ベイズ推論の基礎と信号処理への応用

午前

確率変数・確率分布の基礎
最尤原理と最尤推定
最小二乗法とその解の特性
正則化とミニマムノルム解
L1ノルムの解とスパースネスについての議論

午後１

ベイズ推定の基礎
線形正規モデル
EMアルゴリズム

午後２

ベイズ因子分析
スパース・ベイズ法
変分ベイズ法
質問など

参考文献：「ベイズ信号処理」（共立出版）
「統計的信号処理」（共立出版）

ベイズ信号処理 ―信号・ ×
:su-pub.co.jp/bookdetail/9784320085749
MyWeb Google Google カレンダー Google Scholar News デジスコ nifty 伊豆 デジカメ 銀行 天体写真 交通機関

キーワード検索
検索
詳細検索 »

- 科学一般
- 数学
- 情報・コンピュータ
- 物理学
- 化学・化学工業
- 地球科学・宇宙科学
- 生物学・生物科学
- 医学・薬学
- 生活科学
- 環境科学
- 工学一般
- 土木工学
- 建築学
- 機械工学
- 電気・電子工学
- 経済・経営
- 認知科学・心理学
- 写真集

home » 電気・電子工学 » 音・光・画像・信号処理 » 信号処理 » ベイズ信号処理

ベイズ信号処理―信号・ノイズ・推定をベイズ的に考える―

関原 謙介 著

ISBN 978-4-320-08574-9
判型 A5
ページ数 168ページ
発行年月 2015年04月
本体価格 2,800円

オンライン書店で購入する
- amazon
- 紀伊國屋書店
- honto
- ヨドバシ・ドット・コム
- 楽天ブックス
- セブンネットショッピング
- HonyaClub
- e-hon
- TSUTAYA

店頭在庫を確認する
- 紀伊國屋書店(新宿本店)
- 旭屋倶楽部
- 東京都書店案内

ノイズが重畳した観測データから，信号を選択的に推定する手法についての技術・学問体系は「統計的信号処理」と呼ばれる。

好評既刊『統計的信号処理』では大学学部生のレベルを対象として統計的信号処理を平易に解説しており，ベイズ推定法もその初歩までを解説しているが，読者の方々からはベイズ推定，特にその最近の進歩についてさらに知りたいとの要望が多数寄せられた。

そこで本書ではベイズ推定による信号処理をメインテーマに取り上げた。ベイズ推定は，機械学習，パターン認識，自然言語処理やデータマイニングといった分野でも広く応用され，いわば流行の学問・技術分野といった感さえあるが，本書では統計的信号処理の枠組みの中で，ベイズ推定を基にした信号処理法を，ベイズ推定を初めて学ぶ読者を対象に解説している。

ベイズ推定の初学者でも議論の筋道が追えるよう，数式の展開や導出などをできる限りていねいに説明し，読者が抱える問題の解決に「使える知識」としてベイズ推定を習得できることを目指した。

自己紹介

略歴

1976-2000：日立製作所　中央研究所にて医用画像診断機器（CTスキャン・MRI・MEGなど）の研究開発を行なう.

1996－2000： JSTのプロジェクト「心表象」においてMRIや脳磁場を用いた脳機能計測の研究を行う.

2000-2015：首都大学東京システムデザイン学部教授 生体磁気を用いた生体機能計測・イメージングの研究を行う.

2015- :東京医科歯科大学先端技術医療応用学講座特任教授．株式会社シグナルアナリシス代表

専門

信号処理，再構成問題
特に，生体信号における信号処理・逆問題解法

http://www.comp.sd.tmu.ac.jp/sekihara/

http://www.signalanalysis.jp/

推定問題の基礎

確率・統計の基礎から正則化まで

- 確率変数・確率分布の基礎
- 最尤原理と最尤推定
- 最小二乗法とその解の特性
- 正則化とミニマムノルム解
- L1ノルムの解と，解のスパースネスについての議論

確率変数，確率分布，期待値，分散

- それが取る各値に対しそれぞれ確率が与えられている変数を確率変数と呼ぶ.
- 確率変数 x が連続値をとり，

$$P(a \le x \le b) = \int_a^b p(x)dx$$

なる$p(x)$が存在するとき， xは連続型の確率分布を持つといわれる． $p(x)$ は x の確率密度分布と呼ばれる.

- 確率変数の期待値と分散

期待値: $E(x) = \int_{-\infty}^{\infty} xp(x)dx = \mu$　　分散: $V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = E[(x-\mu)^2]$

- 多（２）変数の確率分布: $P(a \le x \le b, c \le y \le d) = \int_c^d \int_a^b p(x,y)dxdy$
- 周辺化：同時確率分布から一方の確率変数を積分消去してもう一方のみの確率分布とすること

例: $g(x) = \int_{-\infty}^{\infty} p(x,y)dy$, および $h(y) = \int_{-\infty}^{\infty} p(x,y)dx$

共分散

確率変数 x と y に対して以下を共分散と呼ぶ.

$$Cov(x,y) = E[(x-\mu_x)(y-\mu_y)], \quad \text{ここで}, \mu_x = E(x),\ \mu_y = E(y)$$

以下が成立：

$$V(x+y) = V(x) + V(y) + 2Cov(x,y)$$
$$Cov(x,y) = E(xy) - E(x)E(y)$$

確率変数の独立：

確率変数 x と y に対し，$p(x,y) = p_1(x)p_2(y)$となるとき，x と y は独立である.

確率変数 x と y が独立な場合以下が成立.

$$Cov(x,y) = 0$$
$$V(x+y) = V(x) + V(y)$$
$$E(xy) = E(x)E(y)$$

N個の確率変数 $x_1,\ldots,x_N$ に対し，以下が成立．

$$E(x_1+\cdots+x_N)=E(x_1)+\cdots+E(x_N)$$

さらに，$x_1,\ldots,x_N$ が独立なら，

$$V(x_1+\cdots+x_N)=V(x_1)+\cdots+V(x_N)$$

N個の確率変数 $x_1,\ldots,x_N$ が独立で同一な分布，

$$E(x_1)=E(x_2)=\cdots=E(x_N)=\mu,\ V(x_1)=V(x_2)\cdots=V(x_N)=\sigma^2$$

に従う場合，

$$E(x_1+\cdots+x_N)=N\mu \text{ および } V(x_1+\cdots+x_N)=N\sigma^2 \text{ である．}$$

確率変数 $x_1,\ldots,x_N$ の算術平均：

$$\overline{x}=\frac{x_1+\cdots+x_N}{N}$$

に対し，

$$E(\overline{x})=\mu,\qquad V(\overline{x})=\frac{\sigma^2}{N} \text{ が成立．}$$

独立で同一な分布（independently, identically distributed)—I.I.D.と略す場合がある．

正規分布

確率密度: $p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$

$E(x) = \int_{-\infty}^{\infty} xp(x)dx = \mu$, および $V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$ は簡単に確認できる.

簡略的な表記法: $p(x) = N(x \mid \mu, \sigma^2)$

確率変数 x が,期待値μ, 分散σ^2の正規分布する. $\Rightarrow$ $x \sim N(x \mid \mu, \sigma^2)$

正規分布における重要で便利な性質

1) $y = ax + b$なる確率変数 y に対し, $x \sim N(x \mid \mu, \sigma^2)$ なら $y \sim N(y \mid a\mu + b, a^2\sigma^2)$

2) $x \sim N(x \mid \mu_1, \sigma_1^2)$, $y \sim N(y \mid \mu_2, \sigma_2^2)$で x と y が独立なら

$$z = x + y \sim N(z \mid \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

N個の独立な確率変数 $x_1, x_2, \ldots, x_N$ に対し，

（1）$x_1 \sim N(x_1 \mid \mu_1, {\sigma_1}^2), x_2 \sim N(x_2 \mid \mu_2, {\sigma_2}^2), \cdots, x_N \sim N(x_N \mid \mu_N, {\sigma_N}^2)$なら，

$$z = x_1 + x_2 + \cdots + x_N \sim N(z \mid \mu_1 + \mu_2 + \cdots + \mu_N, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_N^2)$$

（2）$x_1, x_2, \cdots, x_N$ が独立で同一な分布 $x_j \sim N(x_j \mid \mu, \sigma^2)$ をする場合，

$$z = x_1 + x_2 + \cdots + x_N \sim N(z \mid N\mu, N\sigma^2)$$

（3）$x_1, x_2, \cdots, x_N$が独立で同一な分布（正規分布に限らず）をする場合，

$$z = x_1 + x_2 + \cdots + x_N \xrightarrow[N \to \infty]{} N(z \mid N\mu, N\sigma^2)$$ （中心極限定理）

- 実際の観測においてデータに重畳しているノイズは，独立な複数の不規則信号の総和と考えることができる．
- 中心極限定理はデータに重畳しているノイズの確率分布に便宜的に正規分布を仮定する事の妥当性の（ある程度の）根拠となる．

ベクトル型確率変数

確率変数$x_1,\ldots,x_N$に対して，列ベクトル：$\boldsymbol{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$を用いて確率変数$x_1,\ldots,x_N$全体を表す．

これをベクトル型確率変数と呼ぶ．

平均ベクトル：$E(\boldsymbol{x})=\begin{bmatrix} E(x_1) \\ \vdots \\ E(x_N) \end{bmatrix}=\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}=\boldsymbol{\mu}$

