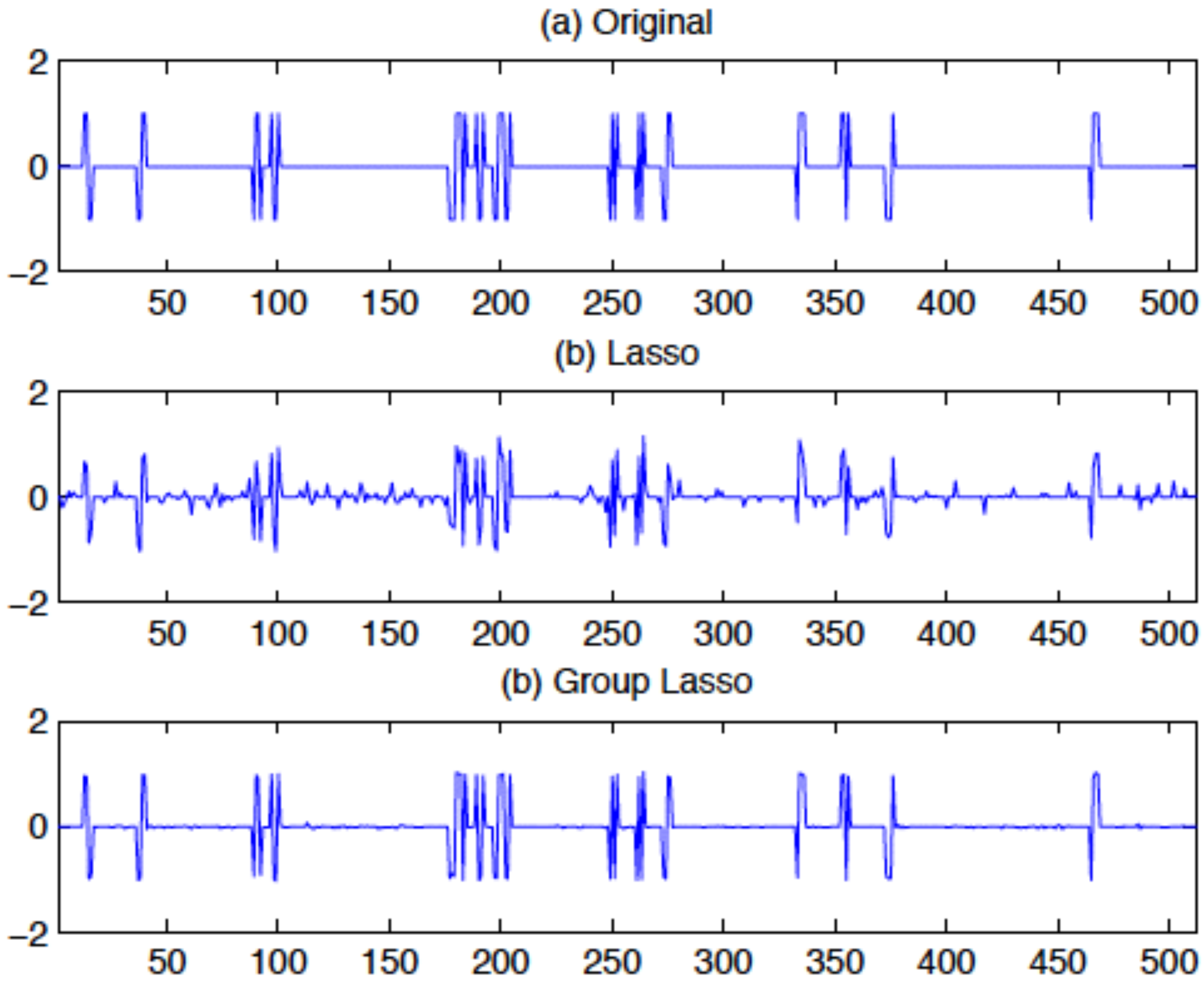
Structured Sparsity

Nataliya Shapovalova

Machine Learning Reading Group

Motivation

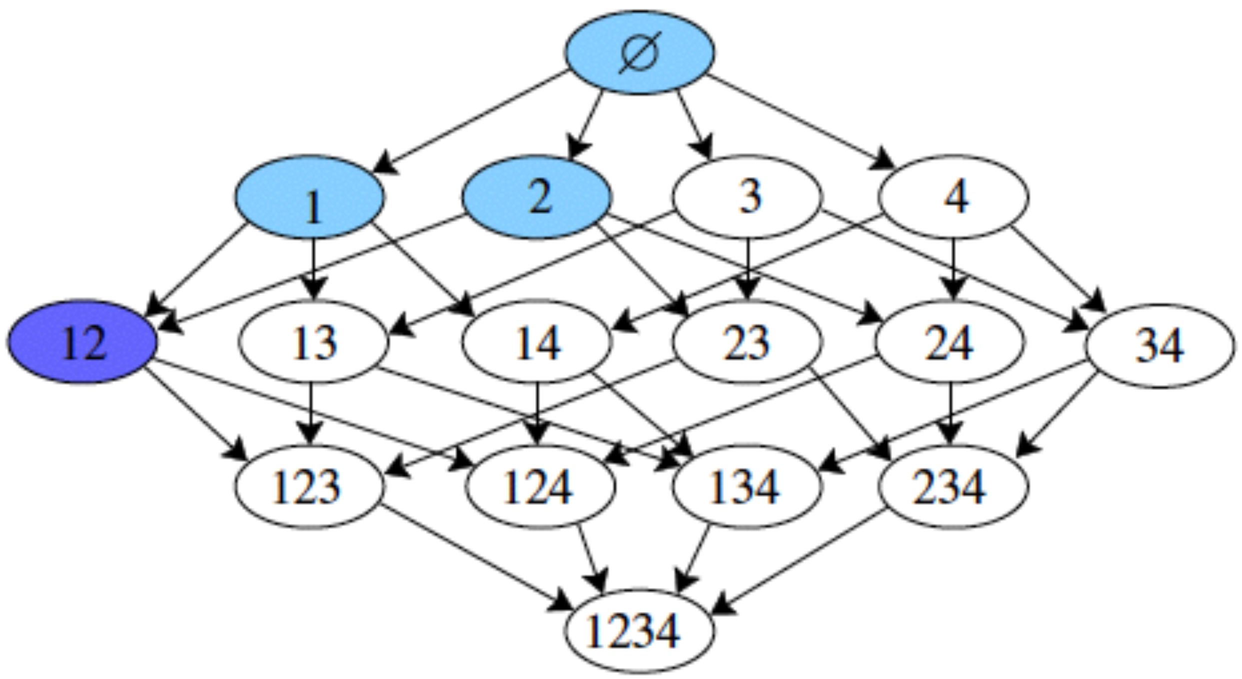
- Compressive sensing:



Motivation

- Hierarchical MKL:

Many kernels can be decomposed as a sum of many "small" kernels indexed by a certain set V: $\boxed{k(x,x') = \sum_{v \in V} k_v(x,x')}$



Outline

- Formulation
- Different types of structured sparsity
- Application: dictionary learning

Formulation

Given data $\{(x_n, y_n)\}_{n=1}^{N} \subseteq \mathcal{X} \times \mathcal{Y}$

Goal: Learn the model parameter

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\frac{1}{N}\sum_{n=1}^{N} L(\mathbf{w}; x_n, y_n)}_{\text{empirical risk}} + \underbrace{\Omega(\mathbf{w})}_{\text{regularizer}}$$

Formulation

Given data $\{(x_n, y_n)\}_{n=1}^{N} \subseteq \mathcal{X} \times \mathcal{Y}$

Goal: Learn the model parameter

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\frac{1}{N} \sum_{n=1}^{N} L(\mathbf{w}; x_n, y_n)}_{\text{empirical risk}} + \underbrace{\Omega(\mathbf{w})}_{\text{regularizer}}$$

- Sparsity hypothesis: not all dimensions of *x* are needed (many features are irrelevant)
- Setting the corresponding weights to zero leads to a sparse *w*

Why Sparsity?

- Easier to interpret
- Generalize better
- Fast to run

Structured regularization

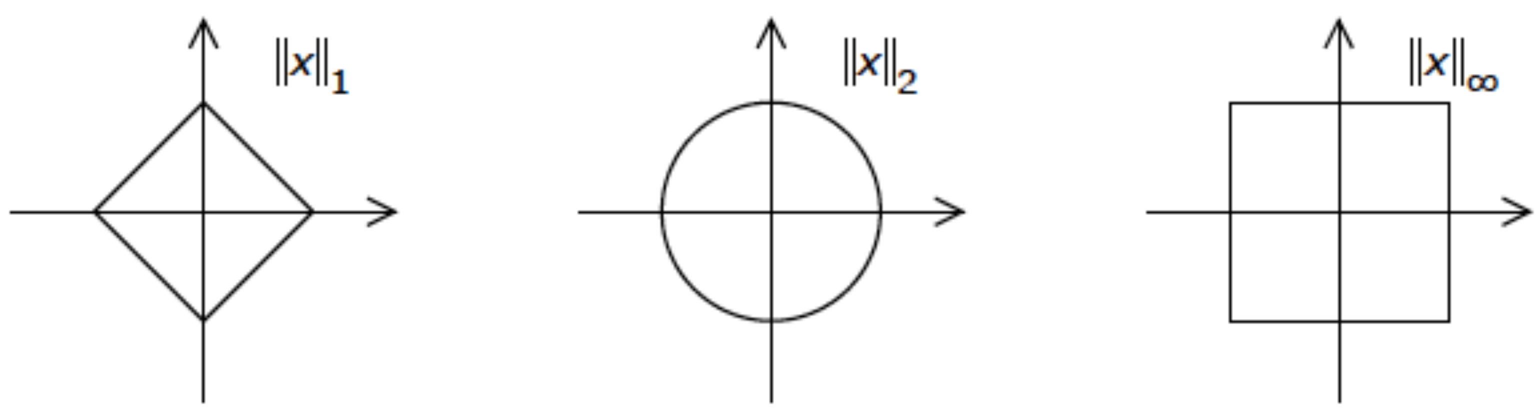
- Non-overlapping groups
  - L1-norm
  - Group Lasso
- Overlapping groups
  - Tree-structured groups
  - Contiguous groups
  - Directed-Acyclic-Graph groups

Structured regularization

- Non-overlapping groups
  - L1-norm
  - Group Lasso
- Overlapping groups
  - Tree-structured groups
  - Contiguous groups
  - Directed-Acyclic-Graph groups

Norms: a quick review

ℓ_p**-norms** $(p \geq 1)$: $\boxed{\|\mathbf{w}\|_p = \left(\sum_i |w_i|^p\right)^{1/p}}$



$$\|\mathbf{w}\|_1 = \sum_i |w_i|, \quad \|\mathbf{w}\|_2 = \sum_i w_i^2, \quad \|\mathbf{w}\|_\infty = \max_i |w_i|$$

- L1-regularization naturally leads to sparse solution

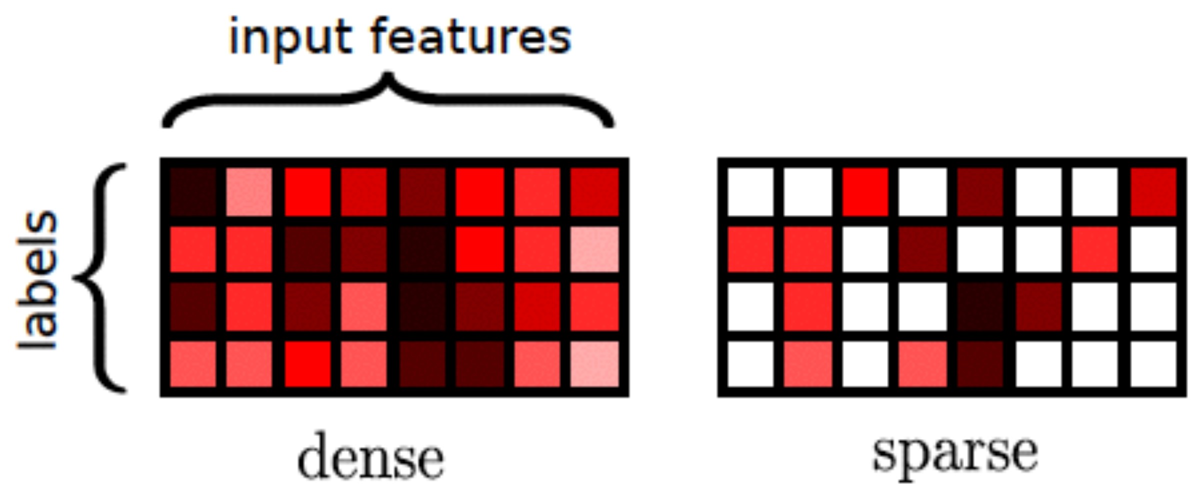
L1-norm

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\frac{1}{N}\sum_{n=1}^{N} L(\mathbf{w}; x_n, y_n)}_{\text{empirical risk}} + \underbrace{\Omega(\mathbf{w})}_{\text{regularizer}}$$

