音響システム特論 第1回
Advanced Topics of Acoustic Systems Day 1

小山 翔一 / Shoichi Koyama

東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

Oct. 5, 2021

講義概要 / Summary

- 目的 / Goals
  - 音響信号を対象としたモデリングや最適化の方法論を学ぶことで，信号処理における発展的な数理的手法を，応用を通して理解することを目指す。
  - Gain understanding of advanced mathematical techniques in signal processing through applications by learning modeling and optimization methods for acoustics signals.
- 概要 / Summary
  - 前半は，音響信号を対象とした統計的推定法やアレイ信号処理，逆問題に関して講義する。具体的には，指向性制御，音源位置推定，空間音響，音場計測・制御などである。後半は，より発展的な内容に関して，主にゲスト講師による講義を行う。
  - The first half is lectures about statistical estimation, array signal processing, and inverse problems for acoustic signals. More specifically, beamforming, source localization, spatial audio, and sound field analysis and control. The second half is mainly talks by guest lecturers about more advanced topics.

講義概要 / Summary

- 火曜 2 限 オンライン講義 / Tuesday 2nd period, Online lecture
- キーワード / Keywords
  - 音響信号処理, 統計的信号処理, アレイ信号処理, スパースモデリング, 逆問題, 指向性制御, 音源位置推定, 空間音響, 音場計測, 音場制御
  - Acoustic signal processing, Statistical signal processing, Array signal processing, Sparse modeling, Inverse problems, Beamforming, Source localization, Spatial audio, Sound field analysis, Sound field control
- 講義は日本語で行いますが，スライドはできるだけ英語にします。 / Lectures are in Japanese, but slides are in English as much as possible.

講義日程 / Schedule

10/5	逆問題と統計的推定法 / Inverse Problem and Statistical Estimation
10/12	信号処理における最適化法 / Optimization in Signal Processing
10/19	休講 / Canceled
10/26	スパースモデリング / Sparse Modeling
11/2	アレイ信号処理 / Array Signal Processing
11/9	波動場のモデリング / Wavefield Modeling
11/16	空間音響の基礎 / Basics of Spatial Audio
11/30	[Guest] 高道慎之介先生 / Prof. Shinnosuke Takamichi
12/7	[Guest] 伊藤信貴先生 (新領域) / Prof. Nobutaka Ito (GSFS)
12/14	休講 / Canceled
12/21	[Guest] 井本桂右先生 (同志社大) / Prof. Keisuke Imoto (Doshisha)
1/4	音場の分析と合成 / Sound Field Analysis and Synthesis
1/11	[Guest] 猿渡洋先生 / Prof. Hiroshi Saruwatari

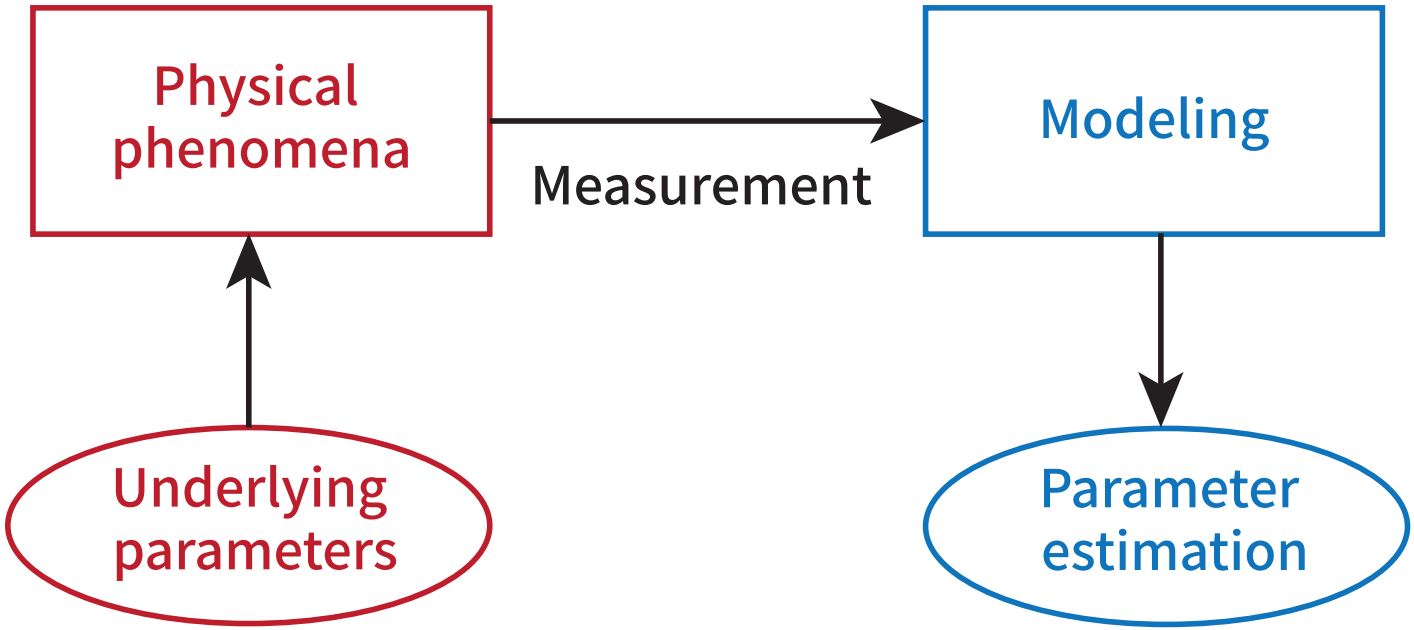
講義資料と成績評価 / Materials and Grades

- 講義資料 / Course materials
  - Downloadable at ITC-LMS or `http://www.sh01.org/ja/teaching/`
- 成績評価 / Grading system
  - レポート課題 (2 回) / Reporting assignments (twice)

Table of Contents

Table of Contents

Inverse Problem



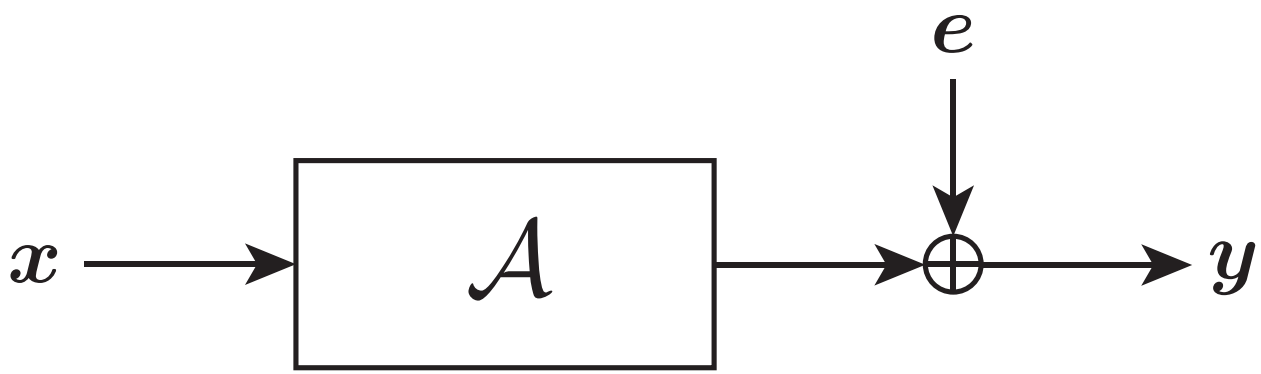
- Many engineering problems can be considered as parameter estimation from physical phenomena.
- This type of problems is generally called Inverse Problem (逆問題).

