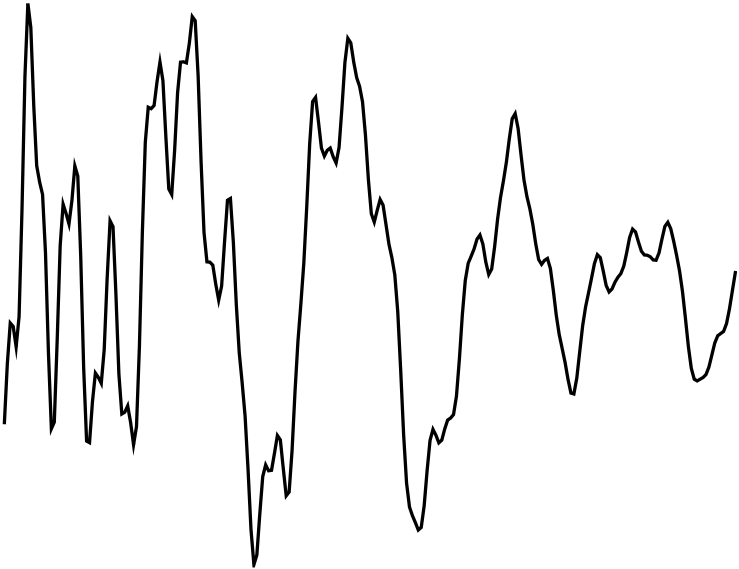
Sparsity in system identification and data-driven control

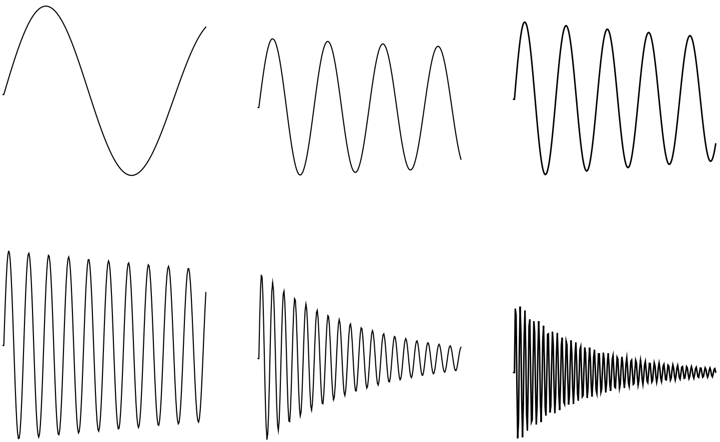
Ivan Markovsky

VRIJE UNIVERSITEIT BRUSSEL

This signal is not sparse in the "time domain"



But it is sparse in the "frequency domain" (it is weighted sum of six damped sines)



Problem: find sparse representation (small number of basis signals)

existence

representation

approximation

System theory offers alternative methods based on low-rank approximation

$$\text{rank of } \begin{bmatrix} y(1) & y(2) & y(3) & \cdots \\ y(2) & y(3) & y(4) & \cdots \\ y(3) & y(4) & y(5) & \cdots \\ \vdots & \vdots & \vdots & \\ y(L) & y(L+1) & y(L+3) & \cdots \end{bmatrix} \leq 12$$

Plan

Sparse signals and linear-time invariant systems

System identification as sparse approximation

Solution methods and generalizations

Sum-of-damped-exponentials signals are solutions of linear constant coefficient ODE

$$y = \alpha_1 \exp_{z_1} + \cdots + \alpha_{\mathrm{n}} \exp_{z_{\mathrm{n}}} \qquad \exp_z(t) := z^t$$

$$\Updownarrow$$

$$p_0 y + p_1 \sigma y + \cdots + p_{\mathrm{n}} \sigma^{\mathrm{n}} y = 0 \quad (\sigma y)(t) := y(t+1)$$

$$\Updownarrow$$

$$y = Cx,\ \sigma x = Ax \qquad x(t) \in \mathbb{R}^{\mathrm{n}} \text{ — state}$$

The solution set of linear constant coefficient ODE is linear time-invariant (LTI) system