共分散行列：

$$E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T]=\begin{bmatrix} E[(x_1-\mu_1)^2] & E[(x_1-\mu_1)(x_2-\mu_2)] & \cdots & E[(x_1-\mu_1)(x_N-\mu_N)] \\ E[(x_2-\mu_2)(x_1-\mu_1)] & \cdot & \cdot & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ E[(x_N-\mu_N)(x_1-\mu_1)] & \cdot & \cdots & E[(x_N-\mu_N)^2] \end{bmatrix}=\boldsymbol{\Sigma}$$

確率変数$x_1,\ldots,x_N$が独立で同一な分布をする場合，$Cov(x_i,x_j)=0,\ V(x_j)=\sigma^2$ であるので，

$$\boldsymbol{\Sigma}=\begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}=\sigma^2\begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}=\sigma^2\boldsymbol{I}$$

多次元正規分布

1次元（スカラー）正規分布：$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$

多次元（ベクトル型）正規分布：$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} \exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]$

（N はベクトル $\boldsymbol{x}$ のサイズ）

確率変数$x_1, x_2, \ldots, x_N$が独立で同一な分布（I.I.D)をする場合，

$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ であるので，$\left|\boldsymbol{\Sigma}\right| = (\sigma^2)^N$ より，

$$
\begin{aligned}
p(\boldsymbol{x}) &= \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})] \\
&= \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp[-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(x_j-\mu_j)^2] \\
&= \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(x_j-\mu_j)^2]
\end{aligned}
$$

多次元正規分布は独立な1次元正規分布の積として表される．

推定

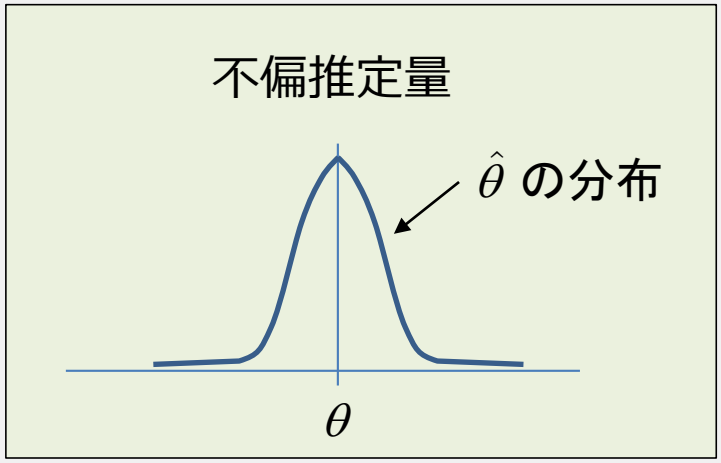
N 回の繰り返し観測で，観測データ $x_1, x_2, \ldots, x_N$ を得るとし，これら観測データを用いて未知量 θ の推定結果 $\hat{\theta}$ を得るとする．

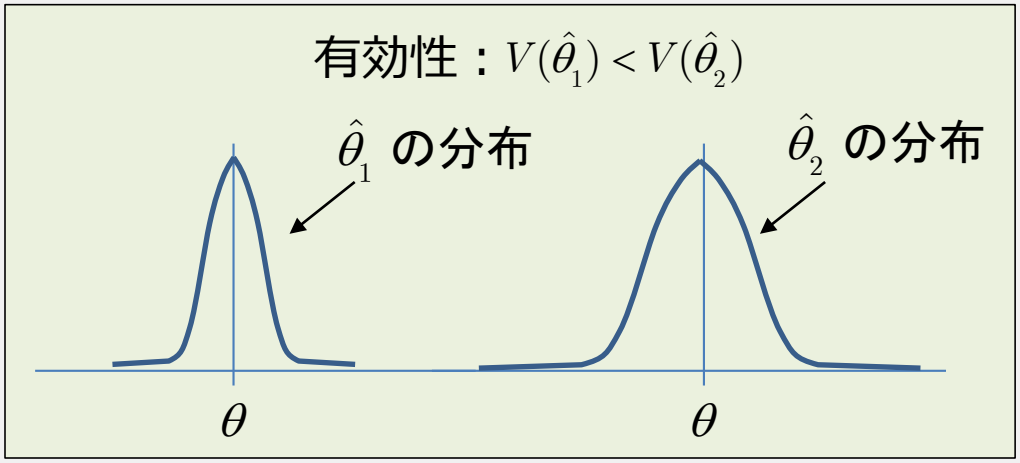
観測データ $x_1, x_2, \ldots, x_N$ は確率変数であるので，これらから求まる $\hat{\theta}$ も確率変数である．したがって，$x_1, x_2, \ldots, x_N$ の値に応じて確率的な値を取る．

推定量の良さを表す指標

（1）不偏性：$E(\hat{\theta}) = \theta$

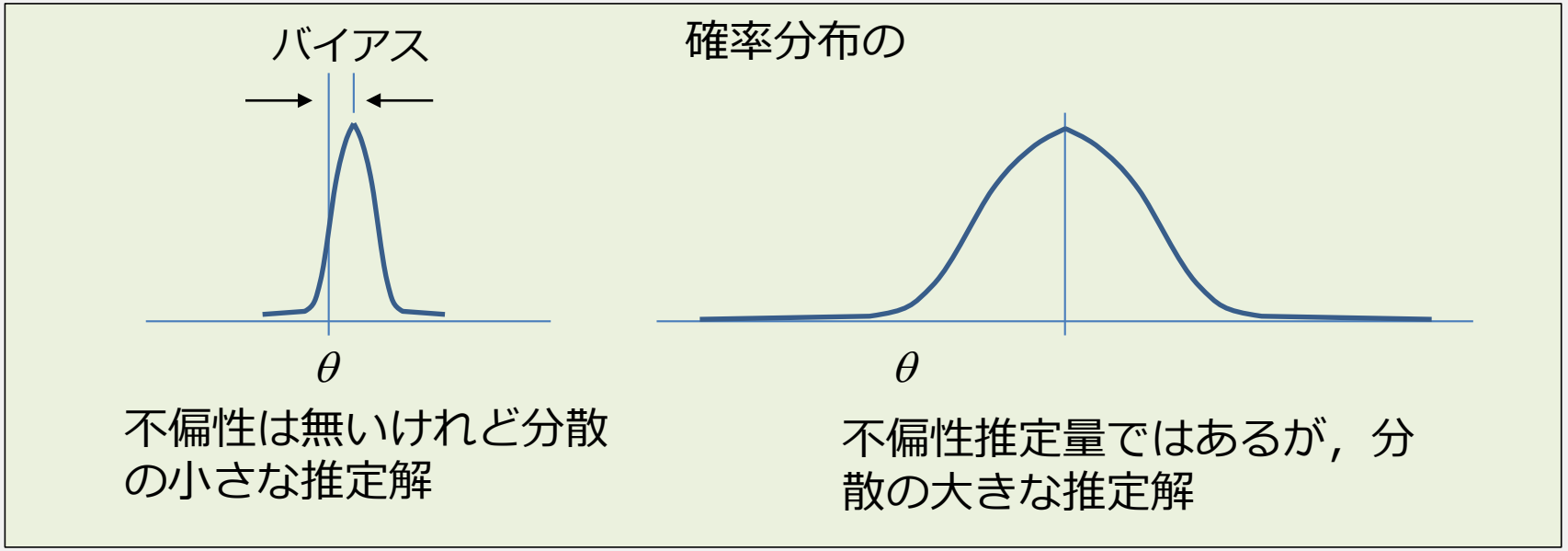
（2）有効性：$V(\hat{\theta}_1) < V(\hat{\theta}_2)$





不偏性と有効性

- 不偏性を持たない推定解の場合には推定結果は真の値とは異なる値を中心に分布する.
- バイアスが大きければ，いくら分散が小さくても意味がないので，統計学では，有効性の概念は不偏推定量にとってのみ意味のあるものと考えられている.
- しかし，バイアスが分散に比べて小さければ，不偏推定量であっても分散の大きな推定量と，不偏推定量ではなくても分散の小さな推定量とどちらがより「良い」推定量であるのかは一概には言えない.



最尤推定法

N 回の繰り返し観測で未知量 θ を推定する

最尤原理

N回の繰り返し観測で，観測値$x_1, x_2, \ldots, x_N$が得られた．

⇓

得られた観測結果$x_1, x_2, \ldots, x_N$は「確率最大のものが実現した．」

すなわち「最も起こりやすいことが起きた．」結果と考える．

N回の繰り返し観測，確率変数$x_1, \ldots, x_N$は独立で，$x_j \sim p(x_j)$とすれば，

観測結果$x_1, \ldots, x_N$が実現する確率：

$$p(x_1, x_2, \ldots, x_N) = \prod_{j=1}^{N} p(x_j)$$

確率分布は未知量θをパラメータとして含む．

尤度関数（likelihood function）：$L(\theta) = \prod_{j=1}^{N} p(x_j)$

最尤推定法：$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)$

対数尤度関数（log-likelihood function）

N回の繰り返し観測を用いた最も単純な観測モデル：

$$
\begin{aligned}
x_1 &= \theta + \varepsilon_1 \\
x_2 &= \theta + \varepsilon_2 \\
&\vdots \\
x_N &= \theta + \varepsilon_N
\end{aligned}
$$

$$
\varepsilon_j \sim N(\varepsilon_j \mid 0, \sigma^2)
$$
$$
\Downarrow
$$
$$
x_j \sim N(x_j \mid \theta, \sigma^2)
$$

尤度関数：$L(\theta) = \prod_{j=1}^{N} p(x_j) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x_j - \theta)^2}{2\sigma^2}]$