General measurement model

- Measurement $\boldsymbol{y} \in \mathbb{R}^M$ (or $\mathbb{C}^M$) is modeled as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{e}$$

  where $\boldsymbol{x} \in \mathbb{R}^N$ (or $\mathbb{C}^N$) is parameter to be estimated (model parameter), $\mathcal{A}$ is forward measurement operator, and $\boldsymbol{e} \in \mathbb{R}^M$ (or $\mathbb{C}^M$) is additive noise.
- Goal is to estimate $\boldsymbol{x}$ from $\boldsymbol{y}$. $\mathcal{A}$ is assumed to be given unless otherwise stated.



Examples of measurement model

- Denoising
  - $\boldsymbol{x}$ is clean speech, and $\boldsymbol{y}$ is noisy speech.
  - Measurement operator is finite-dimensional identity matrix $\boldsymbol{I}$

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}$$

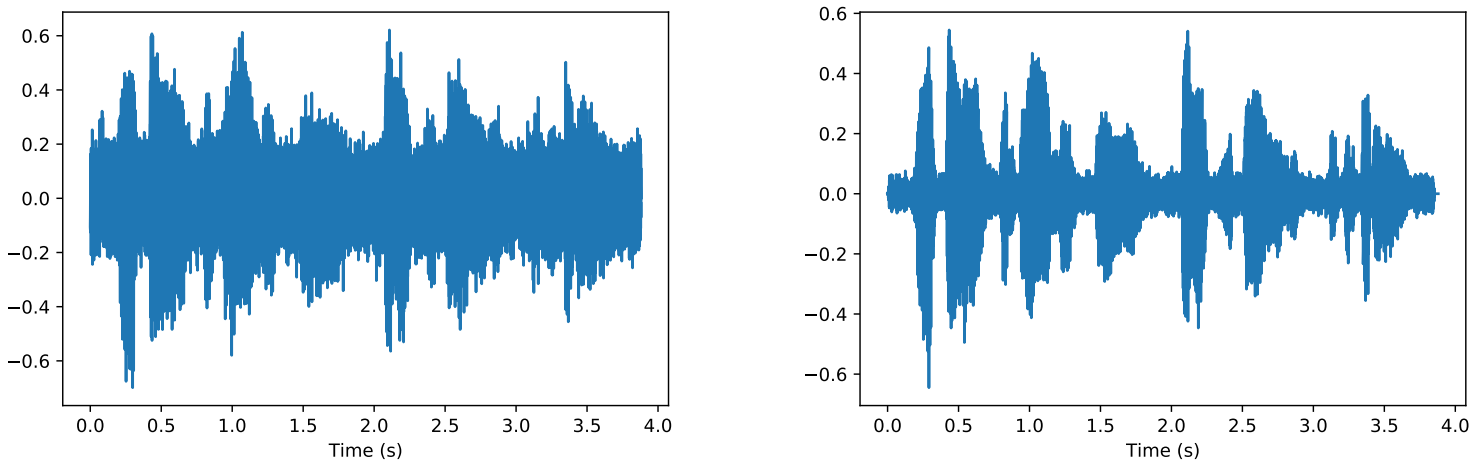


Figure: Left: Noisy speech, Right: Denoised speech

Examples of measurement model

- Fourier analysis
  - Commonly-used (nonparametric) spectral estimation

$$y_n = \sum_{k=1}^{K} a_k \cos\left(\frac{2\pi kn}{N}\right) + \sum_{k=1}^{K} b_k \sin\left(\frac{2\pi kn}{N}\right) + e_n$$

$$(n \in \mathbb{Z}([0, N-1]))$$

Then, Fourier coefficients $\{a_k\}_{k=1}^{K}$ and $\{b_k\}_{k=1}^{K}$ are model parameters, i.e., $\boldsymbol{x} = [a_1, \ldots, a_K, b_1, \ldots, b_K]^\mathsf{T}$.

  - Measurement operator becomes finite-dimensional matrix

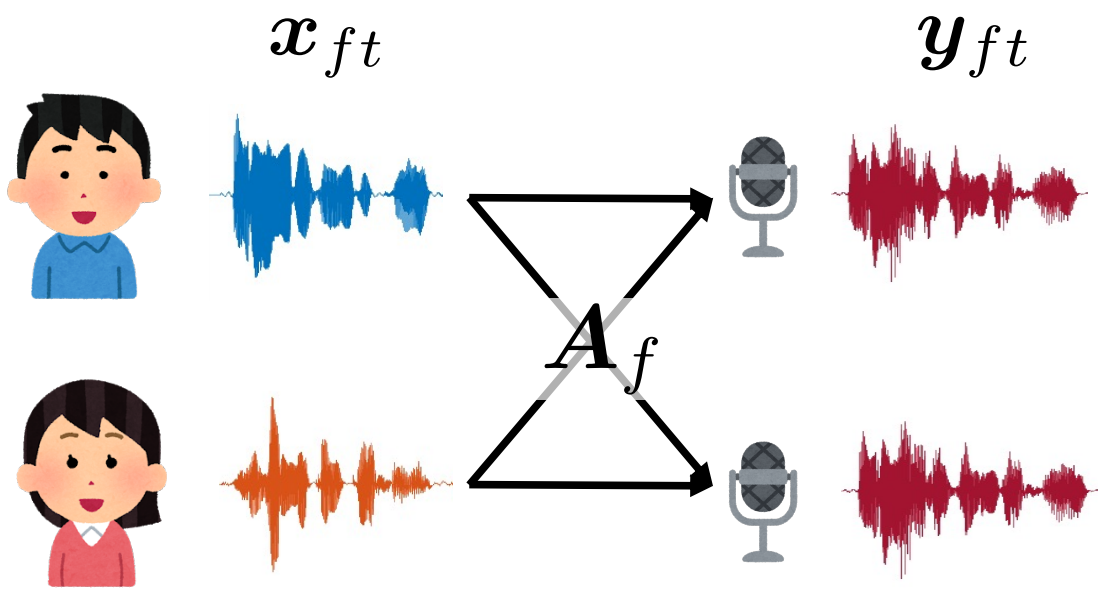
$$\boldsymbol{A} = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \cos\left(\frac{2\pi}{N}\right) & \cdots & \cos\left(\frac{2\pi K}{N}\right) & \sin\left(\frac{2\pi}{N}\right) & \cdots & \sin\left(\frac{2\pi K}{N}\right) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos\left(\frac{2\pi(N-1)}{N}\right) & \cdots & \cos\left(\frac{2\pi K(N-1)}{N}\right) & \sin\left(\frac{2\pi(N-1)}{N}\right) & \cdots & \sin\left(\frac{2\pi K(N-1)}{N}\right) \end{bmatrix}$$

Examples of measurement model

- Source separation
  - Time-frequency-domain source signals $\boldsymbol{x}$ is measured by microphones $\boldsymbol{y}$.
  - Measurement operator becomes finite-dimensional transfer function matrix (mixture matrix).