$\mathtt{n}$-th order autonomous LTI system

$$\mathscr{B} := \{\, y = Cx \mid \sigma x = Ax,\ x(0) \in \mathbb{R}^{\mathtt{n}} \,\}$$

$\dim(\mathscr{B}) = \mathtt{n}$ — complexity of $\mathscr{B}$

$\mathscr{L}_{\mathtt{n}}$ — LTI systems with order $\leq \mathtt{n}$

$y \in \mathscr{B} \in \mathscr{L}_{\mathtt{n}}$ is constrained/structured/sparse

belongs to $\mathtt{n}$-dimensional subspace

is linear combination of $\mathtt{n}$ signals

described by $2\mathtt{n}$ parameters

We assume that sparse representation exists, but we do not know the basis

classical definition of sparse signal y

- y has a few nonzero values
  (we don't know which ones)
- basis: unit vectors

$y \in \mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ with $\mathrm{n} \ll$ # of samples

- y is sum of a few damped sines
  (their frequencies and dampings are unknown)
- basis: damped complex exponentials

The assumption $y \in \mathscr{B} \in \mathscr{L}_\mathrm{n}$ makes ill-posed problems well-posed

noise filtering

- given $y = \bar{y} + \widetilde{y}$, $\widetilde{y}$ — noise
- find $\bar{y}$ — true value

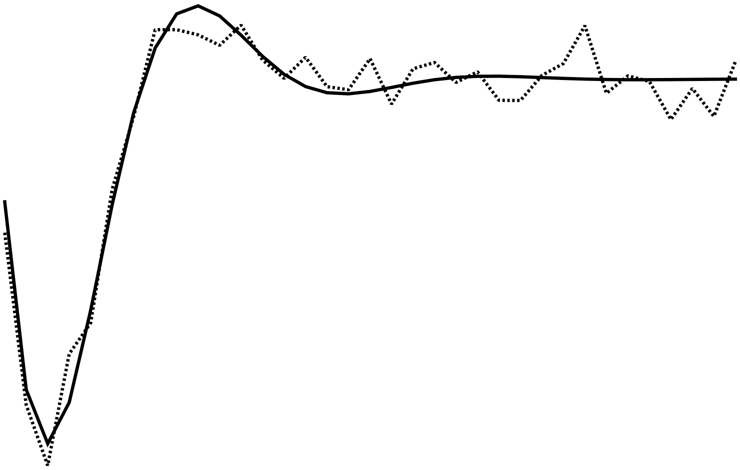
forecasting

- given "past" samples $\big(y(-t+1), \dots, y(0)\big)$
- find "future" samples $\big(y(1), \dots, y(t)\big)$

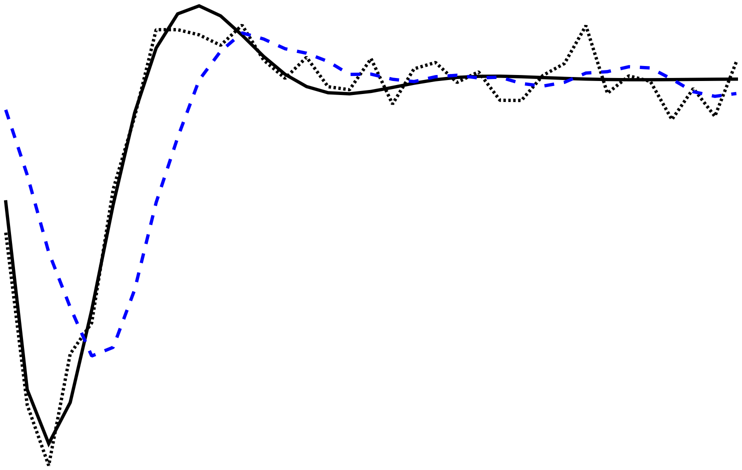
missing data estimation

- given samples $y(t)$, $t \in \mathscr{T}_{\text{given}}$
- find missing samples $y(t)$, $t \in \overline{\mathscr{T}_{\text{given}}}$

Noise filtering: given $y = \bar{y} + \widetilde{y}$, find $\bar{y}$ with prior knowledge $\bar{y} \in \bar{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}$, $\widetilde{y} \sim N(0, \nu I)$