$$\Omega(w) = \sum_i |w_i|$$

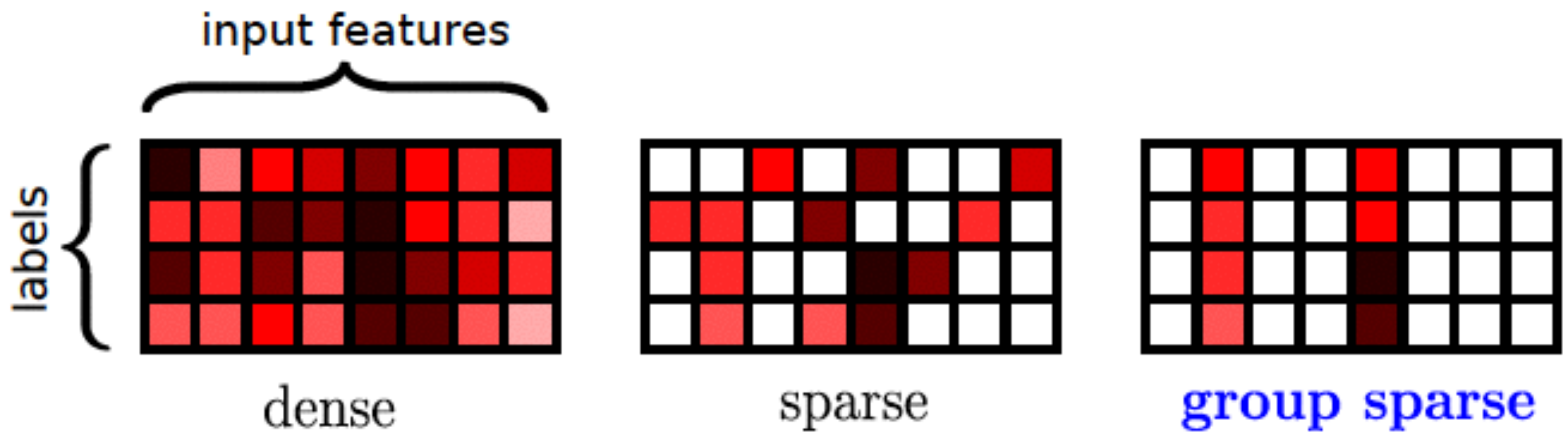
- L1-regularization naturally leads to sparse solution:

Sparsity on a grid



- Sparse solution is not very useful, and we still need all the input features

Sparsity on a grid



- Sparse solution is not very useful, and we still need all the input features

Group sparsity

- D features
- M groups, $G_1, \ldots, G_M$; $G_m \subseteq \{1, \ldots, D\}$

Regularization: L1-norm of L2-norms $\Omega(\mathbf{w}) = \sum_{m=1}^{M} \lambda_m \|\mathbf{w}_m\|_2$

Group sparsity

- D features
- M groups, $G_1, \ldots, G_M$; $G_m \subseteq \{1, \ldots, D\}$

Regularization: L1-norm of L2-norms $\Omega(\mathbf{w}) = \sum_{m=1}^{M} \lambda_m \|\mathbf{w}_m\|_2$

Observation: In L1-norm each feature belongs to exactly one group

Structured sparsity

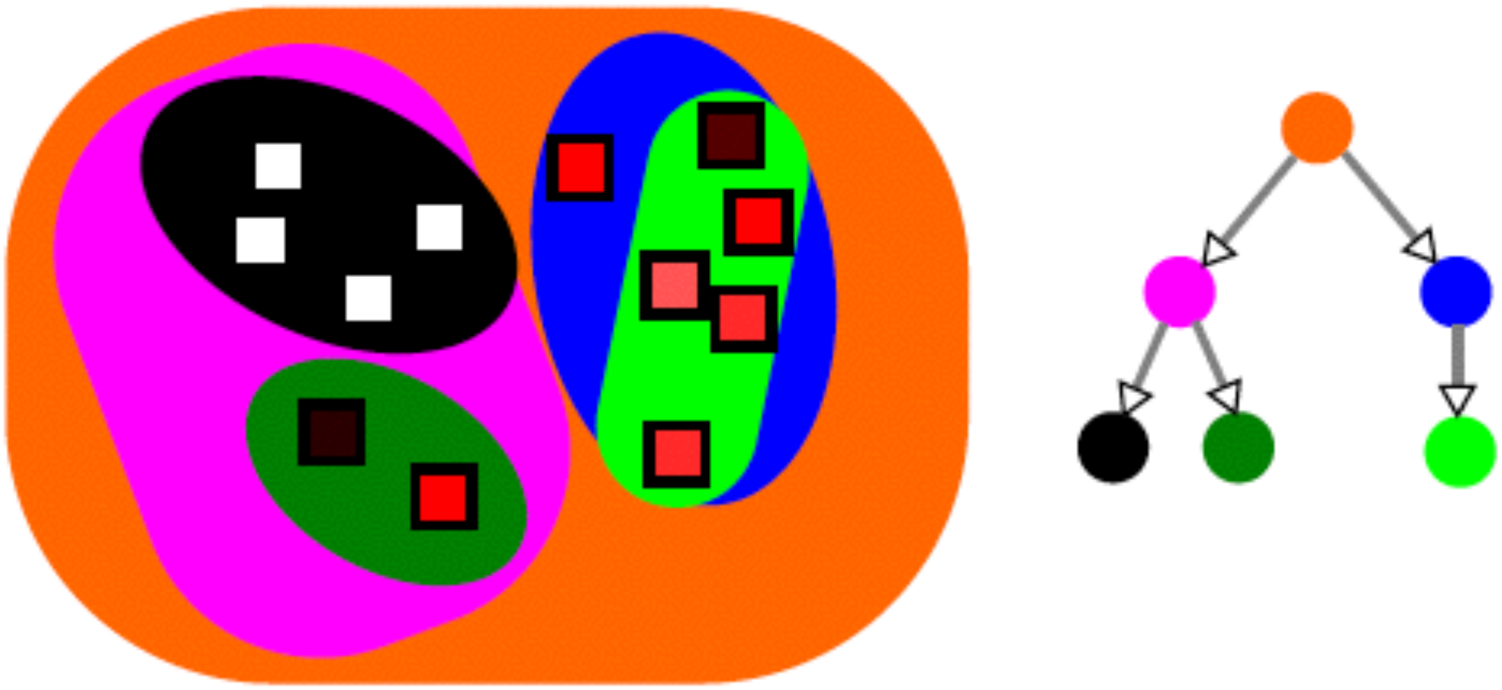
- Structured sparsity cares about the structure of the feature space
- Group-Lasso regularization generalizes well and it's still convex
- Choice of groups: problem dependent, opportunity to use prior knowledge to favour certain structural patterns