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$$

In blind source separation, $\boldsymbol{A}$ is assumed to be unknown.



Examples of measurement model

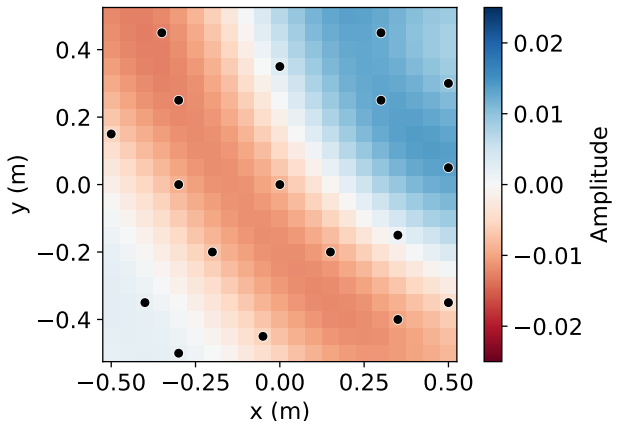
- Interpolation / Inpainting
  - True model parameter $x : \Omega \to \mathbb{R}$ is observed at fixed measurable set $\Omega_0 \subset \mathbb{R}^M$.

$$\boldsymbol{y} = x\big|_{\Omega_0} + \boldsymbol{e}$$

  - When model parameter is finite-dimensional vector $\boldsymbol{x}$, measurement operator becomes finite-dimensional matrix

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{x} + \boldsymbol{e}$$

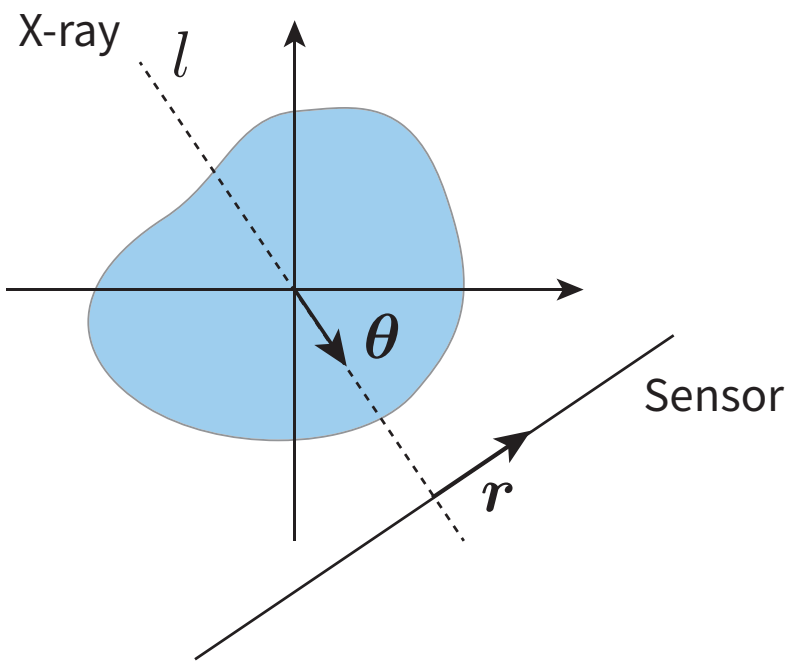
    Here, $\boldsymbol{S}$ is diagonal matrix where $\boldsymbol{S}_{n,n} = 1$ for sampled element and $\boldsymbol{S}_{n,n} = 0$ for not sampled element.



Examples of measurement model

- Computational tomography (CT)
  - X-ray traveling along line $l : s \to \boldsymbol{r} + s\boldsymbol{\theta}$ is detected by sensor, which is attenuated by material on l. Here, $\boldsymbol{\theta} \in \mathbb{S}^1$ and $\boldsymbol{r} \in \boldsymbol{\theta}^\perp$.
  - Model parameter is density of object $x : \mathbb{R}^2 \to \mathbb{R}$. Then, measurement operator is represented as

$$\mathcal{A}(x)(\boldsymbol{r}, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} x(\boldsymbol{r} + s\boldsymbol{\theta}) \mathrm{d}s$$

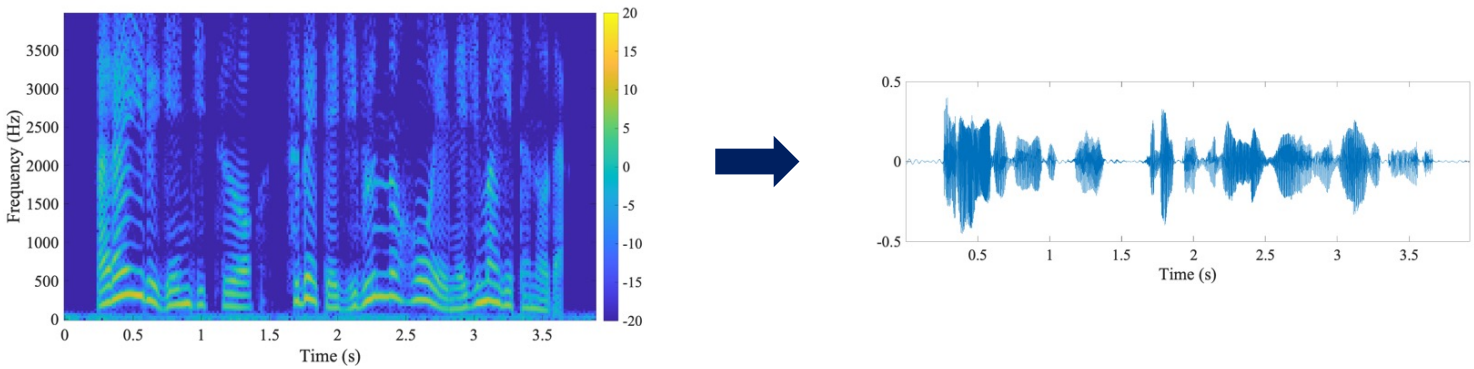


Examples of measurement model

- Phase retrieval
  - Model parameter is time-domain signal vector $\boldsymbol{x}$.
  - Goal is to estimate $\boldsymbol{x}$ only from magnitude spectrogram. Measurement operator becomes

$$\mathcal{A}(x) = |\boldsymbol{A}\boldsymbol{x}|$$

Here, $\boldsymbol{A}$ is STFT matrix, and $|\cdot|$ represents element-wise absolute value.



