9. Latent factor estimation

- mixture models and biconvex formulation
- cost functions, regularization, and constraints
- clustering with prior information
- principal component analysis
- solution and interpretations
- quadratic regularization
- robust low rank approximation
- structural regularization and constraints

Mixture models

inverse problem for general machine learning models:

$$\text{minimize} \quad \sum_{i=1}^m f_\theta(x_i, y_i)$$

- $(x_i, y_i) \in \mathbf{R}^n \times \mathbf{R}$ are data points
- variable $\theta \in \mathbf{R}^n$ are model parameters
- $f_\theta \colon \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}$ is some cost function

mixture model assumes data generated by k different models,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m \sum_{j=1}^k z_{ij} f_{\theta_j}(x_i, y_i) \\ \text{subject to} & z_i \in \{e_1, \ldots, e_k\}, \quad i = 1, \ldots, m \end{array}$$

- variables $\theta_1, \ldots, \theta_k \in \mathbf{R}^n$ and $z_1, \ldots, z_m \in \mathbf{R}^k$
- $z_i \in \mathbf{R}^k$ are *one-hot* labels indicating which of the k models generated the ith data point; sometimes called **latent factors**

rewrite the problem as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m z_i^T r_i \\ \text{subject to} & r_{ij} = f_{\theta_j}(x_i, y_i), \quad i = 1, \ldots, m, \quad j = 1, \ldots, k \\ & z_i \in \{0,1\}^k, \quad \mathbf{card}\, z_i = 1, \quad i = 1, \ldots, m \end{array}$$

- $r_i \in \mathbf{R}^k$ are auxiliary, denote the vector of costs
- last constraint requires z_i to be one-hot

interpretation:

- first separate m data points into k groups, each corresponding to a different model, according to the latent factors z_i
- fit parameters θ_j of each model to the data in different groups

however, solving the inverse problem is very difficult (actually, NP-hard)

Biconvex formulation

relaxation of integer constraints on z_i to convex constraints:

$$0 \preceq z_i \preceq \mathbf{1}, \quad \mathbf{1}^T z_i = 1, \quad i = 1, \ldots, m$$

so we have **biconvex** inverse problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m z_i^T r_i \\ \text{subject to} & r_{ij} = f_{\theta_j}(x_i, y_i), \quad i = 1, \ldots, m, \quad j = 1, \ldots, k \\ & 0 \preceq z_i \preceq \mathbf{1}, \quad \mathbf{1}^T z_i = 1, \quad i = 1, \ldots, m \end{array}$$

- variables $\theta_1, \ldots, \theta_k$ and $z_1, \ldots, z_m$
- convex in θ_j for fixed z_i, and convex in z_i for fixed θ_j
- can be (approximately) solved via alternate convex search
- relaxation is often tight, *i.e.*, the solution z_i is often one-hot

interpretation:

- latent factors z_i lie in the probability simplex

$$\{z \in \mathbf{R}^k \mid 0 \preceq z \preceq \mathbf{1}, \quad \mathbf{1}^T z = 1\} = \mathbf{conv}\{e_1, \ldots, e_k\}$$

  (whose vertices are one-hot vectors)

- relax 'hard' labels to 'soft' labels, $i.e.$, probabilities of each data point belonging to each model

- objective is the *expected* cost of the data points under the soft labels

in ACS iterations, we alternate between

- fit θ_j given data points assigned to each model according to $\tilde{z}_i$

- find best label assignments z_i for each data point given $\tilde{\theta}_j$

Cost functions

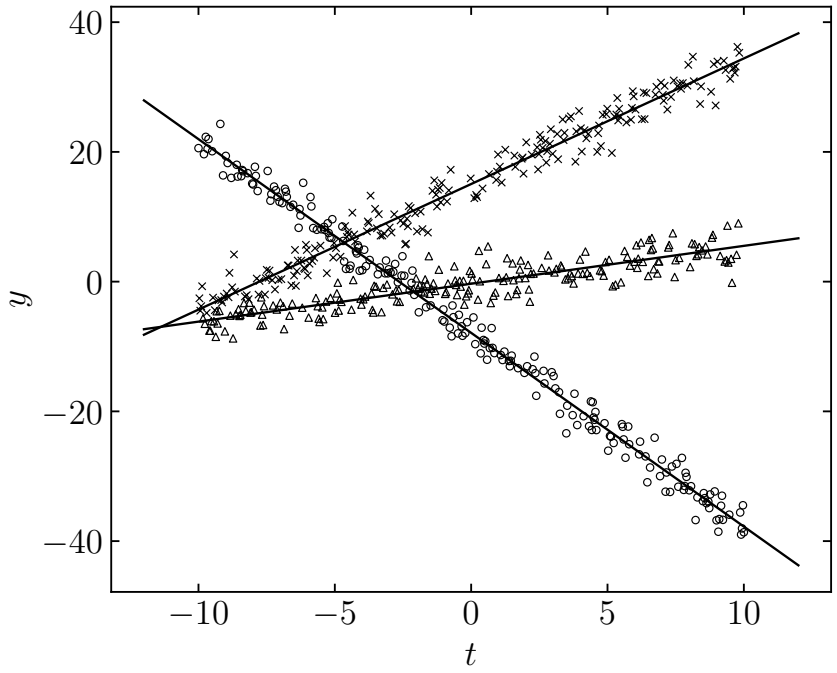
regression models:

$$f_{\theta_i}(x, y) = \phi(x^T\theta_i - y), \quad i = 1, \ldots, k$$

penalty function ϕ: $\mathbf{R} \to \mathbf{R}$ can be quadratic, ℓ_p-norm, Huber, etc.