Structured regularization

- Non-overlapping groups
  - L1-norm
  - Group Lasso
- Overlapping groups
  - Tree-structured groups
  - Contiguous patterns
  - Directed-Acyclic-Graph groups

Tree structured groups

Assumption: if two groups overlap, one contains the other

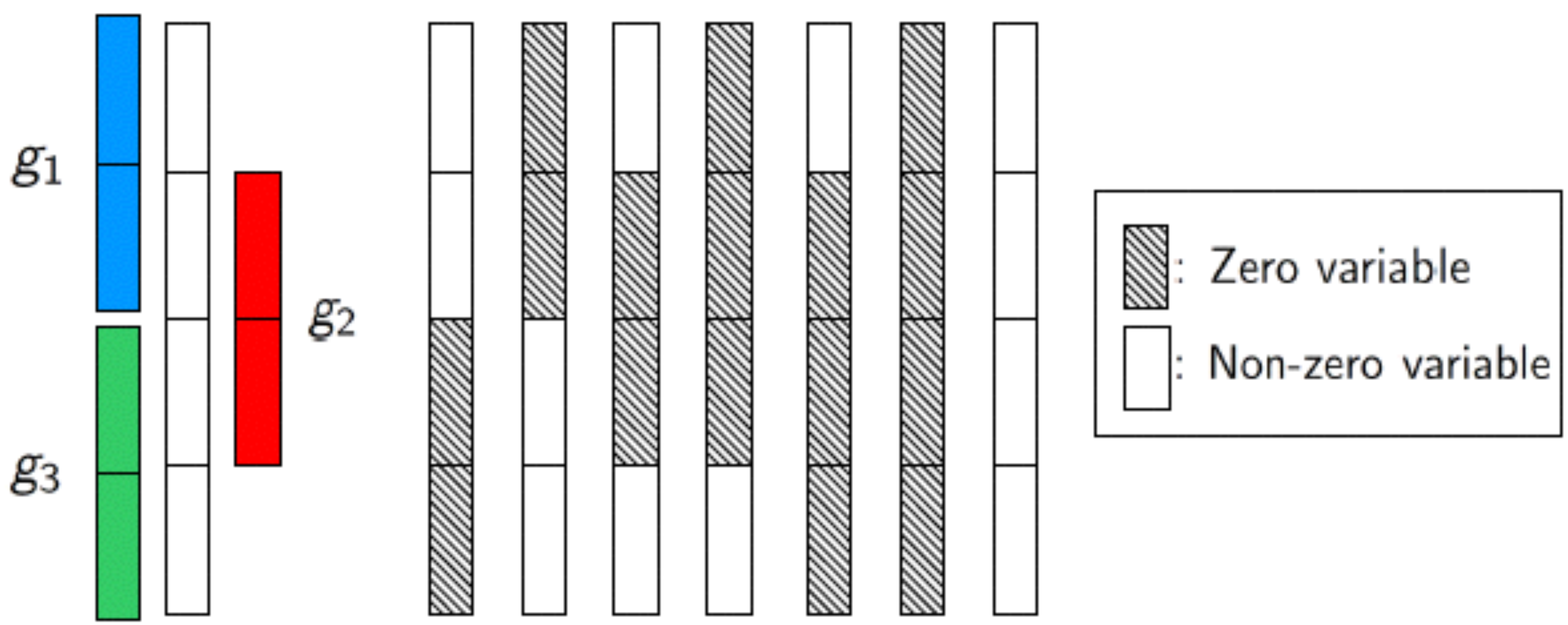


If a group is discarded, all its descendants are also discarded

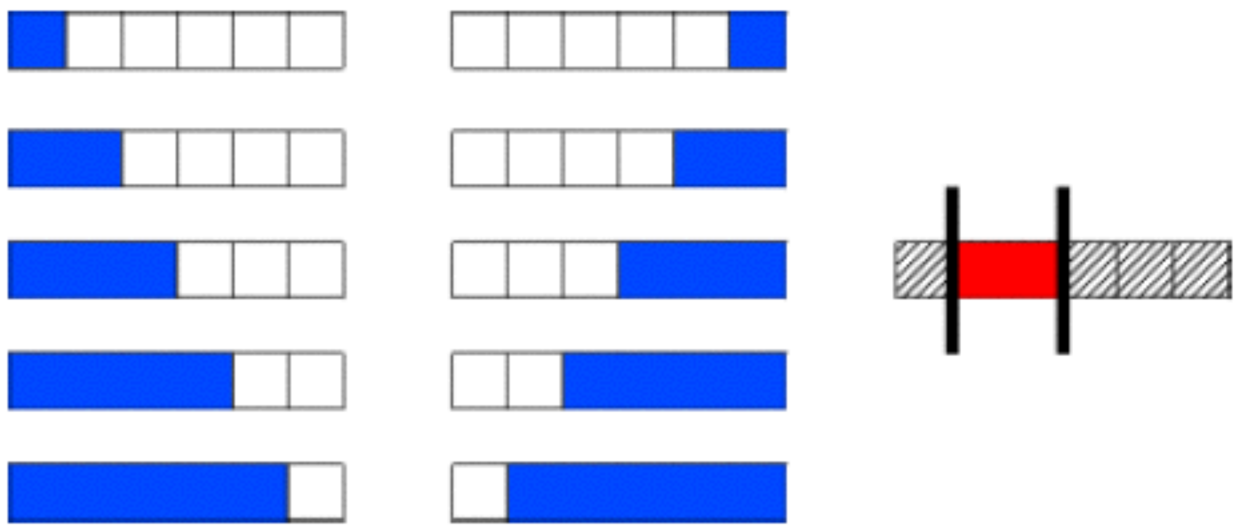
Contiguous patterns

Sets of possible zero patterns and possible non-zero patterns