How to solve inverse problems?

- Well-posed problems must have i) *existence*, ii) *uniqueness*, and iii) *stability* of solution. Problems not having these properties are ill-posed problems.
- To estimate $\boldsymbol{x}$ from data $\boldsymbol{y}$ with given $\mathcal{A}$ in a stable manner, a measure of data discrepancy penalized using regularizer is generally minimized.

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathcal{D}(\mathcal{A}(\boldsymbol{x}), \boldsymbol{y}) + \mathcal{R}_{\boldsymbol{\theta}}(\boldsymbol{x})$$

  Here, $\mathcal{D}$ is data fidelity term and $\mathcal{R}_{\boldsymbol{\theta}}$ is regularization term with parameters $\boldsymbol{\theta}$.
- Various optimization algorithms can be applied, but *closed-form solution* and *linear estimator* are preferable in signal processing applications.

Notations

- Variables
  - x: scalar, $\boldsymbol{x}$: vector, $\boldsymbol{X}$: matrix (or tensor)
  - Subscripts indicate their indexes, e.g., $x_{i,j}$ is (i,j)th entry of matrix $\boldsymbol{X}$.
- Sets of real and complex numbers: $\mathbb{R}$ and $\mathbb{C}$
- Imaginary unit: $\mathrm{j} = \sqrt{-1}$, complex conjugate: $(\cdot)^*$
- Transpose: $(\cdot)^{\mathsf{T}}$, conjugate transpose: $(\cdot)^{\mathsf{H}}$

Table of Contents

Linear discrete model

- Suppose that N model parameters $\boldsymbol{x} = [x_1, \ldots, x_N]^\mathsf{T} \in \mathbb{R}^N$ and M observation data $\boldsymbol{y} = [y_1, \ldots, y_M]^\mathsf{T} \in \mathbb{R}^M$ are related by finite-dimensional measurement matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ as

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$$

  where $\boldsymbol{n} \in \mathbb{R}^M$ is vector of additive noise.
- This model is called *linear discrete model*.
- The problem to be solved is to estimate $\boldsymbol{x}$ from $\boldsymbol{y}$ with given $\boldsymbol{A}$.

Linear discrete model

- Assume additive noise $\boldsymbol{n}$ follows multivariate Gaussian distribution of mean $\boldsymbol{0}$ and covariance $\sigma^2 \boldsymbol{I}$, i.e., $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}|\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

$$p(\boldsymbol{n}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{n}\|^2\right)$$

- Since linear transformation of random variable from Gaussian distribution also normally distributed,

$$p(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2\right)$$

Maximum likelihood estimation

- Suppose $M > N$: *over-determined* (優決定) case
- Maximum likelihood (ML) estimation (最尤推定) of $\boldsymbol{x}$ from $\boldsymbol{y}$: estimate $\boldsymbol{x}$ so that the observed data $\boldsymbol{y}$ is most probable.
- Likelihood function (尤度関数) of $\boldsymbol{x}$, $\mathcal{L}(\boldsymbol{x})$, is defined as

$$\mathcal{L}(\boldsymbol{x}) = p(\boldsymbol{y})$$

- ML estimate $\hat{\boldsymbol{x}}_{\mathrm{ML}}$ is $\boldsymbol{x}$ that maximizes $\mathcal{L}(\boldsymbol{x})$:

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x})$$

- For ease in computation, negative log likelihood is minimized instead of direct maximization of $\mathcal{L}(\boldsymbol{x})$ in general.

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{x}} -\ln \mathcal{L}(\boldsymbol{x})$$

Note that log function is monotonically increasing function.

Maximum likelihood estimation

- Recall that the observation data is normally distributed:

$$p(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2\right)$$

- Then, the negative log likelihood becomes

$$-\ln \mathcal{L}(\boldsymbol{x}) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + C$$

  C denotes constant term that does not include $\boldsymbol{x}$.

- Therefore, ML estimate becomes

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$$

  This problem is equivalent to least-squares problem (最小二乗問題), i.e., minimization of square error between $\boldsymbol{y}$ and $\boldsymbol{A}\boldsymbol{x}$.

Maximum likelihood estimation

- ML estimate is obtained by minimizing the following cost function.

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x})$$
$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$$

- Since $\mathcal{J}$ is quadratic of $\boldsymbol{x}$, desired $\boldsymbol{x}$ satisfies $\partial\mathcal{J}/\partial\boldsymbol{x} = 0$.

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{x}} = \frac{\partial}{\partial \boldsymbol{x}}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 = -2\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) = 0$$

- Therefore, ML estimate is

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}$$

Maximum likelihood estimation

- When the noise is correlated, covariance matrix is not diagonal. Here, assume $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}|\boldsymbol{0}, \boldsymbol{\Sigma})$.

$$p(\boldsymbol{n}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{n}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{n}\right)$$

- ML estimate becomes

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathrm{ML}} &= \arg\min_{\boldsymbol{x}} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \\ &= (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\end{aligned}$$

Matrix differentiation

Inner product

$$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{a} = \boldsymbol{a}, \quad \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{a}^{\mathsf{T}} \boldsymbol{x} = \boldsymbol{a}$$

- Since $\boldsymbol{x}^{\mathsf{T}} \boldsymbol{a} = \sum_i x_i a_i$,

$$\frac{\partial}{\partial x_i} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{a} = a_i$$

- Since $\boldsymbol{a}^{\mathsf{T}} \boldsymbol{x} = \sum_i a_i x_i$,

$$\frac{\partial}{\partial x_i} \boldsymbol{a}^{\mathsf{T}} \boldsymbol{x} = a_i$$

Matrix differentiation

Quadratic form

$$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{x} = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathsf{T}} \right) \boldsymbol{x}$$

- Since $\boldsymbol{x}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{x} = \sum_i \sum_j a_{ij} x_i x_j$,

$$\frac{\partial}{\partial x_k} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{x} = \frac{\partial}{\partial x_k} \sum_i \sum_j a_{ij} x_i x_j = 2 a_{kk} x_k + \sum_{i \neq k} a_{ik} x_i + \sum_{j \neq k} a_{kj} x_j$$

$$= \sum_i a_{ik} x_i + \sum_j a_{kj} x_j$$

- Or,

$$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^{\mathsf{T}} \boldsymbol{A} \boldsymbol{x} = \frac{\partial (\boldsymbol{x}^{\mathsf{T}} \boldsymbol{A}) \boldsymbol{x}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{x}^{\mathsf{T}} (\boldsymbol{A} \boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{x} + \boldsymbol{A} \boldsymbol{x}$$