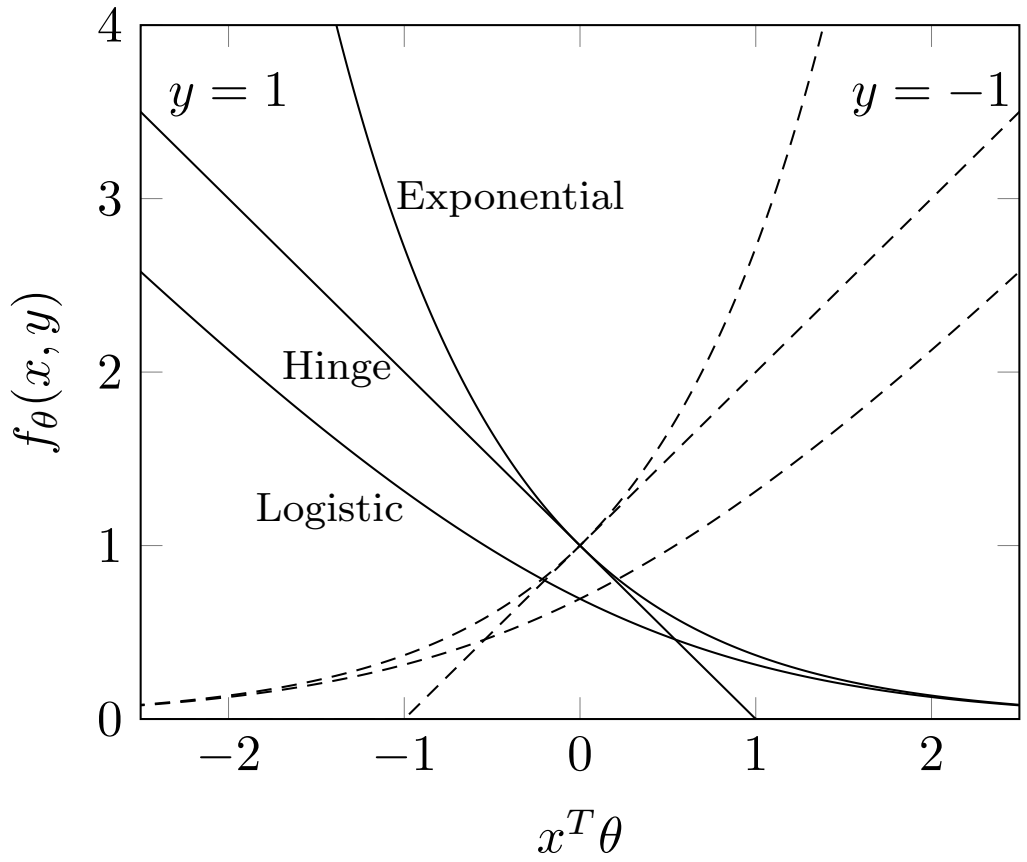
example: mixture of linear regressions

$$\begin{array}{ll} \text{min.} & \sum_{i=1}^m z_i^T r_i \\ \text{s.t.} & r_{ij} = \left(x_i^T\theta_j - y_i\right)^2 \\ & z_i \in \{0,1\}^k, \ \mathbf{card}\, z_i = 1 \\ & i = 1, \ldots, m, \ j = 1, \ldots, k \end{array}$$



classification models:

- logistic regression: $f_{\theta_i}(x, y) = \log(1 + \exp(-yx^T\theta_i))$
- support vector classifier: $f_{\theta_i}(x, y) = \max\{0,\ 1 - yx^T\theta_i\}$
- exponential loss: $f_{\theta_i}(x, y) = \exp(-yx^T\theta_i)$



- multiclass logistic regression: given $(X^{(i)}, y^{(i)})$, $i = 1, \ldots, m$, where

$$X^{(i)} = \left[\begin{array}{ccc} x_1^{(i)} & \cdots & x_q^{(i)} \end{array} \right]^T \in \mathbf{R}^{q \times n}, \qquad y^{(i)} \in \{e_1, \ldots, e_q\} \subseteq \mathbf{R}^q$$

  – $x_j^{(i)} \in \mathbf{R}^n$: feature vector contributing to the jth class

  cost function given by

$$f_{\theta_i}(X, y) = -\log(y^T u / \mathbf{1}^T u), \qquad u = \left[\begin{array}{c} \exp(x_1^T \theta_i) \\ \vdots \\ \exp(x_q^T \theta_i) \end{array} \right]$$

Regularization and constraints

probabilistic properties in z_i: $e.g.$, maximum entropy regularizer

$$\gamma \sum_{i=1}^{m} \sum_{j=1}^{k} z_{ij} \log z_{ij}$$

encourages stochastic z_i, $i.e.$, avoid hard model assignments

parameter alignment: positivity, monotonicity, etc. for different θ_j

time series models: total variation regularizer on $z(t)$ for $t = 1, \ldots, m$,

$$\gamma \sum_{t=1}^{m-1} \|z(t+1) - z(t)\|_1$$

encourages piecewise constant $z(t)$, $i.e.$, the model assignment does not change too often

example: hierarchical logistic regression for time series,

$$
\begin{array}{ll}
\text{minimize} & \sum_{t=1}^{m} z(t)^T r(t) + \gamma \sum_{t=1}^{m-1} \|z(t+1) - z(t)\|_1 \\
\text{subject to} & r_i(t) = \log(1 + \exp(-y(t)x(t)^T \theta_i)), \\
& \qquad\qquad t = 1, \ldots, m, \quad i = 1, 2 \\
& 0 \preceq z(t) \preceq \mathbf{1}, \quad \mathbf{1}^T z(t) = 1, \quad t = 1, \ldots, m \\
& \theta_1 \succeq 0, \quad \theta_{1,1} \geq \cdots \geq \theta_{1,n} \\
& \theta_2 \preceq 0, \quad \theta_{2,1} \leq \cdots \leq \theta_{2,n}
\end{array}
$$

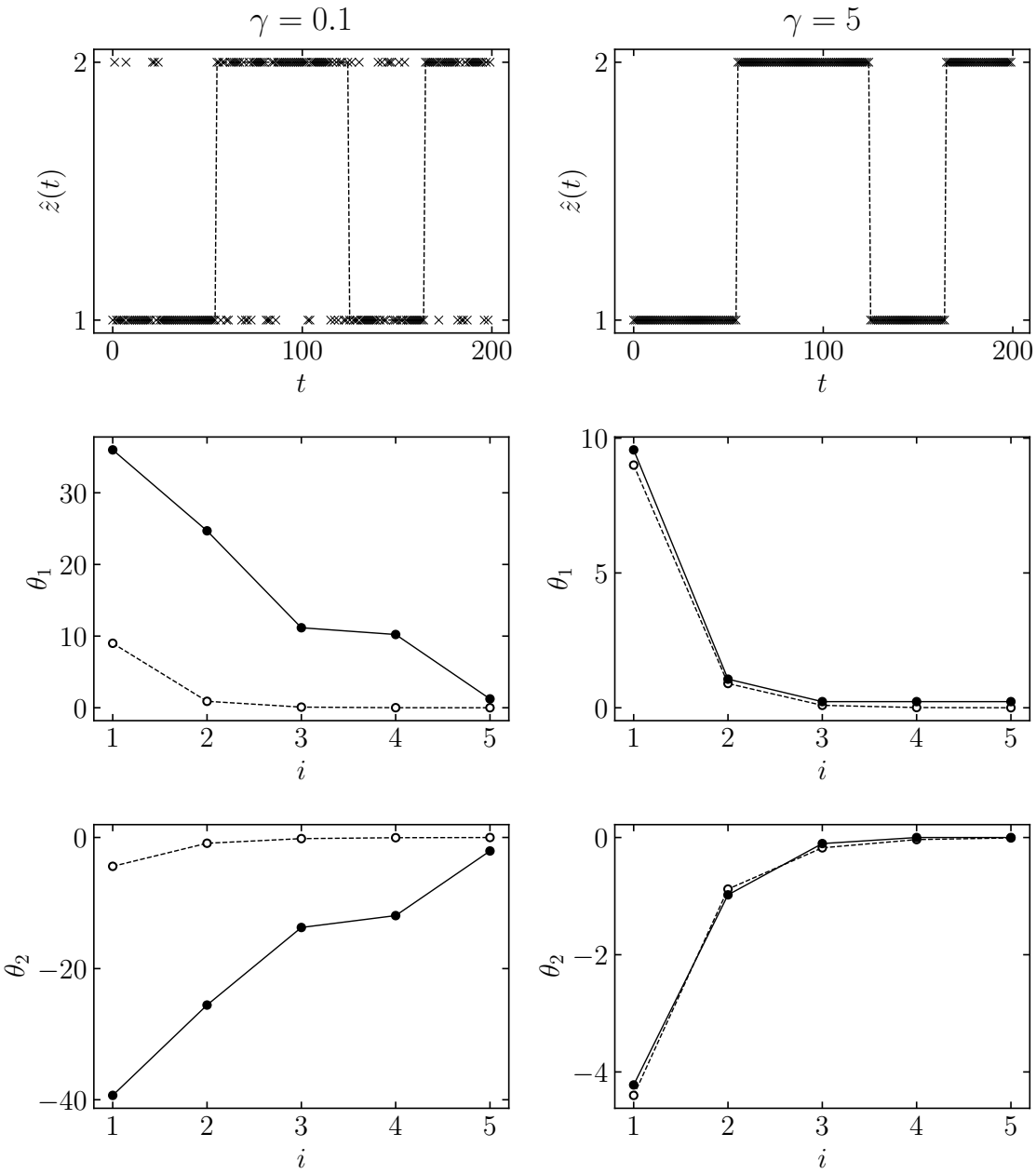
- mixture of two logistic regression models for time series data
- encourage piecewise constant latent factors $z(t)$
- θ_1 nonnegative and nonincreasing
- θ_2 nonpositive and nondecreasing