⇩

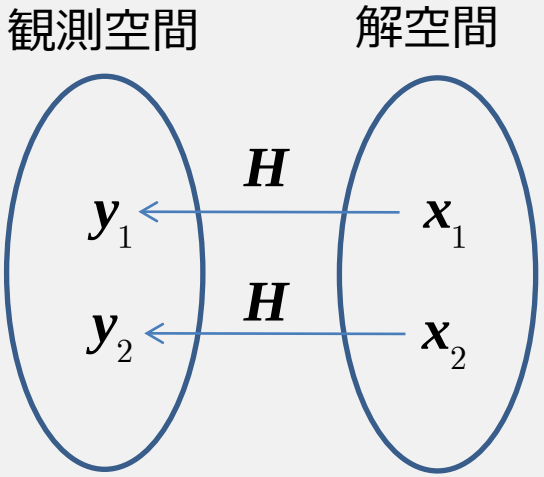
最尤推定解：$\hat{\theta} = \arg\max_{\theta} \log \prod_{j=1}^{N} p(x_j) = \arg\max_{\theta} \left[\sum_{j=1}^{N} -\frac{(x_j - \theta)^2}{2\sigma^2} \right] + C$

⇩

$\frac{\partial}{\partial\theta} \left[\sum_{j=1}^{N} -\frac{(x_j - \theta)^2}{2\sigma^2} + C \right] = 0$ とおいて，最尤推定解 $\hat{\theta} = \frac{1}{N} \sum_{j=1}^{N} x_j$ を得る．

線形離散モデル：線型方程式 $y=Hx$ を解く

未知量 : $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$, 観測データ : $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$ に



線形な関係 :

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} \quad (\boldsymbol{H} : M \times N \text{の行列})$$

が存在する. これに加法的にノイズ $\boldsymbol{\varepsilon}$ が重畳するとして

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

を観測データのモデル（線形離散モデル）と呼ぶ.

「未知量 $\boldsymbol{x}$ が原因で観測結果$\boldsymbol{y}$を生じたと解釈することもできる．」

原因 $\boldsymbol{x}$ を与えて結果 $\boldsymbol{y}$ を推定する ⟶ 順（方向）問題 （forward problem)
(行列$\boldsymbol{H}$ を推定する問題に等しい．)

結果 $\boldsymbol{y}$ を与えて原因 $\boldsymbol{x}$ を推定する ⟶ 逆（方向）問題 （inverse problem)

$M > N \Rightarrow$ 逆問題は優決定（over-determined),
$M < N \Rightarrow$ 逆問題は劣決定 (under-determined)と呼ばれる.

線形最小二乗法

まず，優決定（$M > N$)の場合を仮定

モデル：$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$ の基で，観測データ $\boldsymbol{y}$ から未知な量 $\boldsymbol{x}$ を推定することを考える．

前提： $\boldsymbol{H}$ は既知である．

最尤推定を行う

ノイズ確率分布：$p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon} \mid 0, \sigma^2 \boldsymbol{I})$ を仮定すれば，

正規分布の便利な性質１）より　$\boldsymbol{y} = \boldsymbol{\varepsilon} + \boldsymbol{H}\boldsymbol{x} \sim N(y \mid \boldsymbol{H}\boldsymbol{x}, \sigma^2 \boldsymbol{I})$

観測データの確率分布：

$$p(\boldsymbol{y}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi)^{M/2}(\sigma^2)^{M/2}} \exp[-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2]$$

対数尤度関数：$\log L(\boldsymbol{x}) = \log p(\boldsymbol{y}) = -\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 + C$

線形最小二乗法（２）

対数尤度関数を最大とする $\boldsymbol{x}$ は，以下の F を最小とする $\boldsymbol{x}$ に等しい.

$$F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2$$

この F を最小二乗のコスト関数と呼ぶ.

最尤推定解：　$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F$　である.

最小二乗のコスト関数 F を最小にすることで未知量を求める手法を最小自乗法と呼ぶ.

最小二乗のコスト関数の解釈：

$$F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2$$

推定した $\boldsymbol{x}$ のデータ $\boldsymbol{y}$ への一致度（適合度）と解釈できる.

F を最小にする解はデータを最もよく「説明する」解である.

最小二乗解の導出

コスト関数: $F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2 = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{H}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{x}$

$\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T\boldsymbol{H}^T\boldsymbol{y} = \boldsymbol{H}^T\boldsymbol{y}$, $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{y}^T\boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}^T\boldsymbol{y}$, $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{x} = 2\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{x}$ を考慮すれば,

$\frac{\partial}{\partial \boldsymbol{x}} F = 2\left(-\boldsymbol{H}^T\boldsymbol{y} + \boldsymbol{H}^T\boldsymbol{H}\boldsymbol{x}\right) = 0$ となり,

したがって，以下の最小二乗解を得る.

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

$\boldsymbol{x}$ は線形推定量（後述）である

繰り返し計測モデルでの最小二乗解

$$
\begin{aligned}
y_1 &= \theta + \varepsilon_1 \\
y_2 &= \theta + \varepsilon_2 \\
&\vdots \\
y_M &= \theta + \varepsilon_M
\end{aligned}
\quad \Rightarrow \quad
\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \theta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_M \end{bmatrix}
\quad \Rightarrow \quad
\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}, \ \boldsymbol{H} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}
$$

したがって，

$$
\hat{\boldsymbol{x}} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y} = \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \frac{1}{M} \sum_{j=1}^{M} y_j
$$

$\boldsymbol{H}^T$　　$\boldsymbol{H}$　　$\boldsymbol{H}^T$

線形推定法（linear estimator）

データにある重み行列を掛けて未知量を推定する方法：

すなわち，観測データ $\boldsymbol{y}$ から未知な量$\boldsymbol{x}$ を推定する場合，ある重み行列$\boldsymbol{W}$ を用いて，

$$\hat{\boldsymbol{x}} = \boldsymbol{W}^T \boldsymbol{y}$$

として，$\boldsymbol{x}$ の推定解 $\hat{\boldsymbol{x}}$ を求める方法を線形推定法と呼び，$\hat{\boldsymbol{x}}$ を線形推定量と呼ぶ.

したがって，最小二乗解：

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

は重み行列　$\boldsymbol{W}^T = \left(\boldsymbol{H}^T \boldsymbol{H}\right)^{-1} \boldsymbol{H}^T$　を用いた線形推定法である.

線形推定解における不偏性

ノイズの影響

線形推定量： $\hat{\boldsymbol{x}} = \boldsymbol{W}^T \boldsymbol{y} = \boldsymbol{W}^T (\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}) = \boldsymbol{W}^T \boldsymbol{H}\boldsymbol{x} + \boldsymbol{W}^T \boldsymbol{\varepsilon}$

$$E(\hat{\boldsymbol{x}}) = E(\boldsymbol{W}^T \boldsymbol{H}\boldsymbol{x} + \boldsymbol{W}^T \boldsymbol{\varepsilon}) = \boldsymbol{W}^T \boldsymbol{H}\boldsymbol{x} + \boldsymbol{W}^T E(\boldsymbol{\varepsilon}) = \boldsymbol{W}^T \boldsymbol{H}\boldsymbol{x}$$

であるので，不偏推定量となる条件： $E(\hat{\boldsymbol{x}}) = \boldsymbol{x}$ をみたすには，$\boldsymbol{W}^T \boldsymbol{H} = \boldsymbol{I}$ でなければならない.

線形不偏推定量： $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{W}^T \boldsymbol{\varepsilon}$

線形不偏推定量においては推定結果に含まれる誤差は観測データに含まれるノイズの影響のみである．SN比が大きい極限で推定解は真の解に等しくなる.

最小二乗解場合

$$\hat{\boldsymbol{x}} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}) = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{H}\boldsymbol{x} + (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{\varepsilon}$$

したがって， $\hat{\boldsymbol{x}} = \boldsymbol{x} + (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{\varepsilon}$

線形不偏推定量であり，ノイズゼロの極限で真の解に一致する

最小二乗解の分散

$$\boldsymbol{\Sigma}_x = E\left[(\hat{\boldsymbol{x}}-\boldsymbol{x})(\hat{\boldsymbol{x}}-\boldsymbol{x})^T\right] = E\left[(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon}[(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon}]^T\right]$$

$$= E\left[(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1}\right] = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{H})^{-1} = \sigma^2(\boldsymbol{H}^T\boldsymbol{H})^{-1}$$

ここで，$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\boldsymbol{I}$ を用いた.

$(\boldsymbol{H}^T\boldsymbol{H})^{-1}$:ノイズゲインと呼ばれる.

最小二乗解が線形不偏推定量の中で最も分散の小さな推定量ー有効推定量であることを示すことができる．最小二乗解はBest linear unbiased estimator (BLUE)であると言われる.

最小自乗法はすばらしい方法のように思えるが??

すばらしい方法かどうかは順方向行列 $\boldsymbol{H}$ の特性による.

順方向行列$\boldsymbol{H}$の特異値分解（SVD：singular value decomposition)

over-determined$(M > N)$の場合

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_M \end{bmatrix} \begin{bmatrix} \gamma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \gamma_N \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_N^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_N \end{bmatrix} \begin{bmatrix} \gamma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \gamma_N \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_N^T \end{bmatrix} = \sum_{j=1}^{N} \gamma_j \boldsymbol{u}_j \boldsymbol{v}_j^T$$

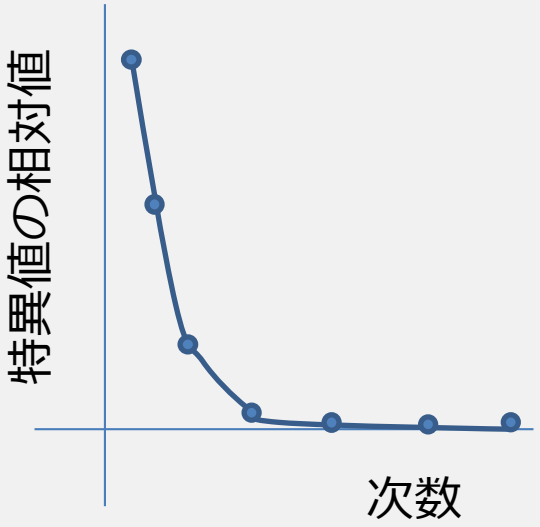
特異値$\gamma_1, \ldots, \gamma_N$（$\geq 0$)は大きさの順に番号付けされているとする.

