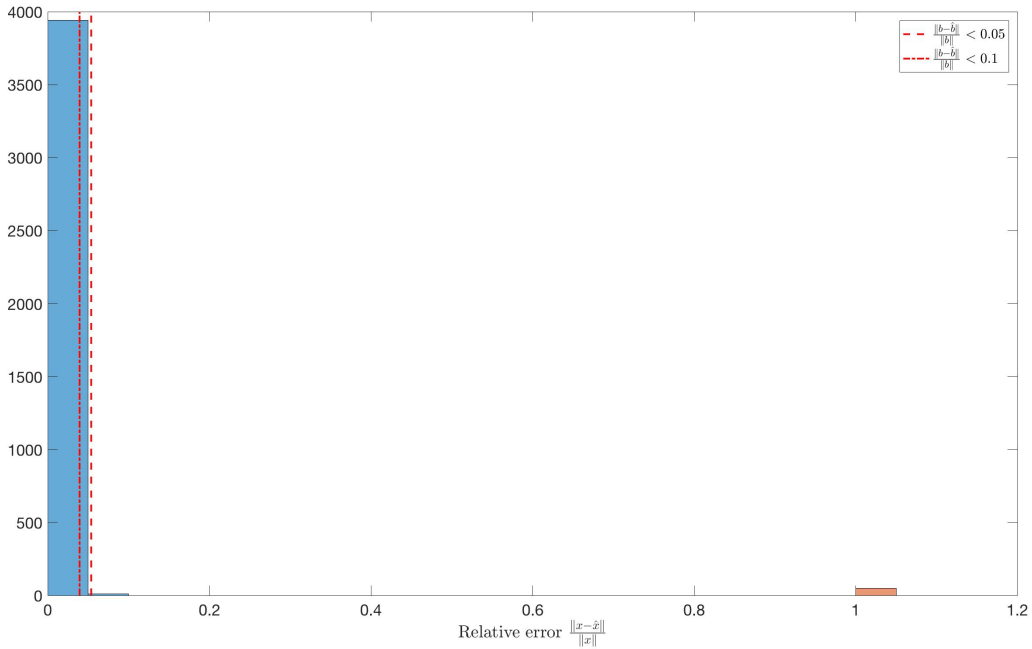
	3904 (97.6%)	34 (0.009%)	1	1	59 (0.01%)
$I_\delta(x, \widehat{x})$	0	1	2	3	4

Fault detection

- Even if some incorrect atoms are identified, the effect on the relative approximation error

$$r(x, \widehat{x}) = \frac{\|b - \mathrm{D}\widehat{x}\|}{\|b\|} = \frac{\|\mathrm{D}(x - \widehat{x})\|}{\|\mathrm{D}x\|}$$

  may be insignificant.
- The number of serious failures of the algorithm can reliably detected by counting the cases where $r(x, \widehat{x}) > \tau$, where $0 < \tau < 1$.
- The computation of the relative error does not require the knowledge of the true x.
- The difference between the data b and the reconstruction $\mathrm{D}\widehat{x}$ is a good proxy.
- Method is robust with respect to τ: in clearly misidentified cases $r(x, \widehat{x})$ is large.



The mark lines are for $\tau = 0.05$ (dashed) and $\tau = 0.1$ (dot-dashed). It is clear that the fault detection mechanism proposed is quite robust with respect to τ.

Dictionary Compression Error

- The data corresponds to a single atom out of a single class.
- Perturb the data by Gaussian white noise with standard deviation 0.03.
- We run the algorithm 5000 times, with and without correcting for DCE.
- Including the DCE leads to a correct cluster identification in 4835 cases, or 97% of the cases.
- Leaving out the DCE, the correct class is identified 4705 times, or 0.94% of the cases.
- Once the correct class is identified, the correct atom is identified in both protocols in approximately 98% of the cases.
- The effect of the DCE is observable but not as crucial as in the next example.