numerical results:

- one-hat label recovery:

$$\hat{z}(t) = \underset{i \in \{1,2\}}{\operatorname{argmax}}\, z_i(t)$$

- ground truth dashed



Clustering

$$\text{minimize} \quad \sum_{i=1}^{m} \min \left\{ \|x_i - \mu_j\|_2^2 \;\middle|\; j = 1, \ldots, k \right\}$$

- variables $\mu_1, \ldots, \mu_k \in \mathbf{R}^n$ denote cluster representatives
- data points $x_1, \ldots, x_m \in \mathbf{R}^n$ are given
- minimize the distance from each data point to its closest representative
- equivalent to

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} \|x_i - \mu_{c_i}\|_2^2 \\ \text{subject to} & c \in \{1, \ldots, k\}^m \end{array}$$

  with variables

  - $\mu_1, \ldots, \mu_k \in \mathbf{R}^n$
  - $c \in \mathbf{R}^m$ denoting cluster assignments

mixture model interpretation:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m z_i^T r_i \\ \text{subject to} & r_{ij} = \|x_i - \mu_j\|_2^2, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k \\ & z_i \in \{0,1\}^k, \quad \mathbf{card}\, z_i = 1, \quad i = 1, \ldots, m \end{array}$$

- variables $\mu_1, \ldots, \mu_k \in \mathbf{R}^n$ and $z_1, \ldots, z_m \in \mathbf{R}^k$

k-**means algorithm**: biconvex relaxation and ACS for clustering problem

---

k-means algorithm.

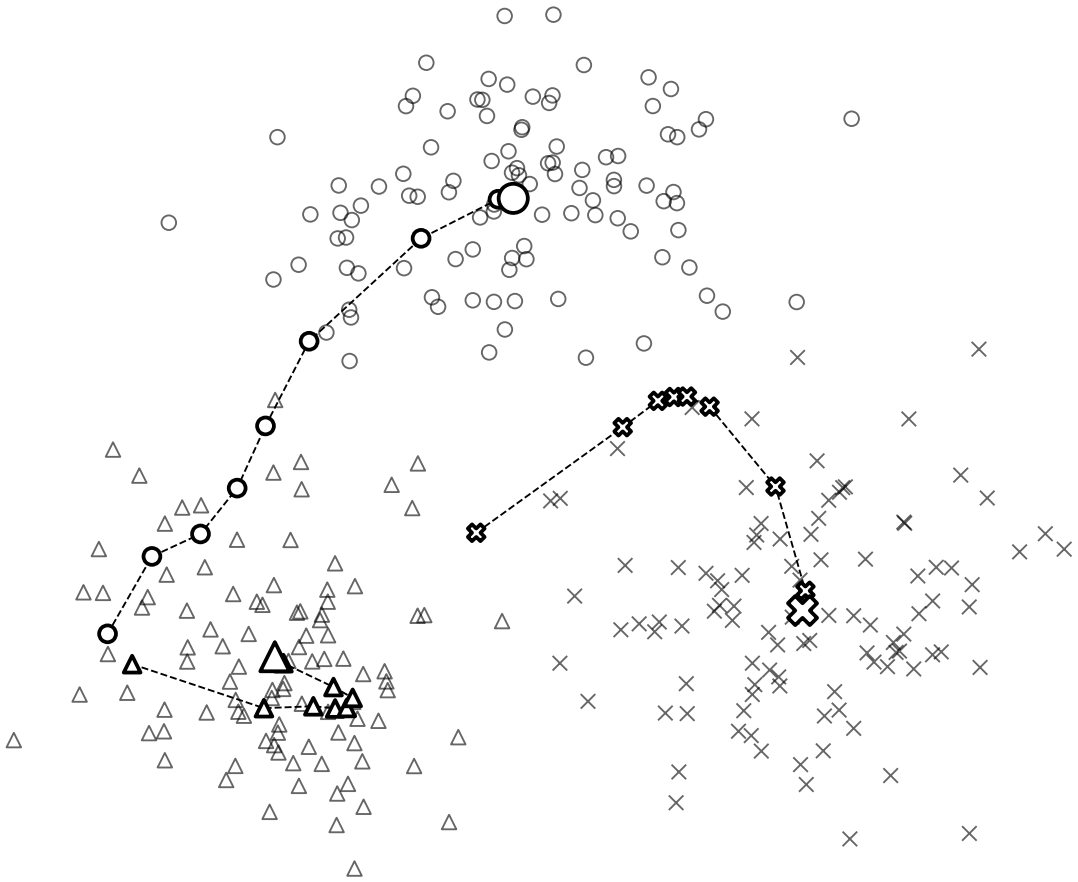
given $x_1, \ldots, x_m \in \mathbf{R}^n$; initial $\tilde{\mu}_1, \ldots, \tilde{\mu}_k \in \mathbf{R}^n$.
repeat

1. *Cluster assignment.* $\tilde{c}_i := \mathrm{argmin}_{j \in \{1,\ldots,k\}} \|x_i - \tilde{\mu}_j\|_2^2$ for $i = 1, \ldots, m$.
2. *Group formation.* $G_j := \{i \in \{1, \ldots, m\} \mid \tilde{c}_i = j\}$ for $j = 1, \ldots, k$.
3. *Representative update.* $\tilde{\mu}_j := (1/|G_j|) \sum_{i \in G_j} x_i$ for $j = 1, \ldots, k$.

until convergence.