ある次数以上の特異値が非常に小さく，ゼロに近くなる

特異値を大きさの順に並べると：

$$\gamma_1, \ldots, \gamma_r, \underbrace{0, \ldots 0}$$

r次数以上の特異値が非常に小さく，ゼロに近くなる



最小二乗解のノイズ耐性

ある次数から先の特異値が非常に小さくなる場合に最小二乗解にどんな影響が出るであろうか.

$\boldsymbol{H}$ の特異値と特異値ベクトルを用いて最小二乗解を記述する.

最小二乗推定の重み： $(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T = \sum_{j=1}^{N}\frac{1}{\gamma_j}\boldsymbol{v}_j\boldsymbol{u}_j^T$

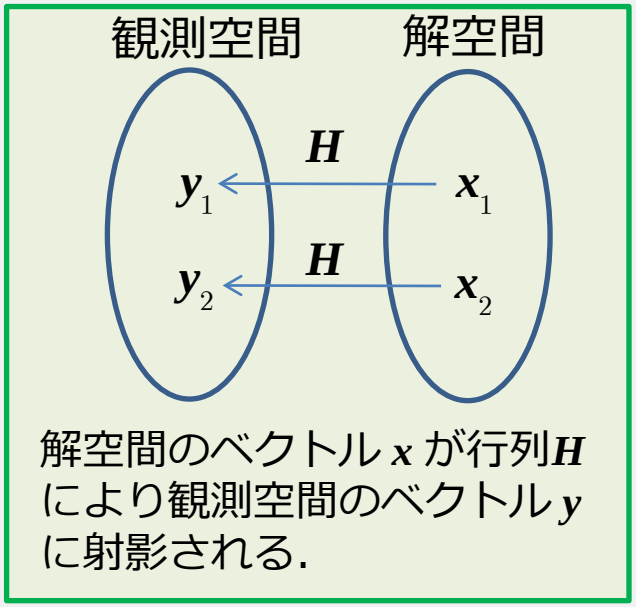
最小二乗解： $\hat{\boldsymbol{x}} = \boldsymbol{x} + (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon} = \boldsymbol{x} + \left[\sum_{j=1}^{N}\frac{1}{\gamma_j}\boldsymbol{v}_j\boldsymbol{u}_j^T\right]\boldsymbol{\varepsilon} = \boldsymbol{x} + \sum_{j=1}^{N}\frac{\left(\boldsymbol{u}_j^T\boldsymbol{\varepsilon}\right)}{\gamma_j}\boldsymbol{v}_j$

ある次数以上の特異値が非常に小さくなる場合，その特異値を含む項はノイズの影響を増幅してしまう.

結果として： $\boldsymbol{x} \ll \sum_{j=1}^{N}\frac{\left(\boldsymbol{u}_j^T\boldsymbol{\varepsilon}\right)}{\gamma_j}\boldsymbol{v}_j$ 最小二乗解はノイズに起因した大きな誤差を含む.

それでは，どんな場合にある次数以上の特異値がゼロに近くなるのか？

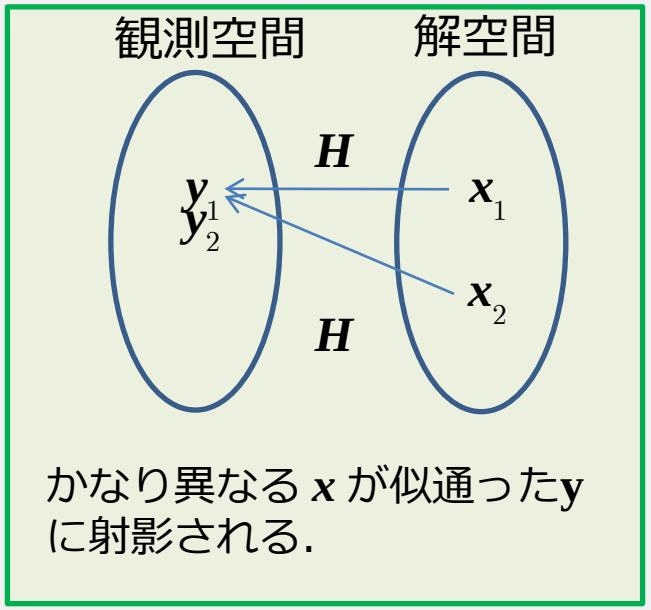
推定に対する更なる考察



解空間のベクトル x が行列 H により観測空間のベクトル y に射影される.

$$y_1 = Hx_1$$

$$y_2 = Hx_2$$



かなり異なる x が似通った y に射影される.

x_1 と x_2 がかなり異なるにも関わらず, $y_1 \approx y_2$ となる場合を考える.

⇓

y_1 と y_2 のわずかな違いから, x_1 なのか x_2 なのかを決めなければならない.

⇓

こんな場合, 未知量 x を 観測データ y から決めることは難しい.

この難しさは, 最小二乗推定のどこに反映されるのか？

⇓

ある次数以上の特異値がゼロに近くなる

順方向行列 $\boldsymbol{H}$ の特性は，計測の物理的な性質から決まり，観測者（ユーザー）がコントロールできない場合が多い.

それでは，ある次数以上の特異値がゼロに近くなるような，順方向行列の場合に，推定はどのように行えばよいか？

データに含まれる誤差（ノイズ）に対してその影響を受けにくい（robust）な方法が望ましい!

正則化の導入

正則化を用いた最小二乗解

最小二乗解：$F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2$ とおいたF を最小とする $\boldsymbol{x}$ を求める．

⇓

求まる推定解は観測データ $\boldsymbol{y}$ に「最もよく一致する」 $\boldsymbol{x}$ である．

⇓

しかし，$\boldsymbol{y}$ に大きなノイズが重畳している場合，観測データとの一致度をあまり追求しても意味がなく，返って観測データに含まれるノイズの影響を受けてしまう．

$\left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2$ のみでなく別の基準$\phi(\boldsymbol{x})$を導入し，コスト関数を以下のように定義する：

$$F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2 + \xi\phi(\boldsymbol{x})$$

$\phi(\boldsymbol{x})$ ：制約条件．解として望ましい性質を関数に表したもの．

ξ：正則化定数．正の定数で，第1項（データとの一致度）と第2項（制約条件）のバランスを取る．

$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F$から解を求めれば，推定解がノイズの影響を受けにくくなることが期待できる．

正則化を用いた最小二乗解（２）

解のノルム$\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$がよく用いられる.

$F = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \xi\|\boldsymbol{x}\|^2$　として，$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F$から解を求めれば以下の公式を得る:

$$\hat{\boldsymbol{x}} = (\boldsymbol{H}^T\boldsymbol{H} + \xi\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

解のノイズ耐性

$$\hat{\boldsymbol{x}} = (\boldsymbol{H}^T\boldsymbol{H} + \xi\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y} = \left[\sum_{j=1}^{N}\frac{\gamma_j^2}{\gamma_j^2+\xi}\boldsymbol{v}_j\boldsymbol{v}_j^T\right]\boldsymbol{x} + \sum_{j=1}^{N}\frac{\gamma_j}{\gamma_j^2+\xi}\left(\boldsymbol{u}_j^T\boldsymbol{\varepsilon}\right)\boldsymbol{v}_j$$

正の定数ξが分母に含まれるため，小さなγ_jを含む項は小さくなり，ノイズ増幅を回避できる.

第1項は $\boldsymbol{x}$ とは異なるため，この解は不偏性を持たない．ただし：

$\left[\sum_{j=1}^{N}\frac{\gamma_j^2}{\gamma_j^2+\xi}\boldsymbol{v}_j\boldsymbol{v}_j^T\right]\boldsymbol{x} \xrightarrow[\xi\to 0]{} \left[\sum_{j=1}^{N}\boldsymbol{v}_j\boldsymbol{v}_j^T\right]\boldsymbol{x} = \boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}$　であるので，　$\xi\to 0$で不偏推定解となる.

正則化定数 ξ は，解のノイズ耐性と解のバイアスのトレードを与える.