LIGO/VIRGO glitches classification

- Observed gravitational wave, or an event of any unknown origin registered by the device, is to be identified based on a precomputed dictionary, as finding a universal predictive parametric model is not feasible.
- Data stream is contaminated by frequent artifacts of non-gravitational origin, known as glitches, that due to their high occurrence rate may lead to false coincidence detections.
- The glitches are transients with particular morphologies that allow them to be classified and identified by using precomputed glitch libraries.

Glitches dictionary

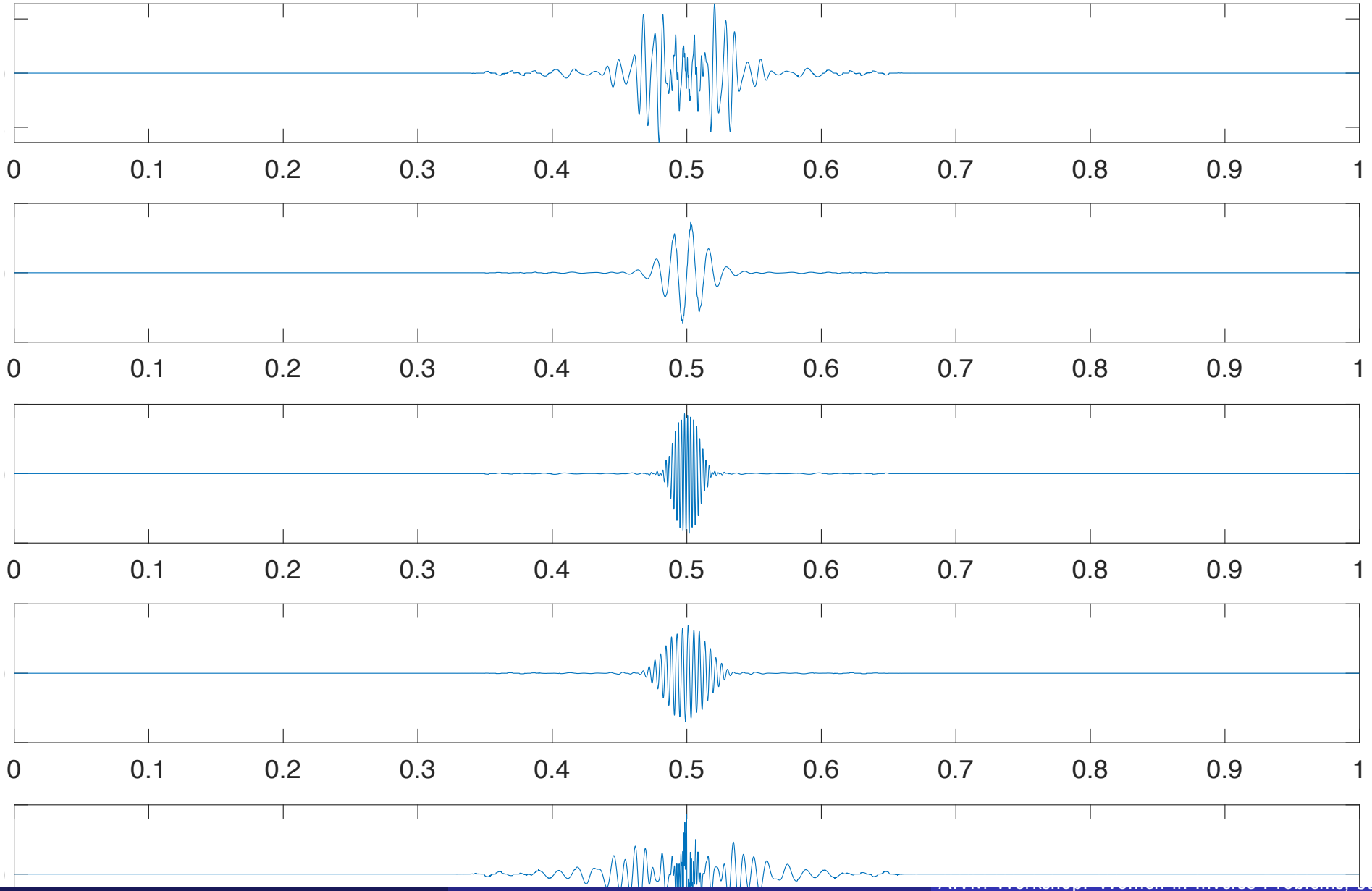
Consider three classes of glitches, sine Gaussian (SG), Gaussian (G) and ring down (RD), parametrized as