---

numerical example:



- path shows results from each iteration
- final cluster representatives shown larger

Clustering with prior information

assignment and relation constraints:

- x_i must in one of the clusters in $S_i \subseteq \{1, \ldots, k\}$:

$$z_{ij} = 0, \quad j \in \{1, \ldots, k\} \setminus S_i$$

- *must-link* constraint: x_p, x_q must in the same cluster,

$$z_p = z_q \quad \Longleftrightarrow \quad z_{pj} = z_{qj}, \quad j = 1, \ldots, k$$

- *cannot-link* constraint: x_p, x_q must in different clusters,

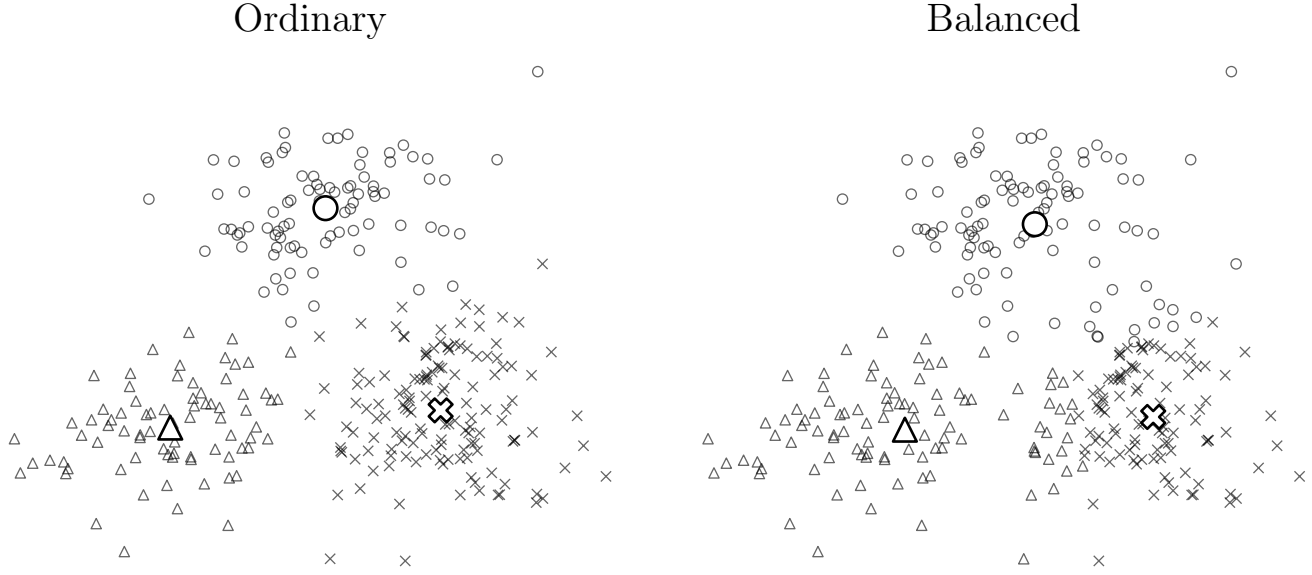
$$z_{pj} + z_{qj} \leq 1, \quad j = 1, \ldots, k$$

  requires only one of z_{pj} and z_{qj} can be nonzero

size constraints: define size measurement $s \in \mathbf{R}^k$ as

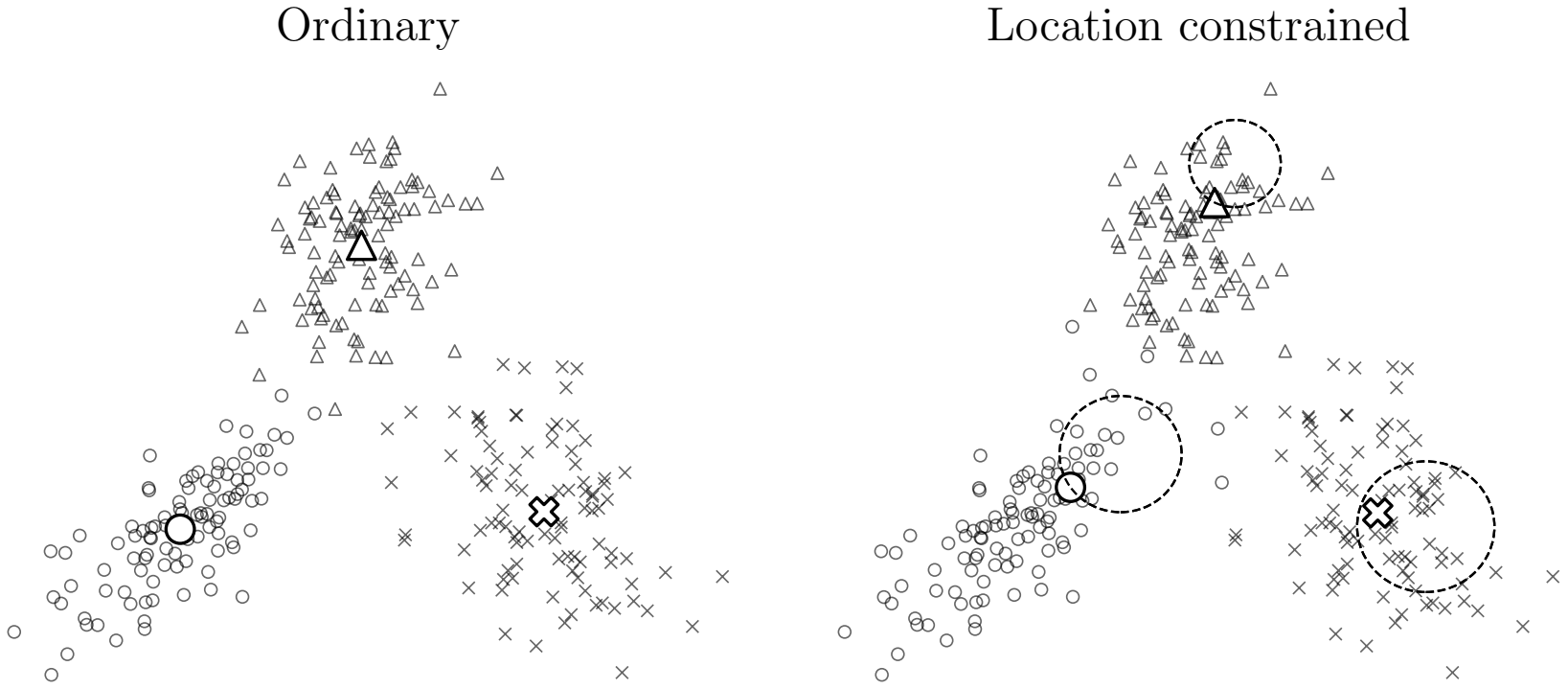
$$s_j = \sum_{i=1}^{m} z_{ij}, \quad j = 1, \ldots, k$$

- bounded size for each cluster: $L_j \leq s_j \leq U_j$, $j = 1, \ldots, k$
- balance constraint: $|s_j - m/k| \leq \epsilon$ for $j = 1, \ldots, k$ (with small $\epsilon > 0$) requires each cluster to have roughly the same number of points



geometric constraints on cluster representatives μ_i:

- polyhedron: $A_i\mu_i \preceq b_i$, $i = 1, \ldots, k$
- norm ball with center c_i and radius r_i: $\|\mu_i - c_i\|_2 \leq r_i$, $i = 1, \ldots, k$



Principal component analysis (PCA)

data matrix $A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbf{R}^{m \times n}$:

- rows $a_i \in \mathbf{R}^n$: (feature vector of) ith data point
- jth column is jth feature across all m data points
- often assume $m \gg n$, *i.e.*, more data points than features

standardization:

$$\tilde{A} = (A - \mathbf{1}\mu^T)D^{-1}$$

- μ_i is mean of ith feature for $i = 1, \ldots, n$
- $D = \mathbf{diag}(\sigma_1, \ldots, \sigma_n)$ with σ_i the standard deviation of ith feature
- often assume A is already standardized, *i.e.*, $\mu = 0$ and $D = I$

PCA finds **low rank approximation** Z of A:

$$\begin{array}{ll} \text{minimize} & \|A - Z\|_F^2 \\ \text{subject to} & \mathbf{rank}\, Z \leq k \end{array}$$

- componentwise quadratic residual matrix penalty:

$$\|A - Z\|_F^2 = \sum_{i=1}^{m} \|a_i - z_i\|_2^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - z_{ij})^2$$

- equivalent to **factorized form**:

$$\text{minimize} \quad \|A - XY\|_F^2$$

  by change of variables $Z = XY$ with $X \in \mathbf{R}^{m \times k}$ and $Y \in \mathbf{R}^{k \times n}$

  – implicit rank constraint since $\mathbf{rank}(XY) \leq k$

Solution via singular value decomposition

PCA is nonconvex problem but has analytical solution via SVD:

$$A = U\Sigma V^T$$

- $U \in \mathbf{R}^{m\times r}$, $V \in \mathbf{R}^{n\times r}$ have orthonormal columns
- $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_r)$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of A
- PCA solution is then given by

$$Z^\star = U_k\Sigma_k V_k^T$$

  with U_k, Σ_k, V_k the first k columns of U, Σ, V

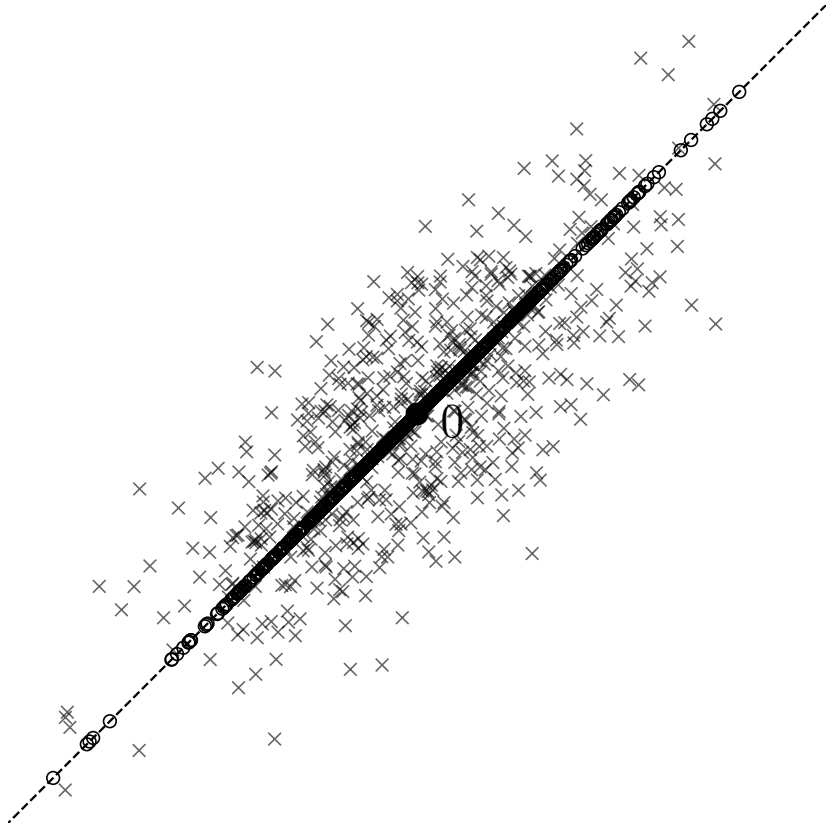
- for factorized form, *a* solution is given by

$$X^\star = U_k\Sigma_k^{1/2}, \qquad Y^\star = \Sigma_k^{1/2}V_k^T$$

  – for any invertible $W \in \mathbf{R}^{k\times k}$, $(X^\star W,\ W^{-1}Y^\star)$ is also a solution

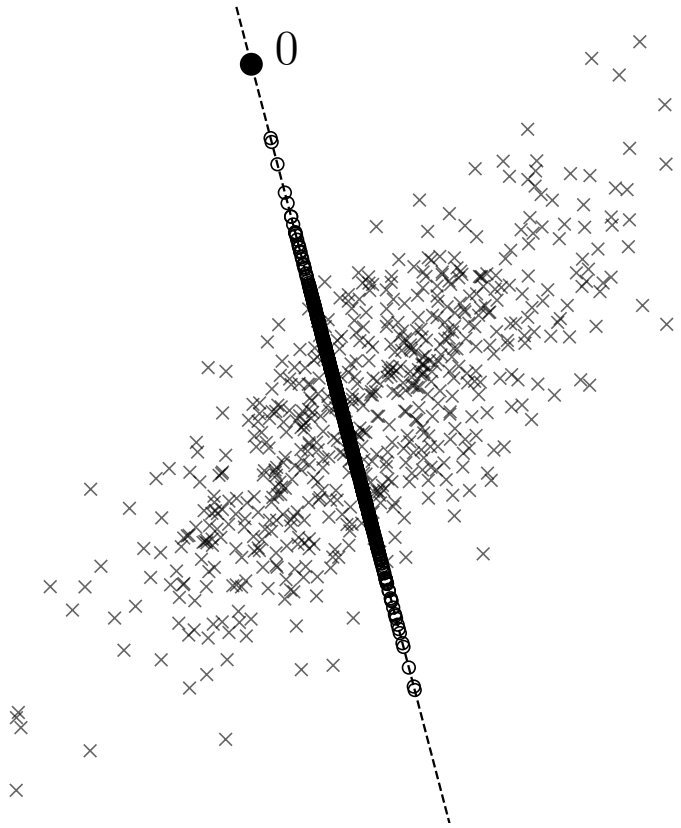
Interpretations

geometric interpretation: objective $\|A - Z\|_F^2 = \sum_{i=1}^m \|a_i - z_i\|_2^2$ finds z_i in a k-dimensional subspace of $\mathbf{R}^n$ that has the smallest sum of squared distances to the data points a_i



- A must be centered, $i.e.$, each column has zero mean
- due to projection only onto some subspace, not any affine set

example of PCA on a noncentered data matrix:



statistical interpretation:

- assuming a_i have zero mean and unit variance (so A standardized)
- find k orthogonal directions in $\mathbf{R}^n$
- such that projections of data points onto these directions have the top k largest variance

MLE interpretation: assuming components a_{ij} of A generated from

$$a_{ij} = x_i^T y_j + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$$

- coefficients $x_i, y_j \in \mathbf{R}^k$ and noise ϵ_{ij} are unknown
- let $X = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}^T \in \mathbf{R}^{m \times k}$, $Y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix} \in \mathbf{R}^{k \times n}$
- assume IID $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$

so log-likelihood is $l(X, Y) = -\dfrac{mn}{2} \log(2\pi\sigma^2) - \dfrac{1}{2\sigma^2} \|A - XY\|_F^2$

latent factor estimation: compress n features into $k < n$ features

- rows x_i^T of X: latent factors of a_i, $i.e.$, representation in new features
- columns y_i of Y: linear maps of latent factors to original features

archetype representation:

- rows $\tilde{y}_i^T$ of Y: archetypes of the data points, $i.e.$, representative features
- rows x_i^T of X: coefficients of the archetypes for each data point

  $i.e.$, *archetype representation* of the data points in terms of $\tilde{y}_i$

Quadratically regularized PCA

$$\text{minimize} \quad \|A - XY\|_F^2 + \gamma(\|X\|_F^2 + \|Y\|_F^2)$$

- variables $X \in \mathbf{R}^{m\times k}$, $Y \in \mathbf{R}^{k\times n}$; $\gamma > 0$ is regularization parameter
- equivalent to the vector form

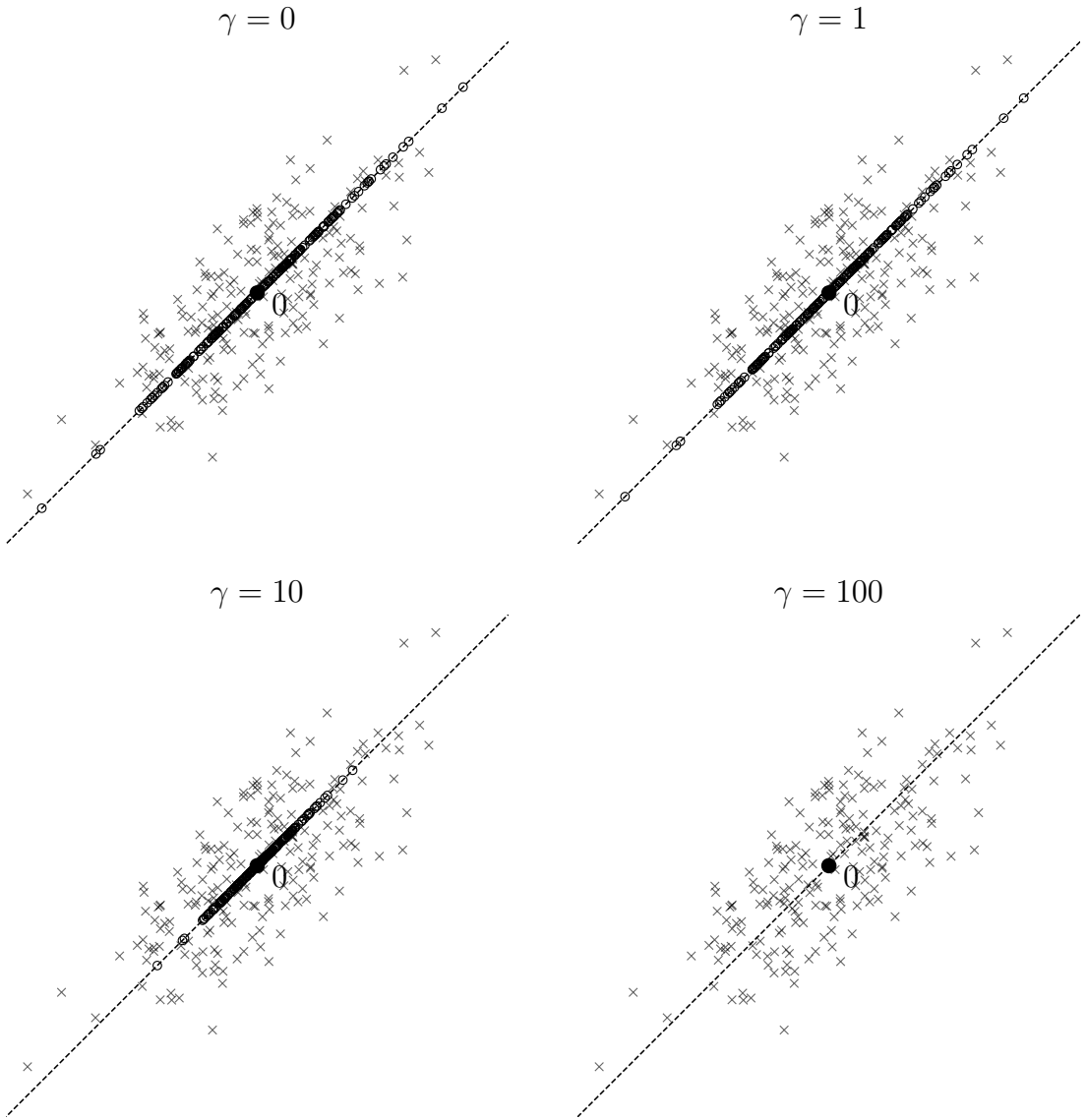
$$\text{minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} - x_i^T y_j)^2 + \gamma\left(\sum_{i=1}^{m}\|x_i\|_2^2 + \sum_{j=1}^{n}\|y_j\|_2^2\right)$$

- analytical solution via SVD: $X^\star = U_k\tilde{\Sigma}_k^{1/2}$ and $Y^\star = \tilde{\Sigma}_k^{1/2}V_k^T$, where $A = U\Sigma V^T$; U_k, V_k are first k columns of U, V, and

$$\tilde{\Sigma}_k = \begin{bmatrix} (\sigma_1 - \gamma)_+ & & 0 \\ & \ddots & \\ 0 & & (\sigma_k - \gamma)_+ \end{bmatrix} \in \mathbf{R}^{k\times k}$$

- still nonunique solutions but bounded

geometric interpretation: project onto a k-dimensional subspace, then shrink coefficients along singular vectors using γ



MAP estimation interpretation:

- assume linear model $a_{ij} = x_i^T y_j + \epsilon_{ij}$ with IID $\epsilon_{ij} \sim \mathcal{N}(0, 1)$
- assume componentwise IID priors $x_{ij}, y_{ij} \sim \mathcal{N}(0, 1/\gamma)$
  (so regularization weight interpreted as inverse variance $\gamma = 1/\sigma^2$)