$F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2 + \xi\phi(\boldsymbol{x})$ における正則化定数 ξ はどのように決めるか

- 正則化定数ξは観測データに含まれるノイズの量に依存する．ノイズが小さければ ξを小さくして観測データとの一致度を優先させることが合理的であるし，ノイズが大きければ観測データとの一致度にこだわっても意味がないため，大きな ξを用いることがノイズ増幅率の点からも合理的である．

- 多くの場合 ξの値は経験的に決められている．つまり，ξのいろいろな値で解を計算してみて最も妥当な解が得られるξを採用する．実用的にはこのやり方で特に問題がない場合が多い．

- ξの理論的な決め方についてはベイズ推定を用いた方法を後に紹介する．

劣決定系における推定解の求め方：

線形離散モデル

未知量：$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$，観測データ：$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$ に

線形な関係：

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$$

が存在する．これに加法的にノイズ$\boldsymbol{\varepsilon}$が重畳するとして

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$$

を観測データのモデル（線形離散モデル）と呼ぶ．

$M > N \Rightarrow$ 逆問題は優決定（over-determined)

$M < N \Rightarrow$ 逆問題は劣決定 (under-determined)

劣決定系における推定解の求め方：

劣決定系の場合，$\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}\|^2=0$ となる，つまり，$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ を満たす無数の$\boldsymbol{x}$ が存在する．

⇓

観測データとの一致度という要請だけでは解を決めることができない．

⇓

観測データとの一致度以外に解が持つ望ましい性質を組み込む．

⇓

$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ を満たす$\boldsymbol{x}$ のなかで，最も望ましい性質を持つものを最適解とする．

望ましい性質として解のノルム最小を再び用いれば

$\hat{\boldsymbol{x}}=\arg\min_{\boldsymbol{x}}\|\boldsymbol{x}\|^2$，subject to $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ から $\boldsymbol{x}$ を求める．

ミニマムノルムの解：　$\hat{\boldsymbol{x}}=\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{H}^T)^{-1}\boldsymbol{y}$　が得られる．

（最小二乗解：$\hat{\boldsymbol{x}}=(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{y}$ ）

ミニマムノルム解の性質

$$\hat{\boldsymbol{x}} = \boldsymbol{H}^T\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\left(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{\varepsilon}\right) = \underbrace{\boldsymbol{H}^T\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}\boldsymbol{x}}_{\neq \boldsymbol{x}} + \boldsymbol{H}^T\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{\varepsilon}$$

ミニマムノルム解は不偏性を持たない．つまり，線形推定量であるが，線形普遍推定量ではない．

先ほどの最小二乗解場合：

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{y} = \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{\varepsilon}\right) = \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{x} + \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon}$$
$$= \boldsymbol{x} + \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{\varepsilon}$$

データのノイズを考慮した場合のミニマムノルム解

ミニマムノルム解 : $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|^2$ subject to $\underbrace{\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}}_{\Uparrow}$ から解を求める.

観測データに一致する $\boldsymbol{x}$

$\boldsymbol{y}$ にノイズが重畳している場合，データとの一致度をあまり追求しても意味がない.

$$\Downarrow$$

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|^2 \quad \text{subject to} \quad \left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 \le d \text{ から解を求める.}$$

$$\Downarrow$$

最適解は制約条件の境界に存在することが多く，以下とほとんど等価である.

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|^2 \quad \text{subject to} \quad \left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 = d$$

$$\Downarrow$$

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F : \quad F = \xi\left\|\boldsymbol{x}\right\|^2 + \left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2$$

$$\Downarrow$$

$$\hat{\boldsymbol{x}} = \boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{H}^T + \xi\boldsymbol{I}\right)^{-1}\boldsymbol{y}$$（正則化ミニマムノルム解と呼ばれる）

ノルム最小の基準で解を求めることの意味

データとの一致度がある範囲内である解が複数ある場合，なるべく「小さな」解を選ぶという考え方

「オッカムのかみそり」と言われる

ウィキペディアより

オッカムの剃刀（オッカムのかみそり、Occam‘s razor）とは、「ある事柄を説明するためには、必要以上に多くを仮定するべきでない」とする指針。もともとスコラ哲学にあり、14世紀の哲学者・神学者のオッカムが多用したことで有名になった。様々なバリエーションがあるが、20世紀にはその妥当性を巡って科学界で議論が生じた。「剃刀」という言葉は、説明に不要な存在を切り落とすことを比喩しており、そのためオッカムの剃刀は思考節約の原理や思考節約の法則、思考経済の法則とも呼ばれる。またケチの原理と呼ばれることもある。

ある事柄を説明するのに複数の説明がある場合に，最も簡単な説明を選ぶ

解の大きさを表すのに，他のやり方は??

ベクトル： $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$ のノルム

ベクトルの「大きさ」を表す“なんらか”の量

L_2-ノルム： $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^{N} x_j^2}$ （代表的）

L_∞-ノルム： $\|\boldsymbol{x}\|_\infty = \max(x_1, \ldots, x_N)$

L_1-ノルム： $\|\boldsymbol{x}\|_1 = \sum_{j=1}^{N} |x_j|$

L_p-ノルム： $\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{j=1}^{N} |x_j|^p}$

L_0-ノルム： $\|\boldsymbol{x}\|_0 = \sum_{j=1}^{N} \theta(x_j)$ where $\theta(x_j) = \begin{cases} 1 & x_j \neq 0 \\ 0 & x_j = 0 \end{cases}$

L_1 -ノルム正則化ミニマムノルム解

L_2 -ノルム正則化

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{j=1}^{N} x_j^2 \quad \text{subject to} \quad \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2 \le d$$ から解を求める.

L_1 -ノルム正則化

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{j=1}^{N} \left| x_j \right| \quad \text{subject to} \quad \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\|^2 \le d$$ から解を求める.

全く同じ理屈で最適化問題のコスト関数は以下のようになる.

$$\hat{\boldsymbol{x}} = \arg\min F : \quad F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\| + \xi \sum_{j=1}^{N} | x_j |$$

この最適化問題はクローズドフォームの解を持たない. 数値計算解を求める.

L_1 -ノルム正則化ミニマムノルム解はスパースな解を与える.

> $\boldsymbol{x}$ の要素 $x_1,\ldots,x_N$ で, ほとんどのものがゼロ, わずかなものがノンゼロとなる解

スパースな解を与える最も直截的な方法：

L_0 -ノルム正則化ミニマムノルム解

ベクトル $\boldsymbol{x}$ のL_0ノルム: $\left\|\boldsymbol{x}\right\|_0 = \sum_{j=1}^{N} \theta(x_j)$

ここで, $\theta(x_j) = \begin{cases} 1 & x_j \neq 0 \\ 0 & x_j = 0 \end{cases}$

L_0 -ノルム正則化

$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{j=1}^{N} \theta(x_j)$ subject to $\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 \leq d$ から解を求める.

ゼロでない要素の数

この解を数値計算で求めようとしても，非常に時間がかかる.
NP-hard問題と呼ばれるものになる.

L_p-正則化の解：コスト関数の形はどう異なるか

コスト関数の一般形

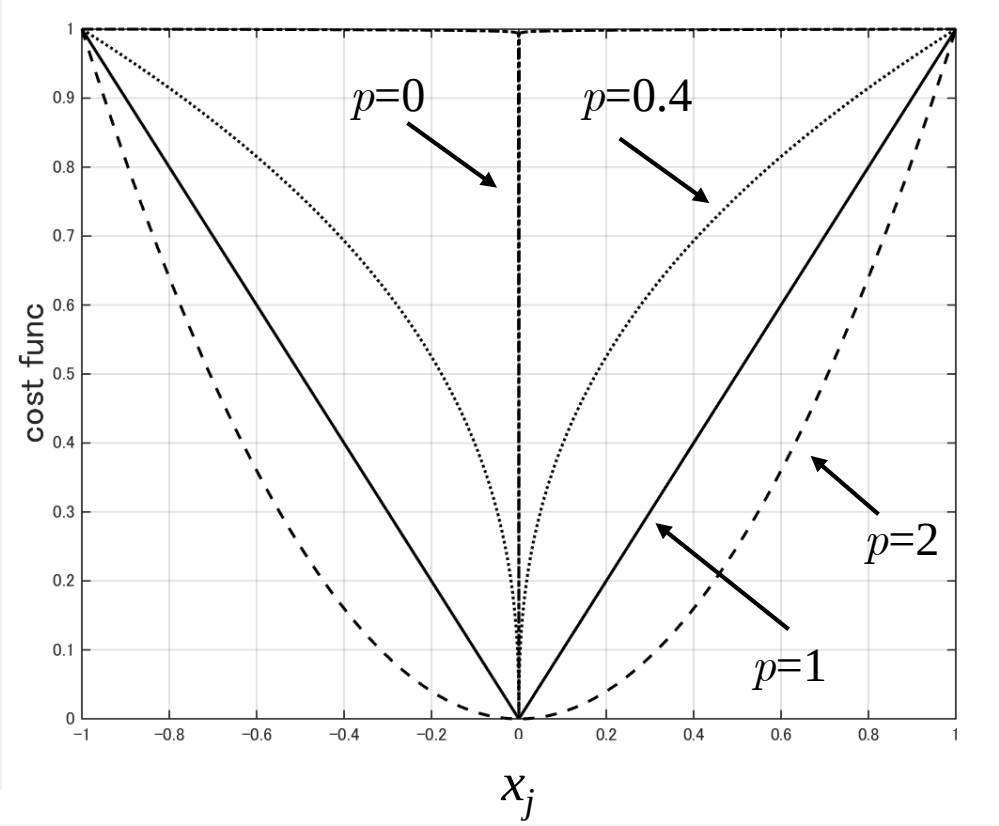
$$\hat{\boldsymbol{x}} = \arg\min F: \quad F = \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \right\| + \xi\phi(\boldsymbol{x})$$