Matrix differentiation

- Trace of matrix also appears in signal processing / machine learning.

$$\mathrm{trace}(\boldsymbol{A}) = \sum_i a_{ii}$$

- Trace allows permutation of multiplication order:

$$\mathrm{trace}(\boldsymbol{AB}) = \mathrm{trace}(\boldsymbol{BA})$$
$$\mathrm{trace}(\boldsymbol{ABC}) = \mathrm{trace}(\boldsymbol{CAB}) = \mathrm{trace}(\boldsymbol{BCA})$$

- Element-wise representation:

$$\text{With } \boldsymbol{AB} = \sum_k a_{ik} b_{kj}, \quad \mathrm{trace}(\boldsymbol{AB}) = \sum_i \sum_j a_{ij} b_{ji}$$

Matrix differentiation

Trace

- Since $\frac{\partial}{\partial a_{ij}}\mathrm{trace}(\boldsymbol{AB}) = b_{ji}$,

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}(\boldsymbol{AB}) = \boldsymbol{B}^{\mathsf{T}}$$

- Since $\frac{\partial}{\partial a_{ij}}\mathrm{trace}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}) = b_{ij}$,

$$\frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{B}) = \boldsymbol{B}$$

- Derivative of $\mathrm{trace}(\boldsymbol{ABA}^{\mathsf{T}})$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}(\boldsymbol{ABA}^{\mathsf{T}}) &= \frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}\left((\boldsymbol{AB})\boldsymbol{A}^{\mathsf{T}}\right) + \frac{\partial}{\partial \boldsymbol{A}}\mathrm{trace}\left(\boldsymbol{A}(\boldsymbol{BA}^{\mathsf{T}})\right) \\ &= \boldsymbol{A}(\boldsymbol{B} + \boldsymbol{B}^{\mathsf{T}})\end{aligned}$$

Least-squares method

- Least-squares method solves the following optimization problem to obtain the estimate $\boldsymbol{x}$:

$$\underset{\boldsymbol{x}}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$$

- Least-squares solution is equivalent to ML estimate for linear discrete model when noise is Gaussian, i.e., $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}|\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{y}$$

Least-squares method

- To investigate robustness of least-squares solution $\hat{\boldsymbol{x}}$ against noise, consider singular value decomposition of $\boldsymbol{A}$ as

$$\boldsymbol{A} = \sum_{j=1}^{N} \gamma_j \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}}$$

  Singular vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M\}$ and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ form a set of orthonormal bases.

- Linear estimator $\boldsymbol{H} := (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}$ becomes

$$\boldsymbol{H} := (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}} = \sum_{j=1}^{N} \frac{1}{\gamma_j} \boldsymbol{v}_j \boldsymbol{u}_j^{\mathsf{T}}$$

Least-squares method

- Least-squares solution is rewritten as

$$\hat{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{x} + \sum_{j=1}^{N} \frac{\boldsymbol{u}_j^{\mathsf{T}}\boldsymbol{n}}{\gamma_j}\boldsymbol{v}_j$$

  The second term represents the effect of noise on the least-squares solution.

- $\hat{\boldsymbol{x}}$ is *unbiased estimator* (不偏推定量) because $\mathbb{E}[\hat{\boldsymbol{x}}] = \boldsymbol{x}$.
- When some singular values are close to 0, the terms of small singular values can amplify the effect of noise.

Regularized least-squares method

- One way to increase robustness against noise is regularization (正則化).
- By using ℓ_2-norm regularization (or Tikhonov regularization), the cost function becomes

$$\mathcal{J}(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x}\|^2$$

  The second term can be regarded as penalty term for the large amplitude of $\boldsymbol{x}$. λ is a constant for balancing data-fidelity and penalty terms.
- The regularized least-squares solution is obtained as

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}$$

Regularized least-squares method

- By using the singular value decomposition of $\boldsymbol{A}$,

$$\hat{\boldsymbol{x}} = \left(\sum_{j=1}^{N} \frac{\gamma_j^2}{\gamma_j^2 + \lambda} \boldsymbol{v}_j \boldsymbol{v}_j^{\mathsf{T}} \right) \boldsymbol{x} + \sum_{j=1}^{N} \frac{\gamma_j^2}{\gamma_j^2 + \lambda} (\boldsymbol{u}_j^{\mathsf{T}} \boldsymbol{n}) \boldsymbol{v}_j$$

  By observing the second term, i.e., the effect of noise, the amplification of the effect noise owing to small γ_j is suppressed by λ.

- Note that regularized least-squares solution $\hat{\boldsymbol{x}}$ is NOT unbiased estimator, i.e., $\mathbb{E}[\hat{\boldsymbol{x}}] \neq \boldsymbol{x}$.

Minimum-norm solution

- Next, suppose $M < N$: *underdetermined (劣決定)* case
- This problem has infinitely many solutions that minimize the cost function of least-squares error $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$. (# of equations $<$ # of unknown variables)
- Preferable solution will minimize ℓ_2-norm of $\boldsymbol{x}$, i.e., $\|\boldsymbol{x}\|^2$, with satisfying $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} \ \|\boldsymbol{x}\|^2 \\ &\text{subject to} \ \ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \end{aligned}$$

Such a solution is referred to as minimum-norm solution.

Minimum-norm solution

- By using *Lagrange's method of underdetermined multipliers* (ラグランジュ未定乗数法), Lagrange function $L(\boldsymbol{x}, \boldsymbol{\theta})$ is defined with Lagrangian multipliers $\boldsymbol{\theta} \in \mathbb{R}^M$ as

$$L(\boldsymbol{x}, \boldsymbol{\theta}) = \|\boldsymbol{x}\|^2 + \boldsymbol{\theta}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$$

- Original constrained optimization problem is converted into unconstrained optimization problem of minimizing L w.r.t. $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

Minimum-norm solution

- By differentiating $L(\boldsymbol{x}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{x}$ and $\boldsymbol{\theta}$,

$$\frac{\partial L(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{x}} = 2\boldsymbol{x} - \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\theta}$$
$$\frac{\partial L(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$$

- By setting $\partial L / \partial \boldsymbol{x} = \boldsymbol{0}$ and $\partial L / \partial \boldsymbol{\theta} = \boldsymbol{0}$,

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^{\mathsf{T}} \left(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} \right)^{-1} \boldsymbol{y}$$

Minimum-norm solution

- When noise is not negligible, data-fidelity constraint should be relaxed. Therefore, an alternative optimization problem is considered as

$$\underset{\boldsymbol{x}}{\text{minimize}} \, \|\boldsymbol{x}\|^2 \;\; \text{subject to} \;\; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 \leq \eta$$

  Here, η is a constant. However, this problem does not have closed-form solution.