$$\begin{aligned} h_{\mathrm{SG}}(t) &= h_0 \sin\left(2\pi f_0(t-t_0)\right) e^{-(t-t_0)^2/2\tau^2}, \\ h_{\mathrm{G}}(t) &= h_0 e^{(t-t_0)^2/2\tau^2}, \\ h_{\mathrm{RD}}(t) &= h_0 \sin\left(2\pi f_0(t-t_0)\right) e^{-(t-t_0)/\tau}\theta(t-t_0), \end{aligned}$$

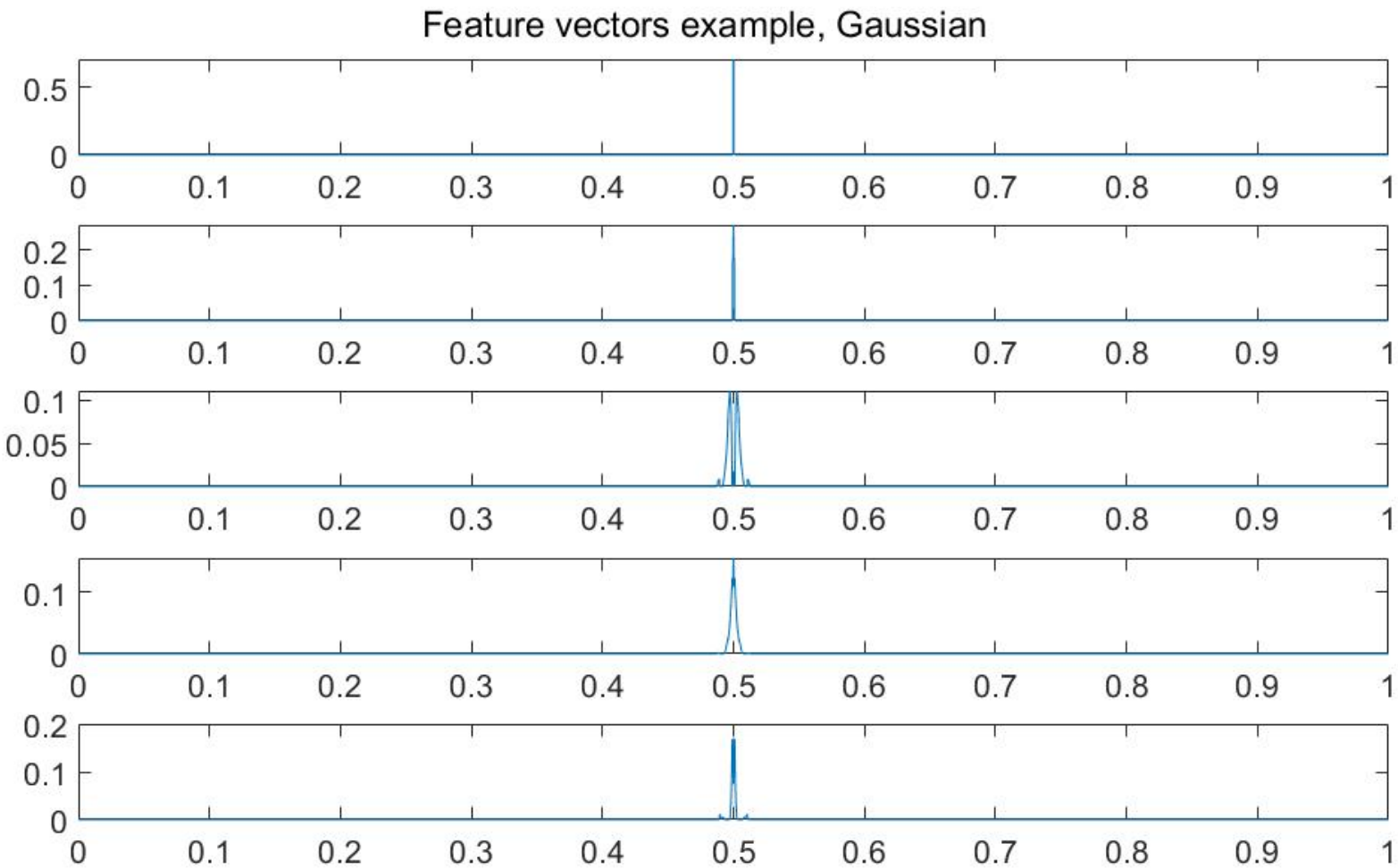
where h_0 is the amplitude, f_0 is the frequency, t_0 is the center time and τ is the characteristic time of the glitch, and θ is the Heaviside function.