L_0-正則化： $\phi(\boldsymbol{x}) = \sum_{j=1}^{N} \theta(x_j)$

L_1-正則化： $\phi(\boldsymbol{x}) = \sum_{j=1}^{N} \left| x_j \right|$

L_2-正則化： $\phi(\boldsymbol{x}) = \sum_{j=1}^{N} x_j^2$

L_p-正則化： $\phi(\boldsymbol{x}) = \sqrt[p]{\sum_{j=1}^{N} \left| x_j \right|^p}$

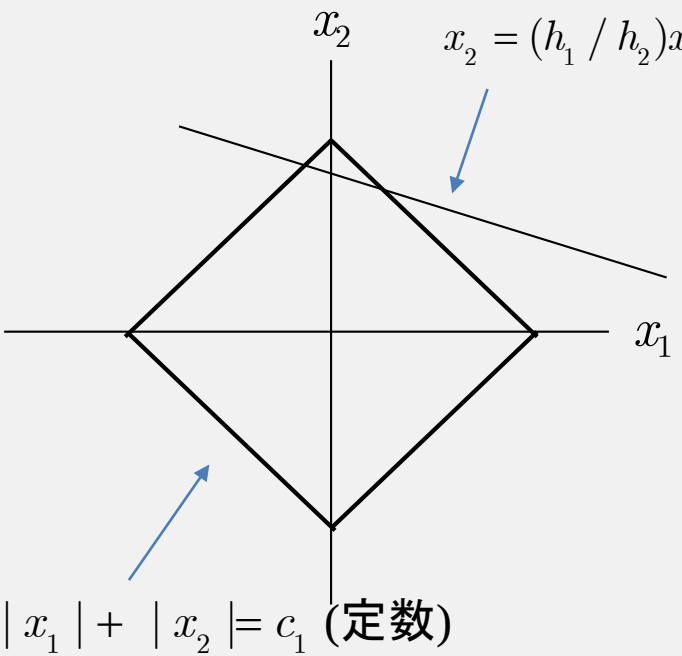


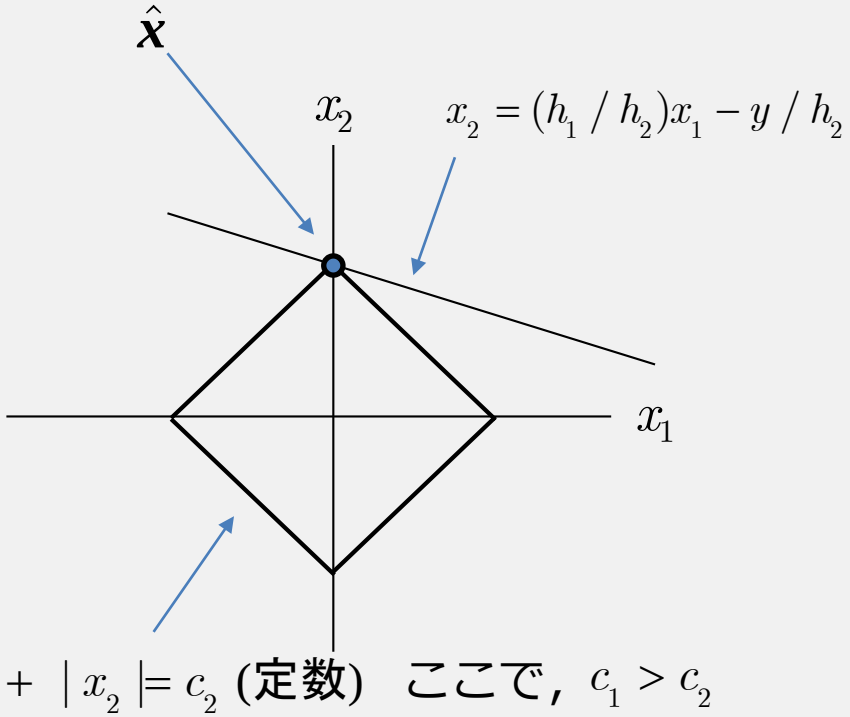
$p > 1$ ：スパースな解とならない
$p \leq 1$ ： スパースな解となる.

2次元問題で考えてみる：L_1-正則化の解

$y = \boldsymbol{H}\boldsymbol{x} = \underbrace{\begin{bmatrix} h_1 & h_2 \end{bmatrix}}_{\boldsymbol{H}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ のモデルで1個の観測値 y から2個の未知量 x_1とx_2 を推定する.

L_1-正則化： $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}(|x_1| + |x_2|)$ subject to $y = h_1 x_1 + h_2 x_2$



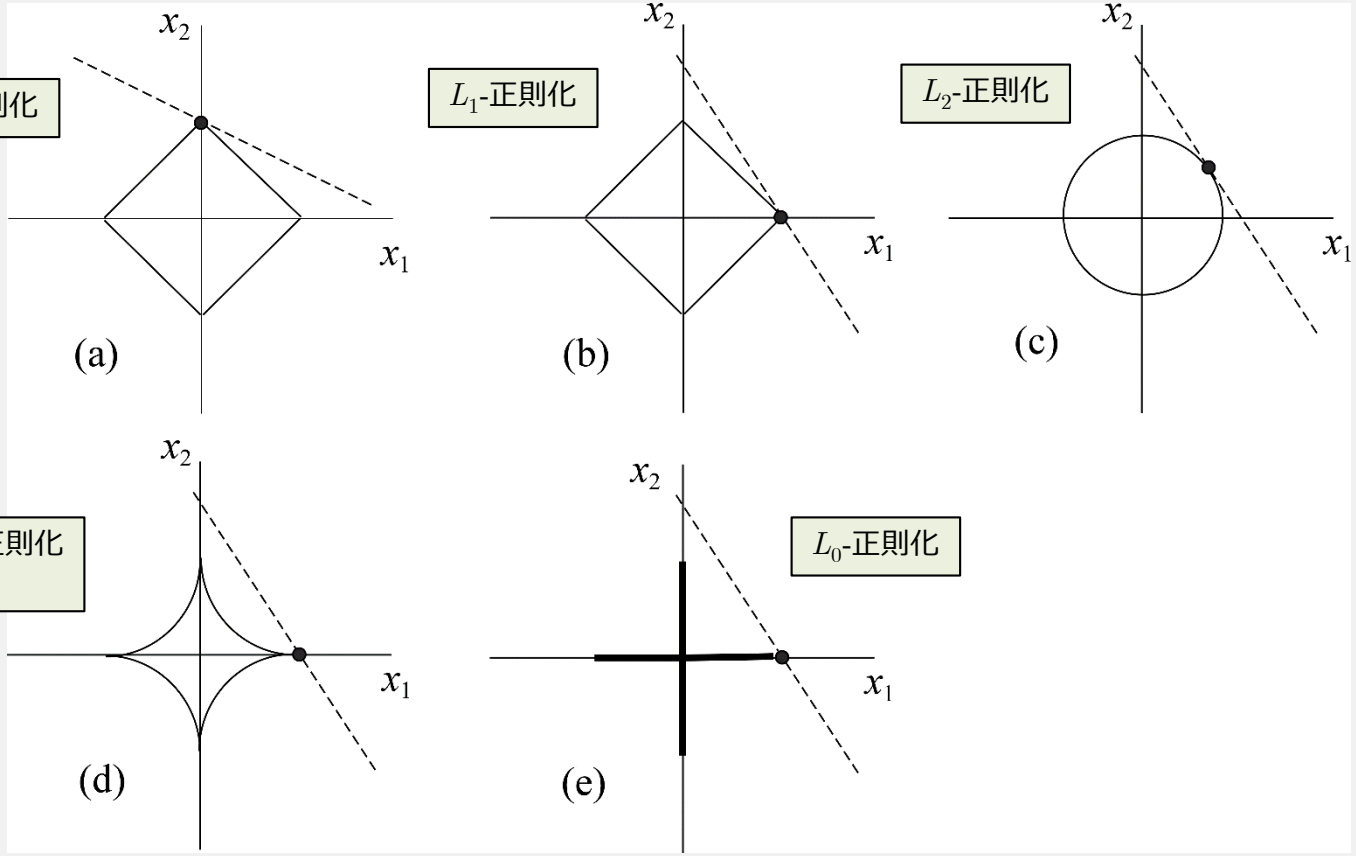


この例では $x_1 = 0,\ x_2 \neq 0$ である解が得られる.

2次元問題で考えてみる：他の制約条件の場合

L_1-正則化： $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}(|x_1| + |x_2|)$ subject to $y = h_1 x_1 + h_2 x_2$

L_2-正則化： $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}(x_1^2 + x_2^2)$ subject to $y = h_1 x_1 + h_2 x_2$



L_p-正則化の場合($p \le 1$)，x_1, x_2 のどちらかがゼロの解（スパースな解）が得られる．

L_2-正則化の場合，x_1, x_2 ともにノンゼロの解（ノンースパースな解）が得られる．

ベイズ推定

ベイズ統計学をもとにした未知量の推定

条件付き確率とベイズ定理

Aが起こる確率：$P(A)$，　Bが起こる確率：$P(B)$，　A とBが両方起こる確率：$P(A,B)$

Bが起た場合にA が起きる確率をB を条件とするA の条件付き確率と呼び$P(A \mid B)$で表す．

$P(A \mid B) = \dfrac{P(A,B)}{P(B)}$ から求まる．

$P(B \mid A) = \dfrac{P(A,B)}{P(A)}$ も成立．

したがって，$P(B \mid A) = \dfrac{P(A \mid B)P(B)}{P(A)}$ （ベイズ定理）が成立．

原因$H_1, H_2, \ldots, H_K$からある結果Aを生じる．

$P(A \mid H_j)$: 原因H_jから結果Aを生じる確率．　←　この確率を求めるのが順問題

$P(H_j \mid A)$: 結果A を生じた原因がH_jであった確率．　←　この確率を求めるのは逆問題

ベイズ定理：$P(H_j \mid A) = \dfrac{P(A \mid H_j)P(H_j)}{P(A)} = \dfrac{P(A \mid H_j)P(H_j)}{\sum_{i=1}^{K} P(A \mid H_i)P(H_i)}$