$$\mathcal{Z} = \Big\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \Big\} \quad \text{and} \quad \mathcal{N} = \Big\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \Big\}$$



Contiguous patterns

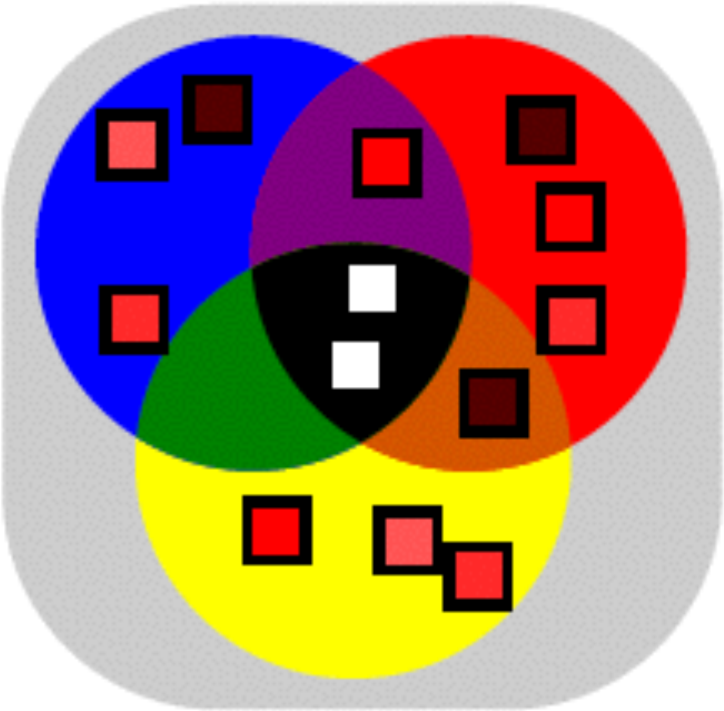


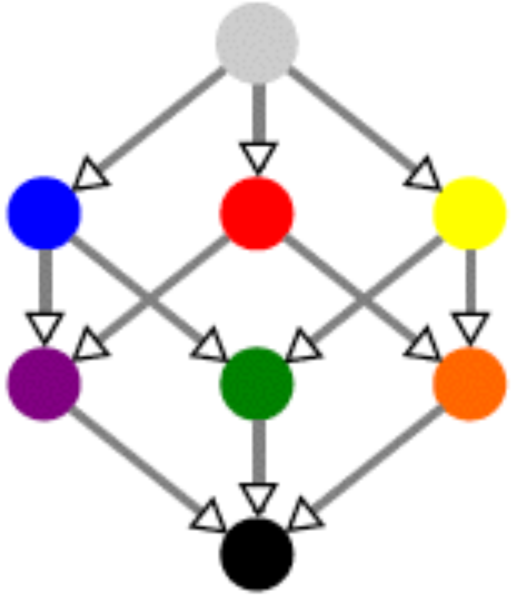
G is the set of blue groups.

Any union of blue groups set to zero leads to the selection of a contiguous pattern (red).

Arbitrary groups

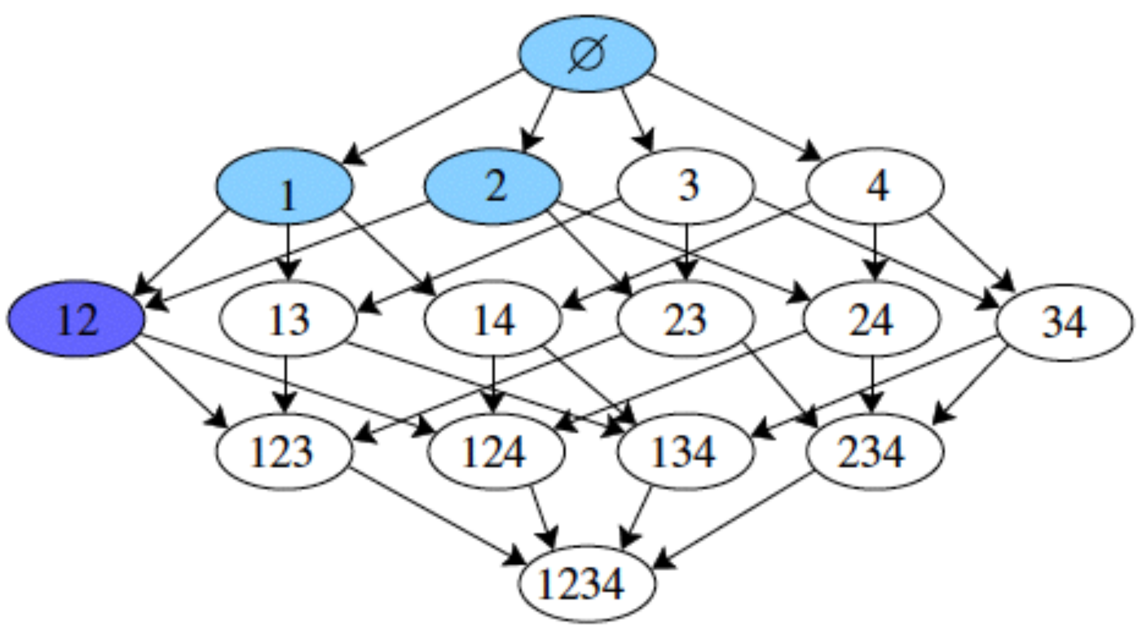
In general: groups can be represented as a directed acyclic graph