- The dictionary is made of 1000 random glitches per class.
- Dimensionality of atoms is significant
- Number of feature vectors changes for the three classes : 50 for sine Gaussian and ring down signals, 11 for Gaussian signals.

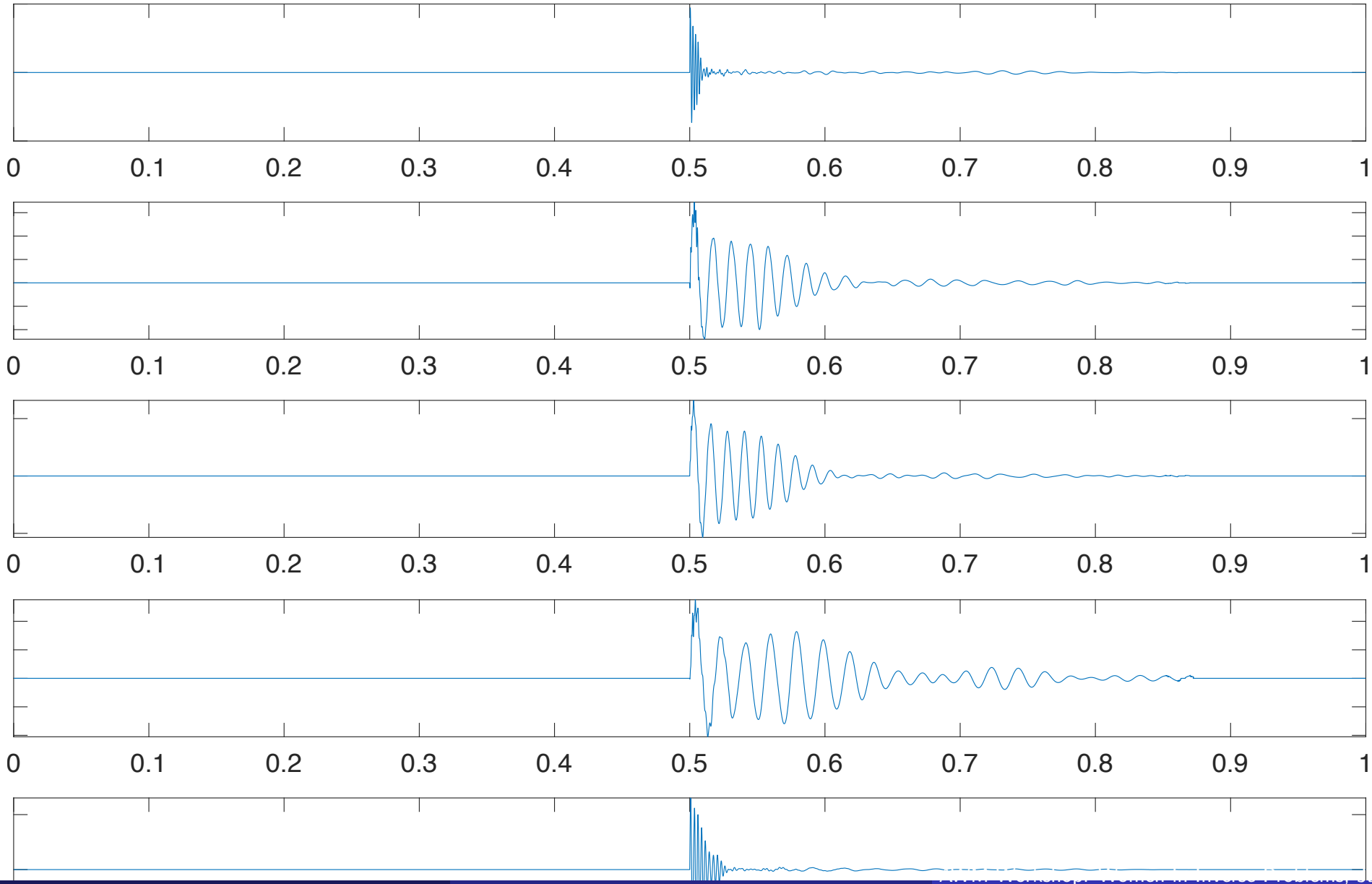
Sine Gaussians feature vectors



Gaussians feature vectors



Ring down feature vectors



To identify the class, for a given data vector b, we find the coefficient vector $h^{(j)}$ for each cluster, $1 \leq j \leq 3$, such that

$$b \approx \mathsf{W}^{(j)} h^{(j)} = \widehat{b}^{(j)},$$

where $h^{(j)}$ is computed using the IAS algorithm with sparsity and positivity constraints, including the DCE correction.Then compare $\widehat{b}^{(j)}$ with b using the structural dissimilarity index,

$$\mathrm{DSSIM}(b, \widehat{b}^{(j)}) = \frac{1}{2}\big(1 - \mathrm{SSIM}(b, \widehat{b}^{(j)})\big),$$

where SSIM is the structural similarity index. The chosen class is the one corresponding to smallest dissimilarity index.

Results

We run the test 20000 times with additive noise $\sigma = 0.02$, and find that the algorithm identifies correctly the class with 100% accuracy (left). Increasing the additive noise level to $\sigma = 0.05$ still get 99.8% correct classification (right).

	SG	G	RD
SG	6812	0	0
G	0	6546	0
RD	0	0	6642

	SG	G	RD
SG	6798	0	0
G	0	6522	37
RD	1	0	6642

.

Importance of DCE