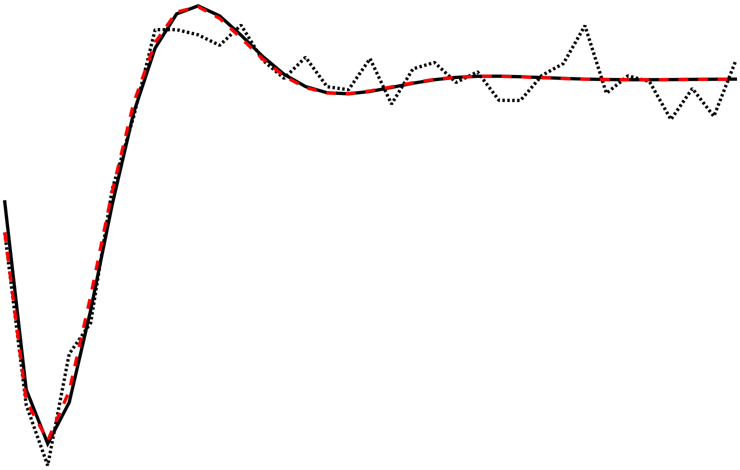


Heuristic: smooth the data by low-pass filter

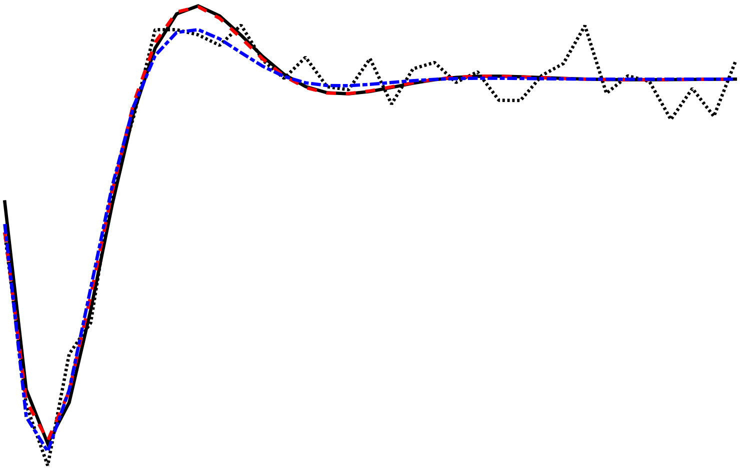


Optimal/Kalman filtering requires a model

The best (but unrealistic) option is to use $\bar{\mathscr{B}}$



Kalman filtering using identified model $\widehat{\mathscr{B}}$, ($i.e.$, prior knowledge $\bar{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}$)



Summary

the assumption $y \in \mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ imposes sparsity

the basis is sum-of-damped-exponentials
with unknown dampings and frequencies

$y \in \mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ "regularizes" ill-posed problems

Plan

Sparse signals and linear-time invariant systems

System identification as sparse approximation

Solution methods and generalizations

System identification is an inverse problem

simulation $\mathscr{B} \mapsto y$

- given model $\mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ and initial conditions
- find the response $y \in \mathscr{B}$

identification $y \mapsto \mathscr{B}$

- given response y and model class $\mathscr{L}_{\mathrm{n}}$
- find model $\mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ that "fits well" y