確率の定義

ベイズ推定では全ての変数は確率変数であると考える

ある事柄Aが起こる確率を$P(A)$と表す．実は，この$P(\mathrm{A})$をどのように定義するかは，いくつかの考え方がある．

代表的な（ベイズ統計学以外の）考え方は，同じ条件のもとで多数回繰り返すことができる実験や観測の結果の１つが事柄Aであり，Aが起こるかどうかは偶然によって決まるものと考える．この時，Aの起こる確率$P(\mathrm{A})$は

$$P(\mathrm{A}) = \frac{\text{Aの起こる場合の数}}{\text{Aも含めて起こりうるすべての場合の数}}$$

と定義すると考える．（頻度主義）

一回限りの出来事に確率を割り当てられない．

例：「明日，雨の降る確率」，「かつて火星に生命が存在した確率」

一方，ベイズ統計学ではある事柄Aが起こる確率$P(\mathrm{A})$は，事柄Aが起こるか否かに関する観測者の確信の度合いであると考える．この考え方を主観確率と呼ぶ．主観確率は人間の考え方に近く，天気予報で用いられる「降水確率」などはこれに近いものであろう．

確率密度分布に対するベイズ定理

条件付き確率密度：$p(x_1 \mid x_2) = \dfrac{p(x_1, x_2)}{p(x_2)}$

ベイズ定理：　$p(x_2 \mid x_1) = \dfrac{p(x_1 \mid x_2)p(x_2)}{p(x_1)} = \dfrac{p(x_1 \mid x_2)p(x_2)}{\int_{-\infty}^{\infty} p(x_1 \mid x_2)p(x_2)dx_2}$

上式の分母は　$\int_{-\infty}^{\infty} p(x_2 \mid x_1)dx_2 = \dfrac{\int_{-\infty}^{\infty} p(x_1 \mid x_2)p(x_2)dx_2}{\int_{-\infty}^{\infty} p(x_1 \mid x_2)p(x_2)dx_2} = 1$　とするために必要な

規格化定数であるため，実際の応用ではベイズ定理は

$$p(x_2 \mid x_1) \propto p(x_1 \mid x_2)p(x_2)$$

として用いられることも多い.

線形離散モデル：線型方程式 $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$ の解法

$$\text{未知量}:\boldsymbol{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix},\quad \text{観測データ}:\boldsymbol{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

の間に　$\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{\varepsilon}$　の関係がある.

$\boldsymbol{H}$ が既知であるとの前提のもとに，観測 $\boldsymbol{y}$ から未知な量 $\boldsymbol{x}$ を求める

未知量$\boldsymbol{x}$も確率変数として考え，以下の確率分布を考慮する.

$p(\boldsymbol{x})$: 未知量$\boldsymbol{x}$についての確率分布. 事前分布(prior probability distribution)と呼ばれる.

$p(\boldsymbol{y}\mid\boldsymbol{x})$: $\boldsymbol{x}$ を与えた場合に観測データ $\boldsymbol{y}$ を得る確率分布. 最尤推定の尤度に等しい.

$p(\boldsymbol{x}\mid\boldsymbol{y})$: データ $\boldsymbol{y}$ を与えた場合に $\boldsymbol{x}$ を得る確率分布. ベイズ推定ではこの確率を基にして $\boldsymbol{x}$ を推定する. 事後分布(posterior probability distribution)と呼ばれる.

ベイズ推定では事前確率分布$p(\boldsymbol{x})$と$p(\boldsymbol{y} \mid \boldsymbol{x})$からベイズの定理

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$$

により事後分布を計算する.

この時，事前分布$p(\boldsymbol{x})$をどう決めるかが問題となる．解$\boldsymbol{x}$ に関して何も先験情報がない場合には$p(\boldsymbol{x}) =$ 定数とせざるを得ず，ベイズ定理は

$$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})$$

となり，ベイズ推定は最尤推定に等しくなる.

$p(\boldsymbol{x}) =$ 定数である事前分布を無情報事前分布(non-informative prior)と呼ぶ.

したがって，ベイズ推定では事前確率分布をどう決めるかが重要である.

事前分布

事前分布をどう決めるか？

- 未知量 $\boldsymbol{x}$ についての先見情報（あらかじめ分かっている情報）に基づいて決める．（先見情報を確率の形で表現する．）
- 確率分布の形まで決めることのできる先見情報が入手できることはまれである．
- 確率分布は計算しやすいように決めることが普通行われる．

> **ベイズの定理**：$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$ **において，**
>
> **与えられた**$p(\boldsymbol{y} \mid \boldsymbol{x})$ **に対して**$p(\boldsymbol{x})$**と**$p(\boldsymbol{x} \mid \boldsymbol{y})$**が同じ分布となる場合，**
>
> **共役事前分布**(conjugate prior)**と呼ぶ．**

（例）$p(\boldsymbol{y} \mid \boldsymbol{x})$ **が正規分布の場合，** $p(\boldsymbol{x})$**と**$p(\boldsymbol{x} \mid \boldsymbol{y})$**はどちらも正規分布となる．**

分布の形は共役事前分布を選ぶことが一般的に行われる．

事前分布

事前確率分布の形は計算しやすいように決めるとして，分布のパラメータはどう決めるか？

例えば，事前分布に正規分布： $p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha \boldsymbol{I})$ を仮定したとして，
α はどう決めるか？

一般的なベイズ法

未知量 $\boldsymbol{x}$ についての先見情報（あらかじめ分かっている情報）に基づいて決める.

経験ベイズ法(empirical Bayes)

分布のパラメータは得られた観測データ $\boldsymbol{y}$ から決める

経験ベイズ法でなければベイズ推定の意味はあまり無い.

事後確率分布$p(\boldsymbol{x} \mid \boldsymbol{y})$を求めた後，$\boldsymbol{x}$ の推定結果をどうやって求めるのか？

- MAP推定解：

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})$$

から，最適推定解 $\hat{\boldsymbol{x}}$ を求める．Maximum a posteriori(MAP) estimate（MAP推定解）と呼ばれる．

- MMSE推定解：

推定解と真の値との誤差を最小にする，すなわち，

$$\hat{\boldsymbol{x}} = \arg\min_{\hat{\boldsymbol{x}}} E[\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}\right\|^2] = \arg\min_{\hat{\boldsymbol{x}}} \int (\hat{\boldsymbol{x}} - \boldsymbol{x})^T (\hat{\boldsymbol{x}} - \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{x}$$

から，最適推定解 $\hat{\boldsymbol{x}}$ を求める．Minimum mean squared error(MMSE) estimate（MMSE推定解）と呼ばれる．

$\frac{\partial}{\partial \hat{\boldsymbol{x}}} \int (\hat{\boldsymbol{x}} - \boldsymbol{x})^T (\hat{\boldsymbol{x}} - \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{x} = 0$ より $\hat{\boldsymbol{x}} = \int \boldsymbol{x} p(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{x}$ を得るので

MMSE解は事後確率分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$の期待値に一致する．

$p(\boldsymbol{x} \mid \boldsymbol{y})$が正規分布の場合，MAP推定解とMMSE推定解は一致する．

精度について

ベイズ推定では，精度＝分散の逆数　が良く用いられる．

分散 σ^2 の代わりに精度 γ を用いれば，

$$1\text{次元正規分布}: p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(x-\mu)^2}{2\sigma^2}] = \sqrt{\frac{\gamma}{2\pi}}\exp[-\frac{\gamma}{2}(x-\mu)^2]$$

共分散行列 Σ の代わりに精度行列 Φ をもちいれば，

$$\text{多次元正規分布}: p(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]$$

$$= \frac{\left|\boldsymbol{\Phi}\right|^{1/2}}{(2\pi)^{N/2}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\boldsymbol{x}-\boldsymbol{\mu})]$$

これらは，N(確率変数 | 期待値, 分散)の表記法のルールから

$$p(x) = N(x \mid 0, \gamma^{-1}) \quad p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \boldsymbol{\Phi}^{-1})$$

と表す．

線形離散モデルに対するベイズ推定解の導出

$$\text{未知量}: \boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \quad \text{観測データ}: \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

の間に $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$ の関係がある．$\boldsymbol{H}$ が既知であるとの前提のもとで，観測 $\boldsymbol{y}$ から未知な量 $\boldsymbol{x}$ をベイズ推定を用いて求めてみよう．

確率モデル：

ノイズの確率分布：　$p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon} \mid 0, \beta^{-1}\boldsymbol{I})$

事前確率分布：　$p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$

ベイズ推定解の求め方：

事前確率分布 $p(\boldsymbol{x})$ と尤度 $p(\boldsymbol{y} \mid \boldsymbol{x})$ からベイズ定理：$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$

により事後分布を計算し，MAP推定解を求める．

正規分布モデルでの事後確率分布の計算

事前確率分布： $p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$

データ確率分布： $p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$

ベイズ定理: $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$**を用いて，事後確率分布**$p(\boldsymbol{x} \mid \boldsymbol{y})$**を求める.**

事後確率分布も正規分布であるので，$p(\boldsymbol{x} \mid \boldsymbol{y}) = N(\boldsymbol{x} \mid \bar{\boldsymbol{x}}, \boldsymbol{\Gamma}^{-1})$ とすれば，求めるのは分布のパラメータ $\bar{\boldsymbol{x}}$ と $\boldsymbol{\Gamma}$ である.