For comparison, we then run the same experiment leaving out the DCE correction. Even with the lower noise level, $\sigma = 0.02$, the class identification goes egregiously astray, giving a 100% misclassification result, with the confusion matrix

	SG	G	RD
SG	0	6754	0
G	6592	0	0
RD	6654	0	0

.

Simplified DCE

Finally, we approximate the DCE covariance matrix by a scaled identity matrix,

$$\mathsf{C}_{\mathrm{DCE}}^{(j)} \approx \mathrm{trace}\big(\mathsf{C}_{\mathrm{DCE}}^{(j)}\big)\mathsf{I}_m.$$

With the noise level $\sigma = 0.02$, the identification of the clusters was 100% correct, and with $\sigma = 0.05$, an identification performance of 99.5% was achieved again.

References

- Llorens-Monteagudo M, Torres-Forné A, Font JA and Marquina A (2019) Classification of gravitational-wave glitches via dictionary learning. Classical and Quantum Gravity **36** 075005
- Pragliola M, Calvetti D and Somersalo E (2022) Overcomplete representation in a hierarchical Bayesian framework. Inverse Problems and Imaging **16**: 19-38. doi: 10.3934/ipi.2021039
- Waniorek N, Somersalo E and Calvetti D (2023) Bayesian hierarchical dictionary learning. Inverse Problems (electronic version: DOI 10.1088/1361-6420/acad21)

The details of this presentation are in a forthcoming manuscript

- A. Bocchinfuso, D. Calvetti and E. Somersalo (2023) Error enhanced layered Bayesian dictionary learning.