Hierarchical MKL:

Many kernels can be decomposed as a sum of many "small" kernels indexed by a certain set V: $\boxed{k(x,x') = \sum_{v \in V} k_v(x,x')}$



- Graph-based structured regularization
  - $\mathrm{D}(v)$ is the set of descendants of $v \in V$:

$$\sum_{v \in V} d_v \|w_{\mathrm{D}(v)}\|_2 = \sum_{v \in V} d_v \left(\sum_{t \in \mathrm{D}(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If v is selected, so are all its ancestors

Application: dictionary learning

- Given data matrix $X = (x_1^\top, \ldots, x_n^\top)^\top \in \mathbb{R}^{n \times p}$, principal component analysis (PCA) may be seen from two perspectives:
  - **Analysis view**: find the projection $v \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $v_1, \ldots, v_k$ such that all x_i have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent

- Sparse extensions
  - Interpretability
  - High-dimensional inference
  - **Two views are differents**

Application: dictionary learning

Sparse PCA:

$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \quad \text{s.t.} \quad \forall j,\ \|\boldsymbol{\alpha}_j\|_2 \leq 1$$

Sparse structured PCA

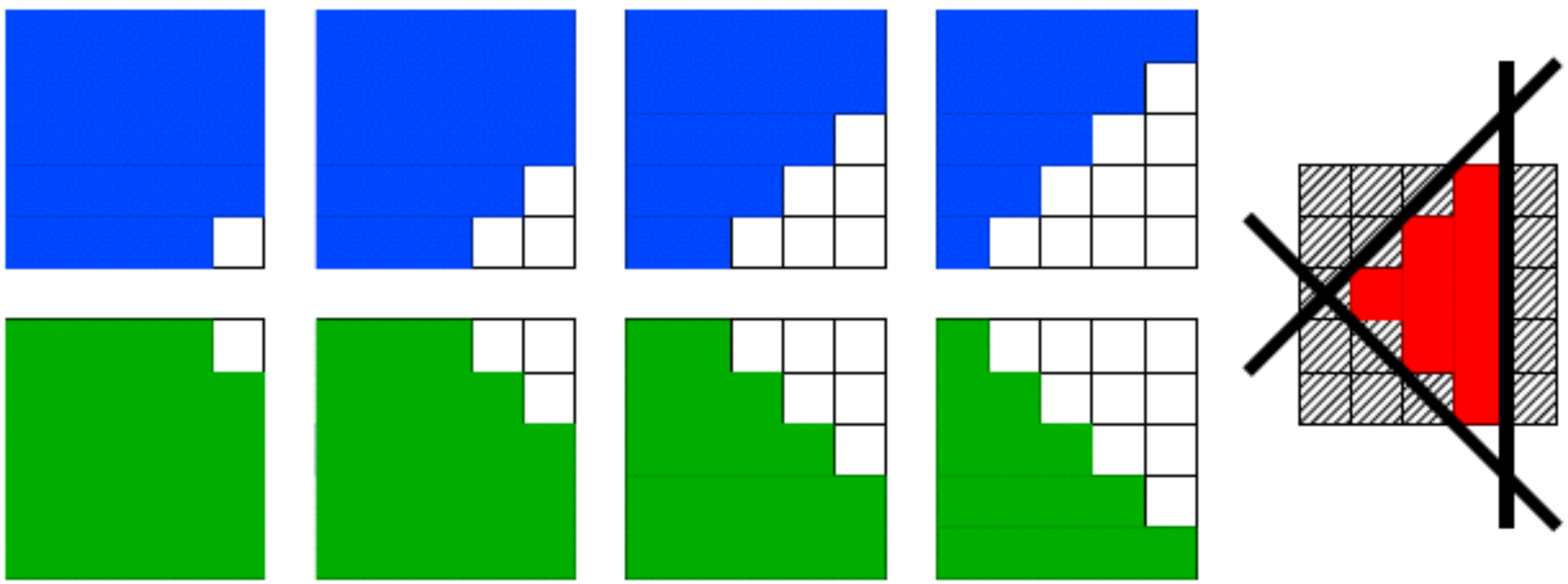
$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \Omega(\mathbf{d}_j) \quad \text{s.t.} \quad \forall j,\ \|\boldsymbol{\alpha}_j\|_2 \leq 1$$