- Instead, inequality constraint is replaced with equality constraint as

$$\underset{\boldsymbol{x}}{\text{minimize}} \, \|\boldsymbol{x}\|^2 \;\; \text{subject to} \;\; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 = \eta$$

Minimum-norm solution

- The Lagrangian function is defined as

$$L(\boldsymbol{x}, \theta) = \|\boldsymbol{x}\|^2 + \theta \left(\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 - \eta \right)$$

- By setting $\lambda = 1/\theta$, optimal solution is obtained as

$$\begin{aligned} \hat{\boldsymbol{x}} &= \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \lambda \|\boldsymbol{x}\|^2 \\ &= \boldsymbol{A}^{\mathsf{T}} \left(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{y} \end{aligned}$$

  The following relation for inverse matrix is used to derived the last line.

$$(\boldsymbol{B}^{-1} + \boldsymbol{C}^{\mathsf{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathsf{T}} = \boldsymbol{B}\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{I})^{-1}$$

- λ is a balancing parameter between first and second terms of $L(\boldsymbol{x}, \theta)$.

Table of Contents

Bayesian inference

- In Bayesian approach, the unknown variable (model parameter) $\boldsymbol{x}$ is also regarded as random variable, which is main difference from ML estimation.
- Based on Bayes' theorem, *posterior probability distribution* (事後確率分布) $p(\boldsymbol{x}|\boldsymbol{y})$, i.e., probability distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$, is represented as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{\int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}} \propto p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$$

$p(\boldsymbol{x})$ is *prior probability distribution* (事前確率分布) and $p(\boldsymbol{y}|\boldsymbol{x})$ is *likelihood* (尤度).

Bayesian inference

How to determine $p(\boldsymbol{x})$...?

- If no prior information on $\boldsymbol{x}$, $p(\boldsymbol{x})$ becomes constant (non-informative prior distribution: 無情報事前分布), and Bayes' estimation corresponds to ML estimation.
- *Conjugate prior* (共役事前分布) is often used for ease in computation. Then, posterior distribution results in the same probability distribution family of prior distribution.
- For example, Gaussian family is self-conjugate w.r.t. Gaussian likelihood function. When likelihood and prior distributions are Gaussian, posterior distribution is also Gaussian.

MAP and MMSE estimation

Two approaches to Bayes' estimation of $\boldsymbol{x}$.

- Maximum a posteriori (MAP) estimation (MAP 推定) is to estimate $\boldsymbol{x}$ so that posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ is maximized.

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y})$$

- Minimum mean square error (MMSE) estimation (MMSE 推定) is to estimate $\boldsymbol{x}$ so that mean square error between true and estimate is minimized.

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \arg\min_{\hat{\boldsymbol{x}}} \mathbb{E}[\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2]$$

MAP and MMSE estimation

- Mean square error is calculated as

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|^2] &= \int\int_{-\infty}^{\infty}(\hat{\boldsymbol{x}}-\boldsymbol{x})^{\mathsf{T}}(\hat{\boldsymbol{x}}-\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}(\hat{\boldsymbol{x}}-\boldsymbol{x})^{\mathsf{T}}(\hat{\boldsymbol{x}}-\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x}\right]p(\boldsymbol{y})\mathrm{d}\boldsymbol{y}\end{aligned}$$

  Note that $p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})$. (*multiplication theorem on probability*: 乗法定理)

- Therefore,

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \arg\min_{\hat{\boldsymbol{x}}}\int_{-\infty}^{\infty}(\hat{\boldsymbol{x}}-\boldsymbol{x})^{\mathsf{T}}(\hat{\boldsymbol{x}}-\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{y})\mathrm{d}\boldsymbol{x}$$

MAP and MMSE estimation

- By setting the derivative of the cost function w.r.t. $\hat{\boldsymbol{x}}$ to 0,

$$\frac{\partial}{\partial \hat{\boldsymbol{x}}} \int_{-\infty}^{\infty} (\hat{\boldsymbol{x}} - \boldsymbol{x})^{\mathsf{T}} (\hat{\boldsymbol{x}} - \boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x} = 2 \int_{-\infty}^{\infty} (\hat{\boldsymbol{x}} - \boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x} = 0$$

  Thus,

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \int_{-\infty}^{\infty} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) \mathrm{d}\boldsymbol{x}$$

  MMSE estimate corresponds to expectation of the posterior distribution.

- Generally complicated, but simply derived in the following two cases:
  - Both x and y are Gaussian. (Equivalent to MAP estimate)
  - $\hat{x}$ is obtained by *linear estimator*, i.e., $\hat{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{y} + \boldsymbol{g}$.

MAP estimation for Gaussian prior

- Assume $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \sigma_x^2 \boldsymbol{I})$

$$p(\boldsymbol{x}) = \frac{1}{(2\pi\sigma_x^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_x^2}\|\boldsymbol{x}\|^2\right)$$

- MAP estimate is obtained as

$$\begin{aligned}
\hat{\boldsymbol{x}}_{\mathrm{MAP}} &= \arg\min_{\boldsymbol{x}} -\ln p(\boldsymbol{x}|\boldsymbol{y}) \\
&= \arg\min_{\boldsymbol{x}} \left[-\ln p(\boldsymbol{x}|\boldsymbol{y}) - \ln p(\boldsymbol{x})\right] \\
&= \arg\min_{\boldsymbol{x}} \left[\frac{1}{\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \frac{1}{\sigma_x^2}\|\boldsymbol{x}\|^2\right] \\
&= \left(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A} + \frac{\sigma^2}{\sigma_x^2}\boldsymbol{I}\right)^{-1} \boldsymbol{A}^{\mathsf{T}}\boldsymbol{y}
\end{aligned}$$

- Correspond to regularized least squares solution when $\lambda = \sigma^2/\sigma_x^2$, i.e., signal-to-noise ratio.

MAP estimation for Gaussian prior

Brief summary of more general case:

- Noise $\boldsymbol{n}$ follows Gaussian distribution of mean $\mathbf{0}$ and precision (inverse of covariance) $\boldsymbol{\Lambda}$ ($:= \boldsymbol{\Sigma}^{-1}$).

$$\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}|\mathbf{0}, \boldsymbol{\Lambda}^{-1})$$

  Then, the likelihood is

$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Lambda}^{-1})$$

- Prior distribution $\boldsymbol{x}$ is Gaussian distribution of mean $\mathbf{0}$ and precision $\boldsymbol{\Lambda}_x$ ($:= \boldsymbol{\Sigma}_x^{-1}$).