Robust low rank approximation

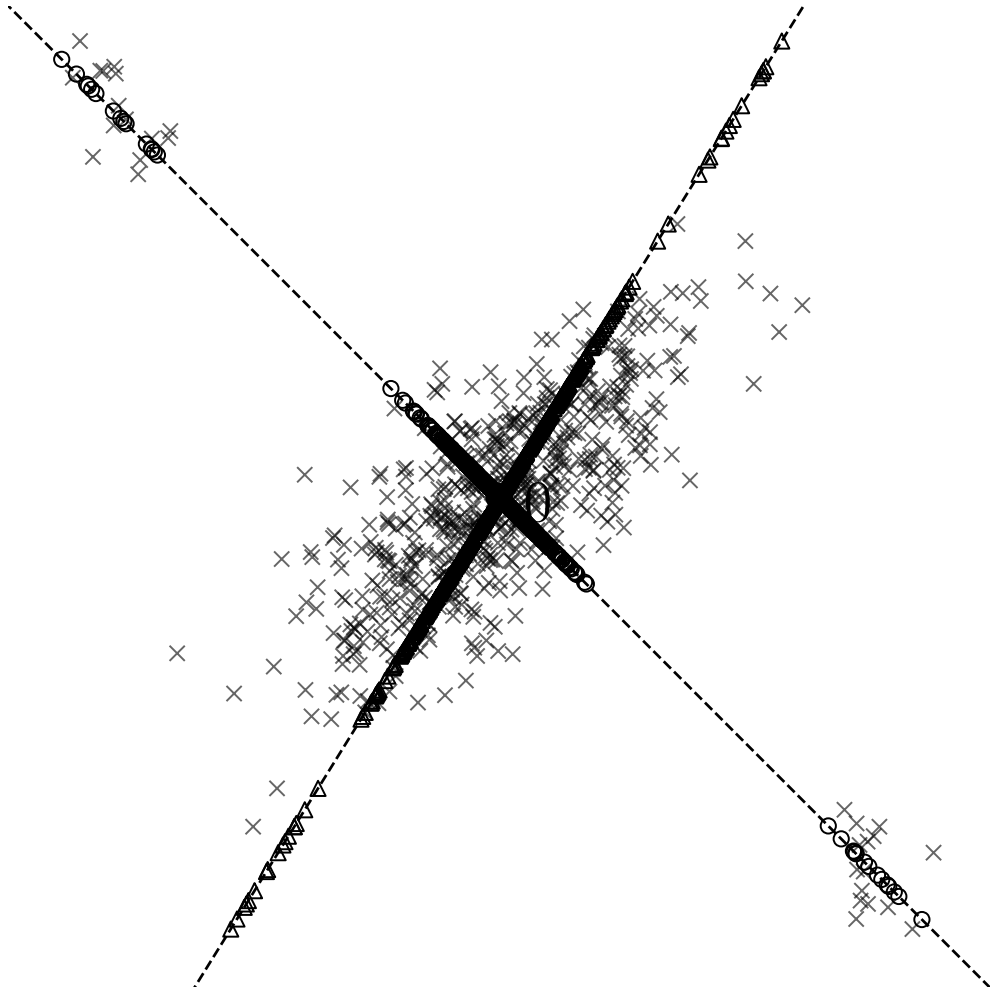
$$\begin{array}{ll} \text{minimize} & \|A - Z\|_{\text{sav}} = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij} - z_{ij}| \\ \text{subject to} & \mathbf{rank}\, Z \leq k \end{array}$$

- componentwise absolute value penalty on residual matrix
- less sensitive to outliers than ordinary PCA
- no analytical solution
- equivalent to factorized form

$$\text{minimize} \quad \|A - XY\|_{\text{sav}} = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij} - x_i^T y_j|$$

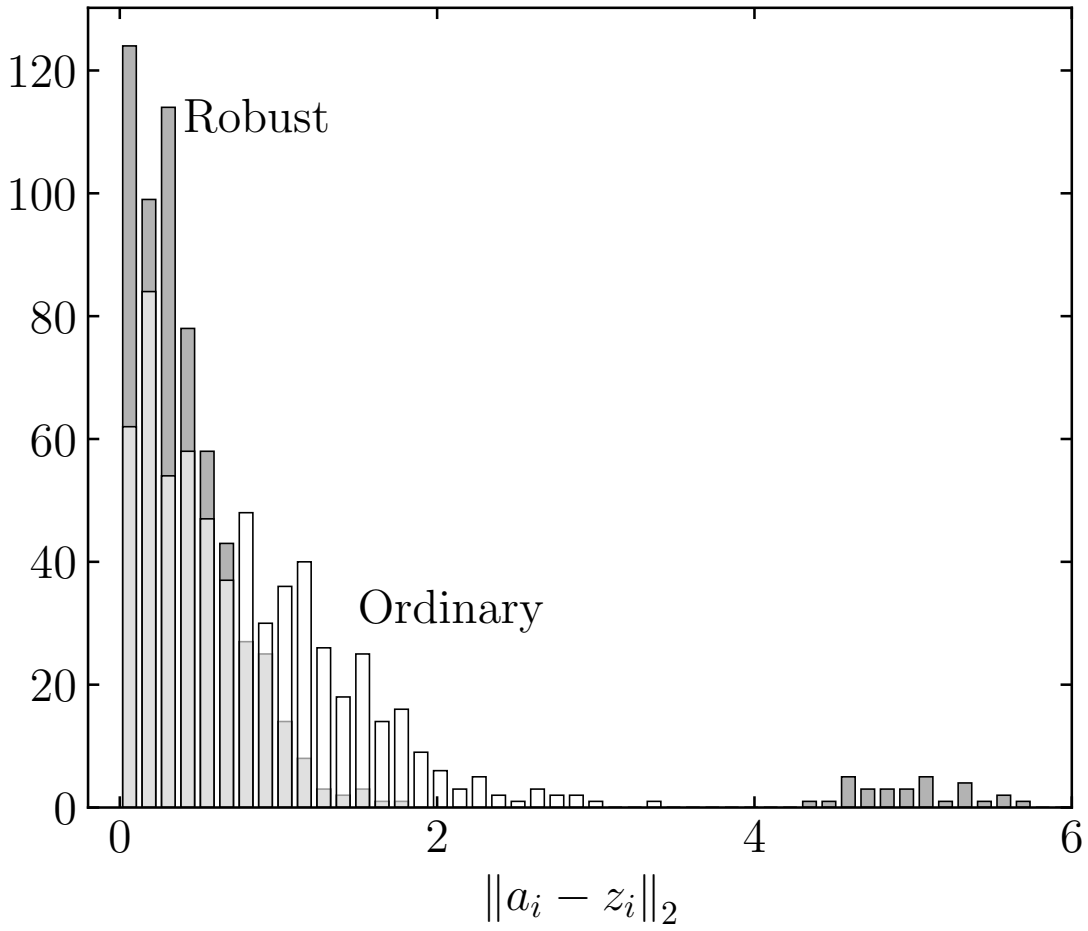
so can be solved via ACS

numerical example: data points (with outliers) in crosses



ordinary PCA in circles; robust approximation in triangles

histogram of distances between data points and approximation



robust PCA:

$$\text{minimize} \quad \|A - XY\|_{\text{sav}} + \gamma(\|X\|_F^2 + \|Y\|_F^2)$$

- equivalent to

$$\begin{array}{ll} \text{minimize} & \|A - Z\|_{\text{sav}} + 2\gamma\|Z\|_* \\ \text{subject to} & \mathbf{rank}\, Z \leq k \end{array}$$

- removing rank constraint gives convex problem

$$\begin{array}{ll} \text{minimize} & \|S\|_{\text{sav}} + 2\gamma\|Z\|_* \\ \text{subject to} & S + Z = A \end{array}$$

  - S denotes the sparse component of A ($e.g.$, outliers)
  - Z denotes the low rank component of A