$$\Omega(\mathbf{d}) = \textstyle\sum_{g\in\mathcal{G}} \|\mathbf{d}_g\|_2$$

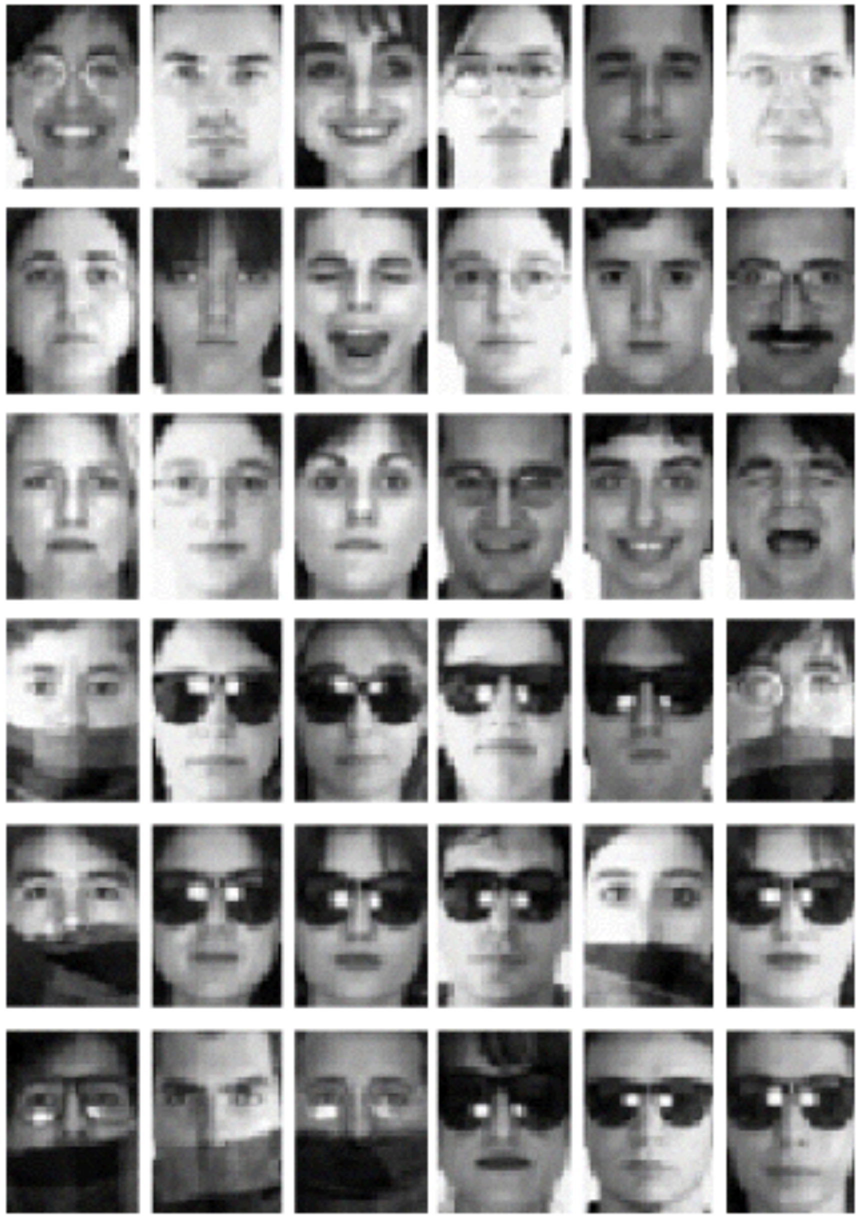
In signal processing $X^\top = \underbrace{V}_{\text{dictionary } D} \underbrace{U^\top}_{\text{decomposition coefficients } \alpha} = D\alpha$

Application: dictionary learning

- $\Omega(\mathbf{d}) = \sum_{g \in \mathcal{G}} \|\mathbf{d}_g\|_2$: Selection of "convex" patterns on a 2-D grids.