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \boldsymbol{\Lambda}_x^{-1})$$

MAP estimation for Gaussian prior

- Since posterior distribution is also Gaussian, $p(\boldsymbol{x}|\boldsymbol{y})$ can be represented with its mean $\bar{\boldsymbol{x}}$ and precision $\boldsymbol{\Gamma}$ as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\bar{\boldsymbol{x}}, \boldsymbol{\Gamma}^{-1})$$

  Exponential part of $p(\boldsymbol{x}|\boldsymbol{y})$ becomes

$$-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\Gamma}(\boldsymbol{x}-\bar{\boldsymbol{x}}) = -\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Gamma}\bar{\boldsymbol{x}} + C$$

  C is constant term not including $\boldsymbol{x}$.

- Based on Bayes' theorem and Gaussian assumption of $p(\boldsymbol{y}|\boldsymbol{x})$ and $p(\boldsymbol{x})$, the exponential part becomes

$$\begin{aligned}
&-\frac{1}{2}\left[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Lambda}_x\boldsymbol{x} + (\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})^{\mathsf{T}}\boldsymbol{\Lambda}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})\right] \\
&= -\frac{1}{2}\boldsymbol{x}^{\mathsf{T}}(\boldsymbol{\Lambda}_x + \boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{y} + C'
\end{aligned}$$

MAP estimation for Gaussian prior

- By comparing coefficients,

$$\boldsymbol{\Gamma} = \boldsymbol{\Lambda}_x + \boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{A}$$
$$\bar{\boldsymbol{x}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{y} = (\boldsymbol{\Lambda}_x + \boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{y}$$

- Another representation of $\bar{\boldsymbol{x}}$:

$$\begin{aligned}\bar{\boldsymbol{x}} &= (\boldsymbol{\Lambda}_x + \boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{y} \\ &= \boldsymbol{\Lambda}_x^{-1}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{\Lambda}_x^{-1}\boldsymbol{A}^\mathsf{T} + \boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{y} \\ &= \boldsymbol{\Sigma}_x\boldsymbol{A}^\mathsf{T}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}\end{aligned}$$

  where $\boldsymbol{\Sigma}_y \coloneqq \boldsymbol{A}\boldsymbol{\Lambda}_x^{-1}\boldsymbol{A}^\mathsf{T} + \boldsymbol{\Lambda}^{-1}$.

- Thus, MAP estimate is obtained as

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \bar{\boldsymbol{x}} = (\boldsymbol{\Lambda}_x + \boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{y}$$

Linear MMSE estimator for Gaussian prior

- When linear estimator $\hat{\boldsymbol{x}} = \boldsymbol{H}\boldsymbol{y} + \boldsymbol{g}$ is assumed, MMSE estimate is obtained as

$$\underset{\boldsymbol{H},\boldsymbol{g}}{\text{minimize}}\, \mathcal{J}(\boldsymbol{H}, \boldsymbol{g}) := \mathbb{E}[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2] = \mathbb{E}[\|\boldsymbol{x} - (\boldsymbol{H}\boldsymbol{y} + \boldsymbol{g})\|^2]$$

- By setting $\partial\mathcal{J}/\partial\boldsymbol{H} = \boldsymbol{0}$ and $\partial\mathcal{J}/\partial\boldsymbol{g} = \boldsymbol{0}$,

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{H}} = -2\mathbb{E}[(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{y} - \boldsymbol{g})\boldsymbol{y}^{\mathsf{T}}] = \boldsymbol{0}$$
$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{g}} = -2\mathbb{E}[\boldsymbol{x} - \boldsymbol{H}\boldsymbol{y} - \boldsymbol{g}] = \boldsymbol{0}$$

- When $\mathbb{E}[\boldsymbol{n}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{0}$,

$$\boldsymbol{g} = \mathbb{E}[\boldsymbol{x}] - \boldsymbol{H}\mathbb{E}[\boldsymbol{y}] = \boldsymbol{0}$$
$$\boldsymbol{H} = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^{\mathsf{T}}](\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}])^{-1} := \boldsymbol{R}_{xy}\boldsymbol{R}_{y}^{-1}$$

Linear MMSE estimator for Gaussian prior

- Finally, MMSE estimate is obtained as

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \boldsymbol{R}_{xy}\boldsymbol{R}_y^{-1}\boldsymbol{y}$$

- When $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{n}|\boldsymbol{0}, \boldsymbol{\Sigma})$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{\Sigma}_x)$, and $\boldsymbol{x}$ and $\boldsymbol{n}$ are uncorrelated,

$$\begin{aligned}\boldsymbol{R}_{xy} &= \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^{\mathsf{T}}] = \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{n})^{\mathsf{T}}] = \boldsymbol{\Sigma}_x\boldsymbol{A}^{\mathsf{T}} \\ \boldsymbol{R}_y &= \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}] = \mathbb{E}[(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{n})(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{n})^{\mathsf{T}}] = \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y\end{aligned}$$

  Thus,

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \boldsymbol{\Sigma}_x\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y}$$

  $\Rightarrow$ Equivalent to MAP estimate for Gaussian prior.

Summary of statistical estimation

- Maximum likelihood (ML) estimation (最尤推定)

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{x}} -\ln p(\boldsymbol{y}|\boldsymbol{x})$$

- Maximum a posteriori (MAP) estimation (最大事後確率推定)

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{x}} -\ln p(\boldsymbol{x}|\boldsymbol{y})$$

- Minimum mean-square error (MMSE) estimation (最小平均二乗誤差推定)

$$\hat{\boldsymbol{x}}_{\mathrm{MMSE}} = \arg\min_{\hat{\boldsymbol{x}}} \mathbb{E}[\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2]$$

Table of Contents

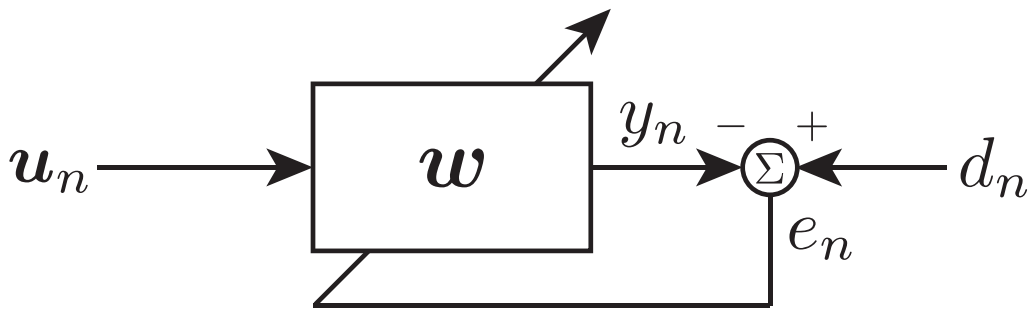
Wiener filter

- Wiener filter is special case of MMSE estimator.
- Input time-series signal $\boldsymbol{u}_n$ and filter coefficients $\boldsymbol{w}$ are

$$\boldsymbol{u}_n = [u_n, u_{n-1}, \dots, u_{n-K+1}]^\mathsf{T}$$
$$\boldsymbol{w} = [w_1, w_2, \dots, w_K]^\mathsf{T}$$