"fits well" is often defined in stochastic setting

assumption $y = \bar{y} + \widetilde{y}$ where

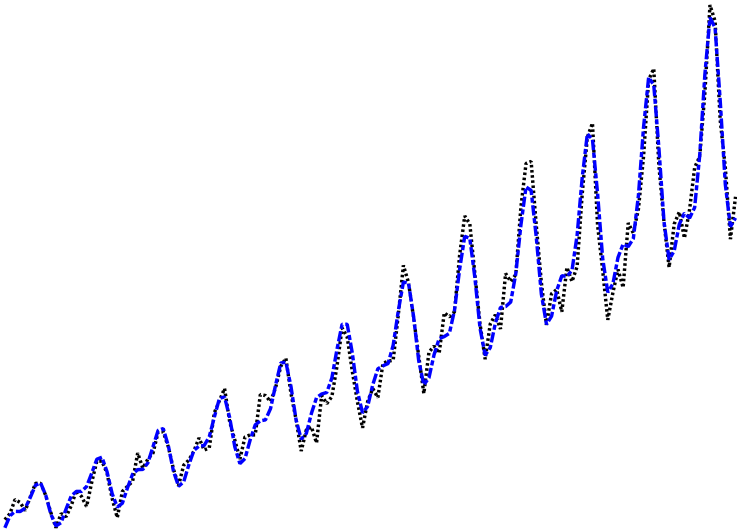
- $\bar{y} \in \bar{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}$ is the true signal
- $\widetilde{y} \sim N(0, \nu I)$ is noise (zero mean white Gaussian)

maximum likelihood estimator

$$\text{minimize} \quad \text{over } \widehat{y} \text{ and } \widehat{\mathscr{B}} \quad \|y - \widehat{y}\|$$
$$\text{subject to} \quad \widehat{y} \in \widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}$$

"The noise model is just an alibi for determining the cost function." L. Ljung

Example: monthly airline passenger data 1949–1960 fit by 6th order LTI model



How well a given model $\mathscr{B}$ fits the data y?

$$\text{error}(y, \mathscr{B}) := \min_{\widehat{y} \in \mathscr{B}} \|y - \widehat{y}\|$$

- likelihood of y, given $\mathscr{B}$
- projection of y on $\mathscr{B}$
- validation error

identification problem:

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}} \quad \text{error}(y, \mathscr{B})$$

The link between system identification and sparse approximation is low rank

$$y \in \mathscr{B} \in \mathscr{L}_{\mathrm{n}}$$

$$\Updownarrow$$

$$\operatorname{rank}\left(\begin{bmatrix} y(1) & y(2) & \cdots & y(T-\mathrm{n}) \\ y(2) & y(3) & \cdots & y(T-\mathrm{n}+1) \\ \vdots & \vdots & & \vdots \\ y(\mathrm{n}+1) & y(\mathrm{n}+2) & \cdots & y(T) \end{bmatrix}\right) \leq \mathrm{n}$$

Hankel structured matrix $\mathscr{H}_{\mathrm{n}+1}(y)$

LTI system identification is equivalent to Hankel structured low-rank approximation

$$\text{minimize} \quad \text{over } \widehat{y} \text{ and } \widehat{\mathscr{B}} \quad \|y - \widehat{y}\|$$
$$\text{subject to} \quad \widehat{y} \in \widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}$$

$$\Updownarrow$$

$$\text{minimize} \quad \text{over } \widehat{y} \quad \|y - \widehat{y}\|$$
$$\text{subject to} \quad \operatorname{rank}\big(\mathscr{H}_{\mathrm{n}+1}(\widehat{y})\big) \leq \mathrm{n}$$

Summary

system identification aims at a map $y \mapsto \mathscr{B}$

the map is defined through optimization problem

equivalent problem: Hankel low-rank approx.
(impose sparsity on the singular values)

Plan

Sparse signals and linear-time invariant systems

System identification as sparse approximation

Solution methods and generalizations

Three solution approaches:

nuclear norm heuristic

subspace methods

local optimization

The nuclear norm heuristic induces sparsity on the singular values

rank: number of nonzero singular values

$\|\cdot\|_*$: ℓ_1-norm of the singular values vector

minimization of the nuclear norm

- tends to increase sparsity $\implies$ reduce rank
- leads to a convex optimization problem

Nuclear norm minimization methods involve a hyper-parameter

$$\begin{array}{ll} \text{minimize} \quad \text{over } \widehat{y} & \|y - \widehat{y}\| \\ \text{subject to} & \|\mathscr{H}_{\mathrm{n}+1}(\widehat{y})\|_* \leq \gamma \end{array}$$

$$\Updownarrow$$

$$\text{minimize} \quad \text{over } \widehat{y} \quad \alpha \|y - \widehat{y}\| + \|\mathscr{H}_{\mathrm{n}+1}(\widehat{y})\|_*$$