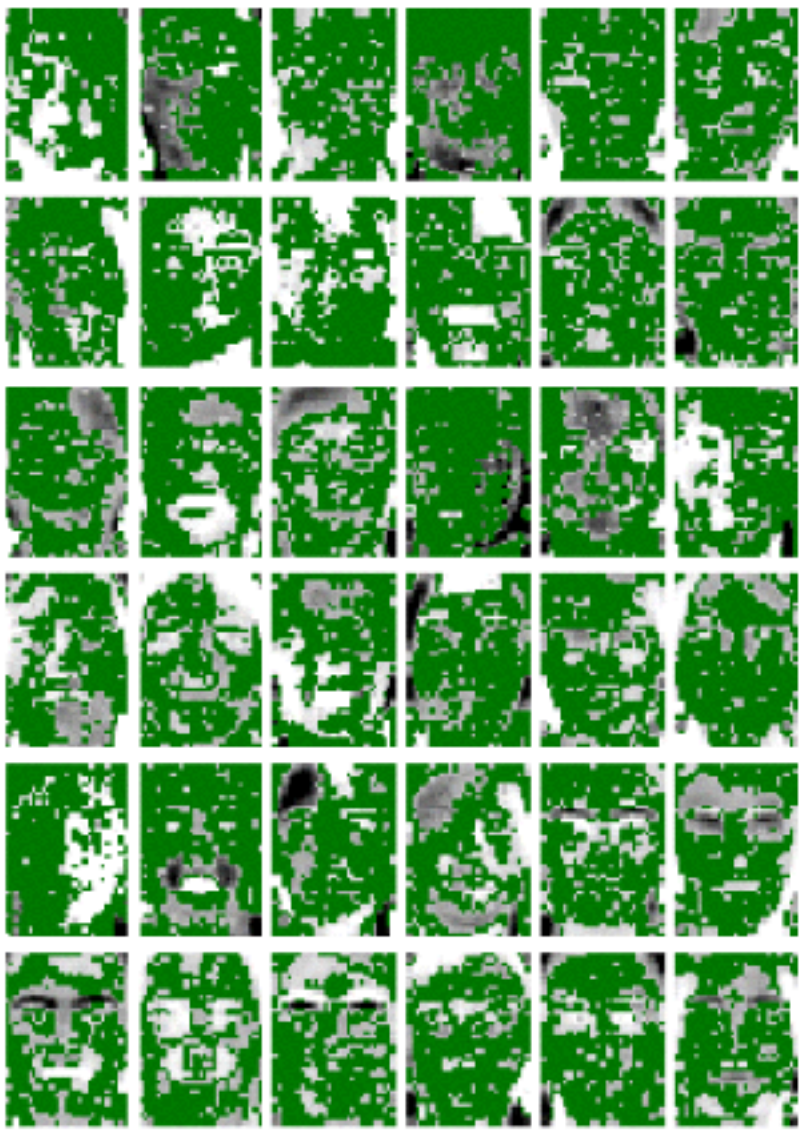


Application: dictionary learning

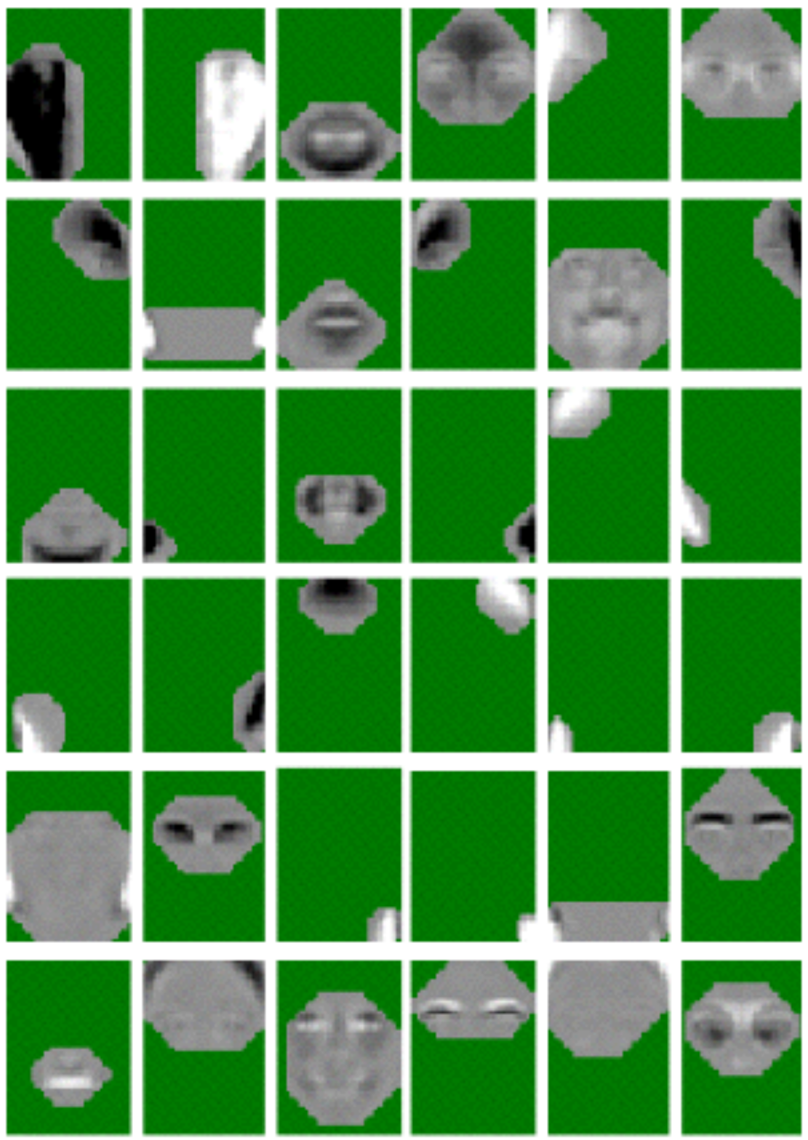


- AR Face database
- 100 individuals (50 W/50 M)
- For each
  - 14 non-occluded
  - 12 occluded
  - lateral illuminations
  - reduced resolution to 38×27 pixels

Application: dictionary learning

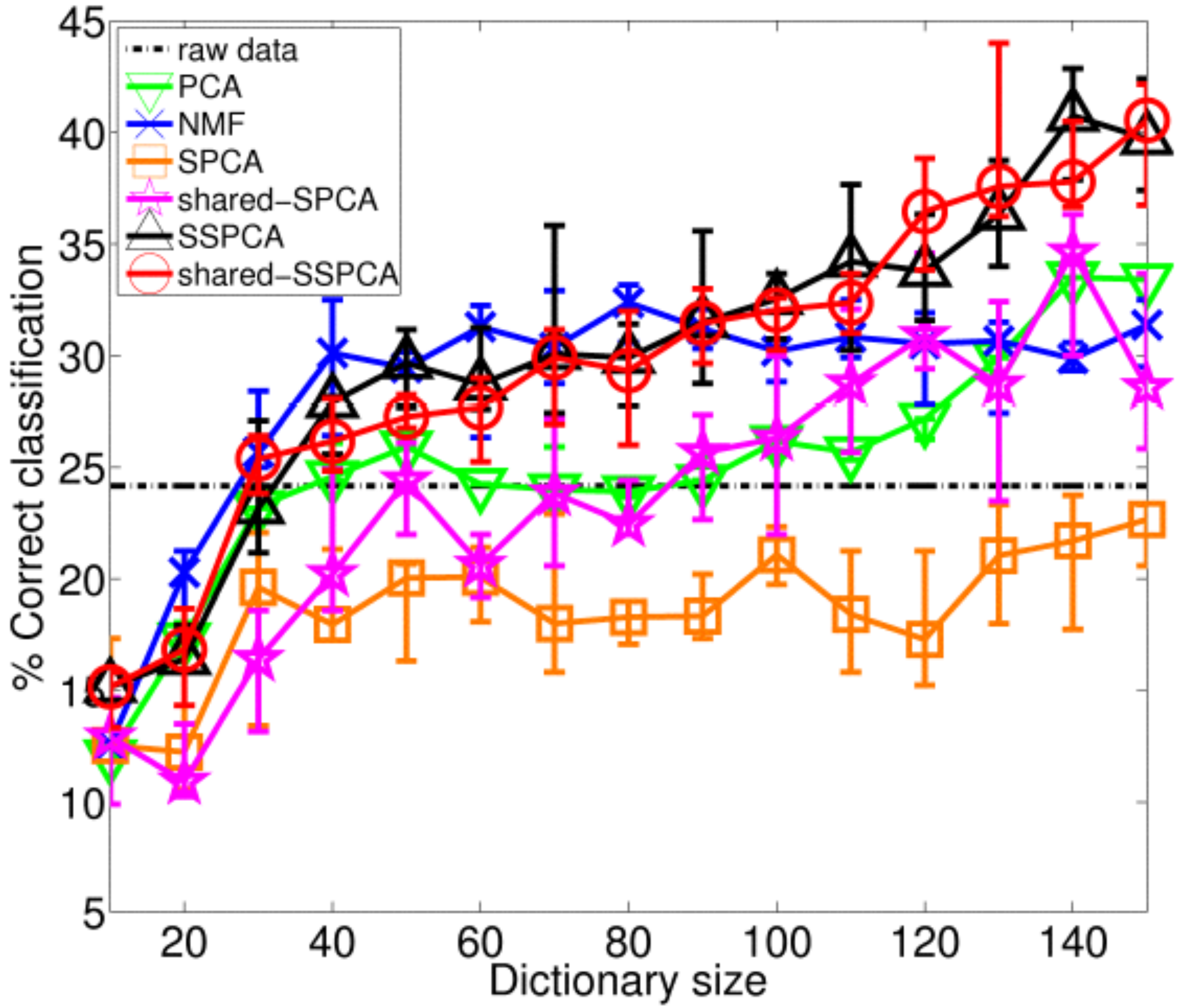


SPCA



SSPCA

k-NN classication based on decompositions



Thank you!