  $\mathbb{E}[u_n] = 0$ is assumed.
- Then, output signal y_n is convolution of $\boldsymbol{u}_n$ and $\boldsymbol{w}$ as

$$y_n = \boldsymbol{w}^\mathsf{T}\boldsymbol{u}_n = \sum_{j=1}^{K} w_j u_{n-j+1}$$



Wiener filter

- Goal is to obtain $\boldsymbol{w}$ such that output y_n corresponds to desired response d_n, i.e., error signal $e_n = d_n - y_n$ becomes 0.
- Cost function of wiener filter is defined as mean square error of e_n as

$$
\begin{aligned}
\mathcal{J} &= \mathbb{E}[|e_n|^2] \\
&= \mathbb{E}[(d_n - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{u}_n)(d_n - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{u}_n)^{\mathsf{T}}] \\
&= \mathbb{E}[|d_n|^2] - \boldsymbol{w}^{\mathsf{T}}\mathbb{E}[\boldsymbol{u}_n d_n] - \mathbb{E}[d_n \boldsymbol{u}_n]\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\mathbb{E}[\boldsymbol{u}_n \boldsymbol{u}_n^{\mathsf{T}}]\boldsymbol{w} \\
&= \sigma_d^2 - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{r}_{ud} - \boldsymbol{r}_{du}\boldsymbol{w} + \boldsymbol{w}^{\mathsf{T}}\boldsymbol{R}_u\boldsymbol{w}
\end{aligned}
$$

Here, $\sigma_d = \mathbb{E}[|d_n|^2]$, $\boldsymbol{r}_{du} = \mathbb{E}[d_n \boldsymbol{u}_n]$, $\boldsymbol{r}_{ud} = \mathbb{E}[\boldsymbol{u}_n d_n]$, $\boldsymbol{R}_u = \mathbb{E}[\boldsymbol{u}_n \boldsymbol{u}_n^{\mathsf{T}}]$.

Wiener filter

- By setting $\partial \mathcal{J} / \partial \boldsymbol{w} = 0$,

$$\boldsymbol{R}_u \boldsymbol{w} = \boldsymbol{r}_{ud}$$

  This equation is called Wiener–Hopf equation.
- Therefore, optimal filter is

$$\hat{\boldsymbol{w}} = \boldsymbol{R}_u^{-1} \boldsymbol{r}_{ud}$$

Noncausal Wiener filter

- By assuming input u_n is weak stationary, its autocorrelation matrix $\boldsymbol{R}_u$ becomes Toeplitz.

$$\boldsymbol{R}_u = \begin{bmatrix} r_u(0) & r_u(1) & \cdots & r_u(K-1) \\ r_u(-1) & r_u(0) & \cdots & r_u(K-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_u(1-K) & r_u(2-K) & \cdots & r_u(0) \end{bmatrix}$$

  Here, $r_u(n') \coloneqq \mathbb{E}[u_n u_{n-n'}]$.

- Wiener–Hopf equation is rewritten as

$$\sum_{k=1}^{K} w_k r_u(n-k+1) = r_{du}(n) \quad (n = 0, \ldots, K-1)$$

  where $r_{du}(n') \coloneqq \mathbb{E}[d_n u_{n-n'}]$.

Noncausal Wiener filter

- By extending the range of n to $(-\infty, \infty)$, frequency-domain Wiener–Hopf equation is derived as

$$W(\omega)S_u(\omega) = S_{du}(\omega)$$

  $W(\omega)$, $S_u(\omega)$, and $S_{du}(\omega)$ are Fourier transform of w_k, $r_u(n)$, and $r_{du}(n)$, respectively.
- Suppose that u_n is sum of desired signal d_n and noise v_n:

$$u_n = d_n + v_n$$

  and d_n and v_n are uncorrelated.
- Then, noncausal Wiener filter is derived as

$$W(\omega) = \frac{S_{du}(\omega)}{S_u(\omega)} = \frac{S_d(\omega)}{S_d(\omega) + S_v(\omega)}$$

Wiener filter for denoising

- Spectral subtraction method in speech enhancement:
  - $S_v(\omega)$: Estimated from input signal in interval that speech activity is absent.
  - $S_d(\omega)$: Approximated as subtraction of noise power from input power.

$$S_d(\omega) = \max(S_u(\omega) - S_v(\omega), 0)$$

  - Example code:
    - `https://colab.research.google.com/drive/1XE3F5h1fvy5jpFVWEK8iz_WV1vxZMdhF?usp=sharing`

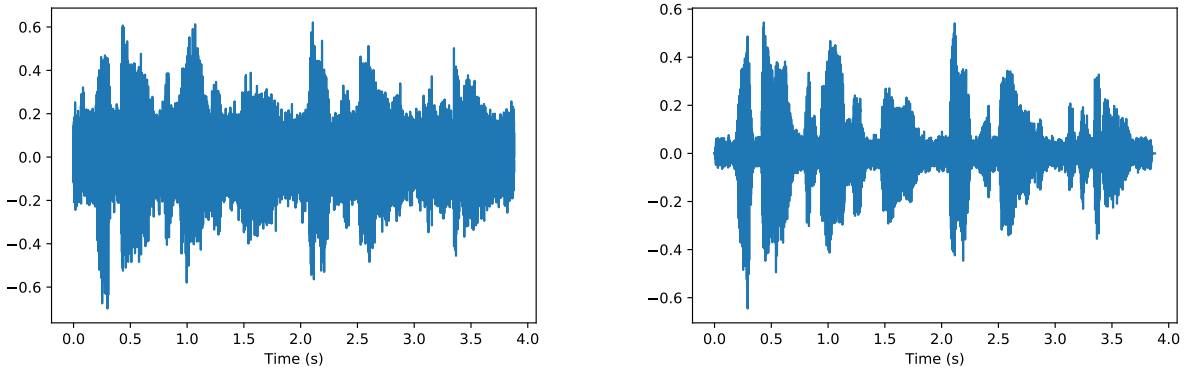


Figure: Left: Noisy speech, Right: Denoised speech

References

C. Bishop (2006)
Pattern Recognition and Machine Learning
Springer-Verlag, New York.

関原謙介 (2015)
ベイズ信号処理 – 信号・ノイズ・推定をベイズ的に考える
共立出版, Tokyo.

S. Boll (1979)
Suppression of acoustic noise in speech using spectral subtraction
IEEE Trans. ASSP, vol. 27, no. 2, pp. 113–120.