γ/α — determines the rank of $\mathscr{H}_{\mathrm{n}+1}(\widehat{y})$

we want $\alpha_{\text{opt}} = \max\{\, \alpha \mid \text{rank}\left(\mathscr{H}_{\mathrm{n}+1}(\widehat{y})\right) \leq \mathrm{n} \,\}$

α_{opt} can be found by bijection

Originally the subspace identification methods were developed for exact data

$\mathscr{L}_{\mathrm{n}}$ — class of LTI systems of order $\leq \mathrm{n}$

state space representation

$$\mathscr{B} := \{\, y = Cx \mid \sigma x = Ax,\ x(0) \in \mathbb{R}^{\mathrm{n}} \,\}$$

exact identification problem $y \mapsto (A, C)$

- given $y \in \mathscr{B} \in \mathscr{L}_{\mathrm{n}}$ — exact data
- find (A, C) — model parameters

Two steps solution method

1. rank revealing factorization

$$\mathscr{H}_L(y) = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{L+1} \end{bmatrix}}_{\mathscr{O}} \underbrace{\begin{bmatrix} x(0) & Ax(0) & A^2x(0) & \cdots & A^{T-L}x(0) \end{bmatrix}}_{\mathscr{C}}$$

2. shift equation

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{L-1} \end{bmatrix} A = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^{L} \end{bmatrix} \quad \Longleftrightarrow \quad \mathscr{O}(1{:}L-1,:)A = \mathscr{O}(2{:}L,:)$$

$T = 2\mathrm{n}+1$ samples suffice, $\quad L \in [\mathrm{n}+1, T-\mathrm{n}]$

For noisy data, subspace methods involve unstructured low-rank approximation

do steps 1 and 2 approximately:

1. singular value decomposition of $\mathscr{H}_L(y)$
2. least squares solution of the shift equation

L is hyper-parameter that affects the solution $\widehat{\mathscr{B}}$

Local optimization using variable projections

"double" optimization

$$\min_{\widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}} \left(\min_{\widehat{y} \in \widehat{\mathscr{B}}} \|y - \widehat{y}\| \right)$$

"inner" minimization

$$\text{error}(y, \widehat{\mathscr{B}}) = \|\Pi_{\widehat{\mathscr{B}}} y\|$$

"outer" minimization

$$\min_{\widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}} \text{error}(y, \widehat{\mathscr{B}})$$

Representation of an LTI system as kernel of polynomial operator

$$p_0 y + p_1 \sigma y + \cdots + p_\mathrm{n} \sigma^\mathrm{n} y = 0 \quad (\sigma y)(t) := y(t+1)$$

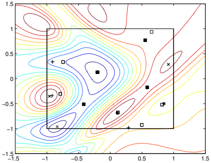
$p(\sigma) y = 0$, where $p(z) = p_0 + p_1 z + \cdots + p_\mathrm{n} z^\mathrm{n}$

model parameter $p = \begin{bmatrix} p_0 & p_1 & \cdots & p_\mathrm{n} \end{bmatrix}$

Parameter optimization problem

optimization over a manifold

$$\min_{\widehat{\mathscr{B}} \in \mathscr{L}_{\mathrm{n}}} \ \mathrm{error}(y, \widehat{\mathscr{B}}) \quad \Longleftrightarrow \quad \min_{\|p\|=1} \ \mathrm{error}(y, p)$$



optimization over Euclidean spaces

$$p \neq 0 \quad \Longleftrightarrow \quad p = \begin{bmatrix} x & 1 \end{bmatrix} \Pi$$

Π permutation

- Π fixed $\quad \rightsquigarrow \quad$ total least-squares
- Π can be changed during the optimization

Three generalizations

systems with inputs

missing data estimation

nonlinear system identification

Dealing with missing data

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{y} \quad \|y - \widehat{y}\|_v \\ \text{subject to} & \operatorname{rank}\big(\mathscr{H}_{\mathrm{n}+1}(\widehat{y})\big) \le \mathrm{n} \end{array}$$