Structural regularization and constraints

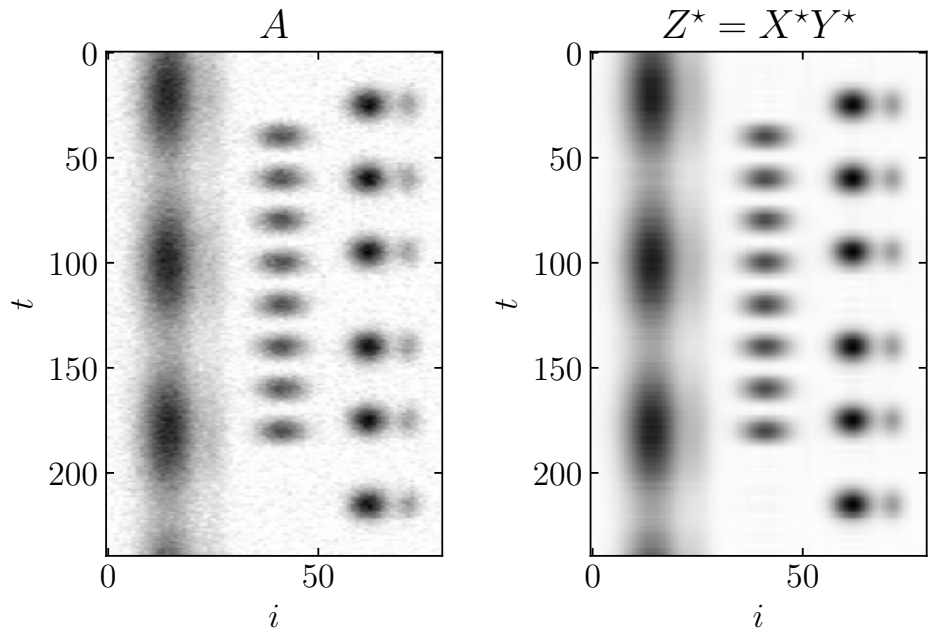
nonnegative matrix factorization (NMF):

$$\begin{array}{ll} \text{minimize} & \|A - XY\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n \left(a_{ij} - x_i^T y_j\right)^2 \\ \text{subject to} & x_i \succeq 0, \quad i = 1, \ldots, m \\ & y_i \succeq 0, \quad i = 1, \ldots, n \end{array}$$

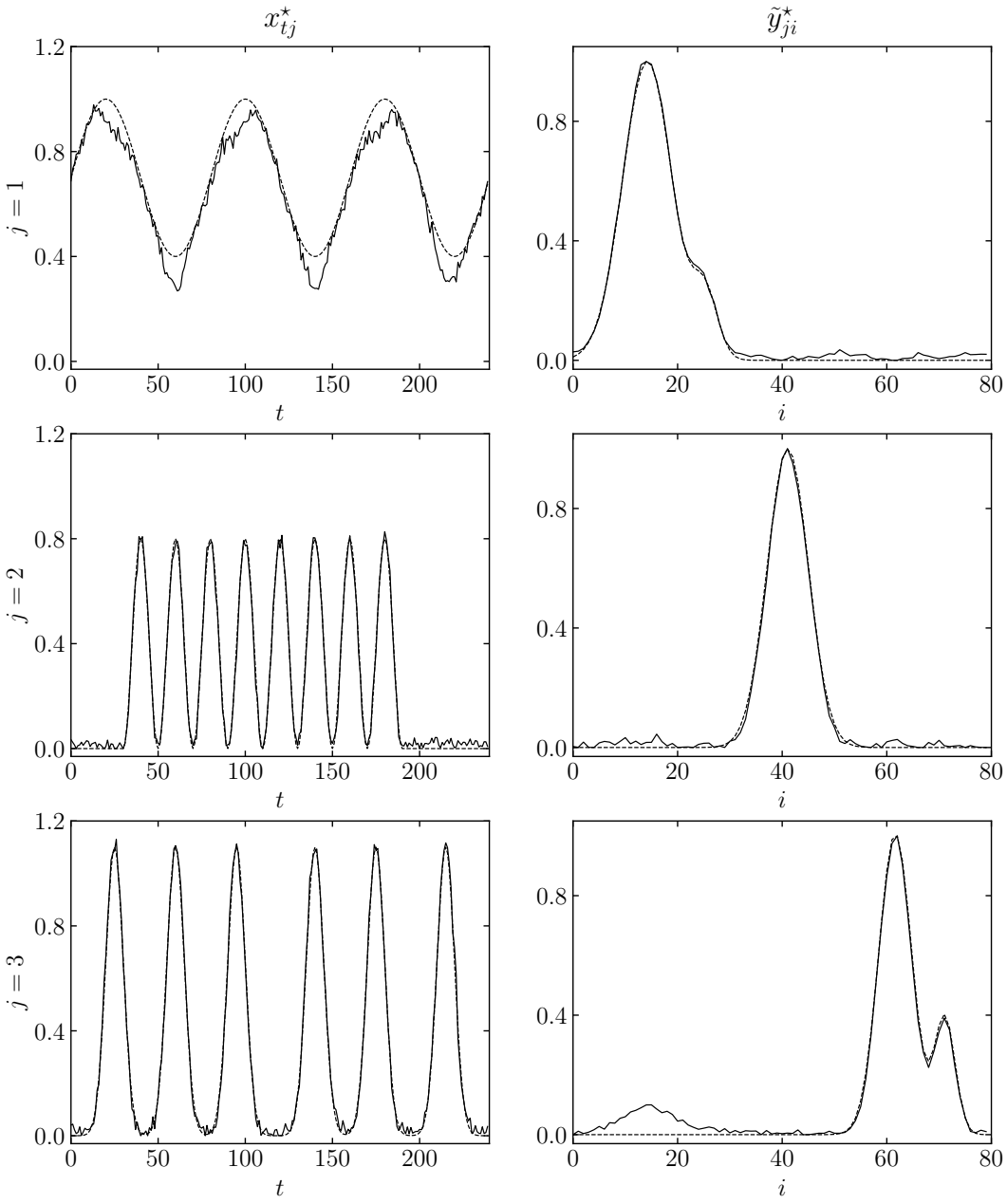
- no analytical solution, but can be solved via ACS

example:

- (corrupted) data $A \in \mathbf{R}_+^{m \times n}$
- m: number of time steps
- n: number of features
- NMF result $Z^\star = X^\star Y^\star$ with $X^\star \in \mathbf{R}_+^{m \times 3}$ and $Y^\star \in \mathbf{R}_+^{3 \times n}$



- right: archetypes
- left: coefficients for each time step of archetypes



sparse PCA:

$$\text{minimize} \quad \|A - XY\|_F^2 + \gamma \sum_{i=1}^m \|x_i\|_1$$

- latent factors x_i encouraged to be sparse
- or each point is represented by a few archetypes

example: *dictionary learning* problem

$$\begin{array}{ll} \text{minimize} & \|A - XY\|_F^2 + \gamma \|X\|_{\text{sav}} \\ \text{subject to} & \|Y\|_F^2 \leq \beta \end{array}$$

- Y called *dictionary*, $i.e.$, bounded set of archetypes

feature selection:

$$\text{minimize} \quad \|A - XY\|_F^2 + \gamma \sum_{i=1}^n \|y_i\|_2$$

- encourage columnwise sparsity in Y, $i.e.$, only a few of original n features are used in the archetypes