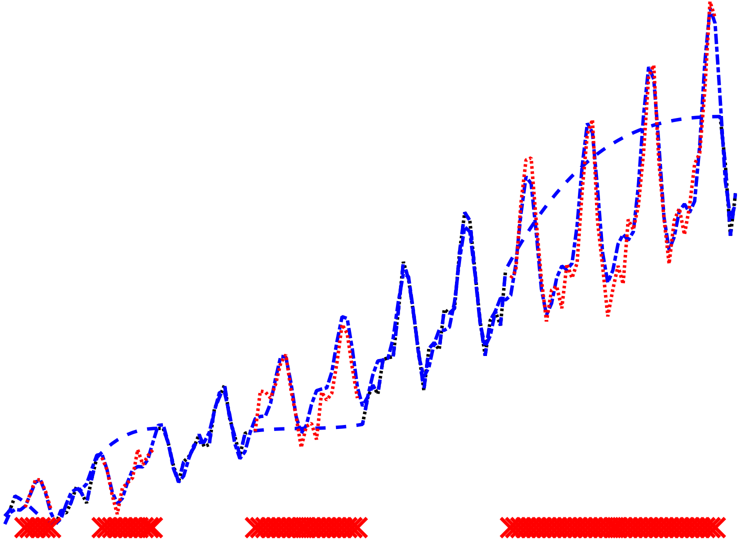
weighted 2-norm approximation

$$\|y - \widehat{y}\|_v := \sqrt{\textstyle\sum_{k,t} \; v^k(t) \big(y^k(t) - \widehat{y}^k(t)\big)^2}$$

with element-wise weights

$$\begin{array}{lll} v^k(t) \in (0, \infty) & \text{if } y^k(t) \text{ is noisy} & \text{approximate } y^k(t) \\ v^k(t) = 0 & \text{if } y^k(t) \text{ is missing} & \text{interpolate } y^k(t) \\ v^k(t) = \infty & \text{if } y^k(t) \text{ is exact} & \widehat{y}^k(t) = y^k(t) \end{array}$$

Example: piecewise cubic interpolation vs LTI identification on the "airline passenger data"



Conclusion

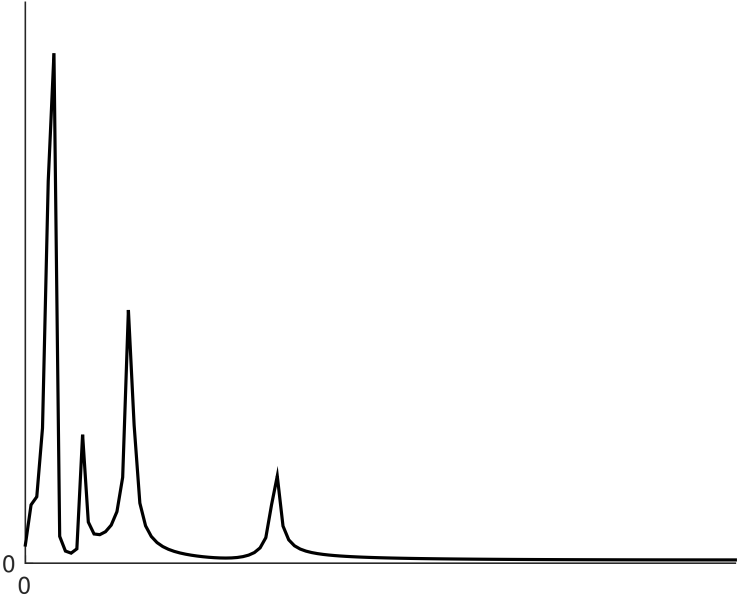
y is response of LTI system $\iff$ y sparse

LTI identification $\iff$ low-rank approx.

solution methods

- convex relaxation (nuclear norm)
- subspace (SVD + least squares)
- local optimization

DFT analysis suffers from the "leakage"



Gridding the frequency axis and using ℓ_1-norm minimization has limited resulution

given signal y

select "dictionary" $\Phi(t) = \begin{bmatrix} \sin(\omega_1 t) & \cdots & \sin(\omega_N t) \end{bmatrix}$

minimize over a $\quad \|a\|_1 \quad$ subject to $\quad y = \Phi a$