定数倍を無視した以下の式を用いる：

$$p(\boldsymbol{x}) \propto \exp[-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}]$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp[-\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})]$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto \exp[-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}})^T\boldsymbol{\Gamma}(\boldsymbol{x} - \bar{\boldsymbol{x}})]$$

具体的な計算：

$$p(\boldsymbol{x}) \propto \exp[-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}]$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp[-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})]$$

$$\Downarrow$$

ベイズ定理: $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) \propto \exp[-\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}) - \frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}]$

事後分布：$p(\boldsymbol{x} \mid \boldsymbol{y}) \propto \exp[-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T\boldsymbol{\Gamma}(\boldsymbol{x}-\bar{\boldsymbol{x}})]$

マゼンダの部分を比較すれば，

$$-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T\boldsymbol{\Gamma}(\boldsymbol{x}-\bar{\boldsymbol{x}}) = -\frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}) - \frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}$$

カッコを払い$\boldsymbol{x}$についてまとめれば，

$$-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Gamma}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{\Gamma}\bar{\boldsymbol{x}} + C = -\frac{1}{2}\boldsymbol{x}^T(\alpha\boldsymbol{I} + \beta\boldsymbol{H}^T\boldsymbol{H})\boldsymbol{x} + \boldsymbol{x}^T\beta\boldsymbol{H}^T\boldsymbol{y} + C'$$

$\boldsymbol{x}$の係数（赤と青の部分）を比較し，以下を得る．

$$\boldsymbol{\Gamma} = \alpha\boldsymbol{I} + \beta\boldsymbol{H}^T\boldsymbol{H}$$

$$\bar{\boldsymbol{x}} = \beta\boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{y} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\alpha}{\beta}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

$\Leftarrow$ $\boldsymbol{x}$ のベイズ推定解

正規分布の確率モデルでベイズ推定解を導く
ーコスト関数の考察からー

MAP推定解は $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})$ を最大にする $\boldsymbol{x}$ である.

コスト関数： $F = \log p(\boldsymbol{x} \mid \boldsymbol{y}) = \log p(\boldsymbol{y} \mid \boldsymbol{x}) + \log p(\boldsymbol{x})$

ノイズの分布を正規分布：$p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon} \mid 0, \beta^{-1}\boldsymbol{I})$と仮定すれば,

$p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$　であるので,

$$\log p(\boldsymbol{y} \mid \boldsymbol{x}) = \log \exp[-\frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})] = -\frac{\beta}{2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2$$

ベイズ推定解におけるコスト関数は以下のように表現できる：

$$F = \frac{\beta}{2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 - \log p(\boldsymbol{x})$$

観測データとの一致度　　解に対する制約条件

- ベイズ推定は，解の制約条件を事前分布の形で組み込むことができる.

正規分布の確率モデルでベイズ推定解を導く
ーコスト関数の考察からー

コスト関数： $F = \frac{\beta}{2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 - \log p(\boldsymbol{x})$

事前分布に正規分布：$p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$ を仮定すれば：

$$\log p(\boldsymbol{x}) = \log \exp[-\frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}] = -\frac{\alpha}{2}\left\|\boldsymbol{x}\right\|^2 \text{より}$$

コスト関数：$F = \frac{\beta}{2}\left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 + \frac{\alpha}{2}\left\|\boldsymbol{x}\right\|^2 \approx \left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\right\|^2 + \frac{\alpha}{\beta}\left\|\boldsymbol{x}\right\|^2$ を得る.

F を $\boldsymbol{x}$ で微分してゼロとおけば:

ミニマムノルム解 $\hat{\boldsymbol{x}} = \beta\boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{y} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\alpha}{\beta}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$ が導かれる

- ノルム最小の条件を「未知量 $\boldsymbol{x}$ は分散σ^2の正規分布に従う」として導いたことに相当.
- 正則化定数はどうやって決めたらよいであろうか？（最尤推定の範囲では決める方法は無かった.）

ハイパーパラメータ

モデル：$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$ と以下の確率分布の仮定のもとで，

事前確率分布： $p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$

ノイズの分布： $p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon} \mid 0, \beta^{-1}\boldsymbol{I})$

未知量 $\boldsymbol{x}$ のベイズ推定解は

$$\overline{\boldsymbol{x}} = \beta \boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{y} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\alpha}{\beta}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

と求まる．

しかし上式は（通常は）未知な量である α と β を含む．α と β は未知パラメータ $\boldsymbol{x}$ の確率分布を記述するパラメータなのでハイパーパラメータと呼ばれる．

どうやってハイパーパラメータを推定するのか??

$\boldsymbol{x}$ だけでなく，ハイパーパラメータ $\boldsymbol{\theta} = \{\alpha, \beta\}$ も未知量である．

ベイズ原理主義的な取り組み方

事後分布をもとめ，MAPあるいはMMSE推定解を求める

未知量 $\boldsymbol{x}$ とハイパーパラメータ $\boldsymbol{\theta}$ の両方のパラメータに対する結合事後分布：

$$p(\boldsymbol{x},\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x},\boldsymbol{\theta})}{\iint p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}d\boldsymbol{\theta}}$$

を求め，これを最大とする $\boldsymbol{x}$ と $\boldsymbol{\theta}$ を求める．

リーズナブルな $p(\boldsymbol{x},\boldsymbol{\theta})$ を仮定できない，などの問題があり，この結合事後分布を求める事は一般的にはむずかしい．

さてどうするか？　次善の策：

2つの方針が知られている：

1）周辺尤度：$\log p(\boldsymbol{y} \mid \boldsymbol{\theta})$ の最大化により，ハイパーパラメータ $\boldsymbol{\theta}$ を推定する．

1．周辺尤度を計算する．

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta}) d\boldsymbol{x} = \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{\theta}) d\boldsymbol{x}$$

2．周辺尤度の変わりに平均データ尤度を計算する．➡ EMアルゴリズム

2）変分ベイズ法

条件付確率 $p(\boldsymbol{x} \mid \boldsymbol{y})$ と $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ は独立との仮定を置いて，結合事後分布

$$p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y}) p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

とファクトライズし，結合事後分布を近似した確率分布を求める．

方針１－周辺尤度とは

$\boldsymbol{\theta}$ のMAP推定解: $\hat{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta} \mid \boldsymbol{y})$ を求めたい.

$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$と計算できるが, $\boldsymbol{\theta}$ についての事前分布$p(\boldsymbol{\theta})$を決めることは難しい.

無情報事前分布 $p(\boldsymbol{\theta}) = const$ を仮定して：

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta})$$

であるので，実際には $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta})$ を用いざるをえない.

⇑

周辺尤度(marginal likelihood)あるいはエビデンス(evidence)と呼ばれる

実際に周辺尤度を計算してみると

周辺尤度は以下で計算できる.

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta}) d\boldsymbol{x} = \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{\theta}) d\boldsymbol{x}$$

正規モデルの場合：

$$p(\boldsymbol{x} \mid \alpha) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \beta) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$$

周辺尤度：

$$p(\boldsymbol{y} \mid \alpha, \beta) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \beta) p(\boldsymbol{x} \mid \alpha) d\boldsymbol{x} = \int N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I}) N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})\ d\boldsymbol{x}$$

と計算できる．しかし，右辺の積分はやはり計算がやっかいである.

2つの方針

1. やっかいな積分を計算する．（後で述べる）
2. この積分を計算せずにすむ方法を見つける ⇒ EMアルゴリズム

EMアルゴリズムー平均データ尤度

周辺尤度を近似した，もっと計算の楽な量は無いのか？

完全データ尤度と呼ばれる以下の量を考える

$$\log p(\boldsymbol{y}, \boldsymbol{x} \mid \alpha, \beta) = \log p(\boldsymbol{y} \mid \boldsymbol{x}, \beta) + \log p(\boldsymbol{x} \mid \alpha)$$

積分を含まず計算しやすいが，未知な量 $\boldsymbol{x}$ を明示的に含むため計算できない.

何か $\boldsymbol{x}$ の代わりになる量を用いて完全データ尤度に近いものを計算できないか？

われわれが知っている $\boldsymbol{x}$ に関する最善の知識は事後分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$ である.

１つの案：$\boldsymbol{x}$ の代わりに $\boldsymbol{x}$ のMAP推定解 $\hat{\boldsymbol{x}}$ を用いて，$\log p(\boldsymbol{y}, \hat{\boldsymbol{x}} \mid \alpha, \beta)$ を求める.

もっといい案：完全データ尤度を事後分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$ で重み付けして $\boldsymbol{x}$ で積分したものを用いる：

$$\Theta(\alpha, \beta) = \int \log p(\boldsymbol{x}, \boldsymbol{y} \mid \alpha, \beta) p(\boldsymbol{x} \mid \boldsymbol{y}) d\boldsymbol{x} = E[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \alpha, \beta)]$$

この $\Theta(\alpha, \beta)$ を平均データ尤度と呼ぶ.

事後分布での平均

平均データ尤度：

$$\Theta(\alpha,\beta)=\int \log p(\boldsymbol{y},\boldsymbol{x}\mid\alpha,\beta)p(\boldsymbol{x}\mid\boldsymbol{y})d\boldsymbol{x}=E[\log p(\boldsymbol{y},\boldsymbol{x}\mid\alpha,\beta)]$$

周辺尤度：

$$\log p(\boldsymbol{y}\mid\alpha,\beta)=\log\int p(\boldsymbol{y}\mid\boldsymbol{x},\beta)p(\boldsymbol{x}\mid\alpha)d\boldsymbol{x}=\log\int N(\boldsymbol{y}\mid\boldsymbol{H}\boldsymbol{x},\beta^{-1}\boldsymbol{I})N(\boldsymbol{x}\mid 0,\alpha^{-1}\boldsymbol{I})\ d\boldsymbol{x}$$

平均データ尤度を周辺尤度の代わりに計算し，ハイパーパラメータの値を求めるアルゴリズムをExpectation Maximization (EM)アルゴリズムと呼ぶ.

平均データ尤度の計算には事後分布が必要.

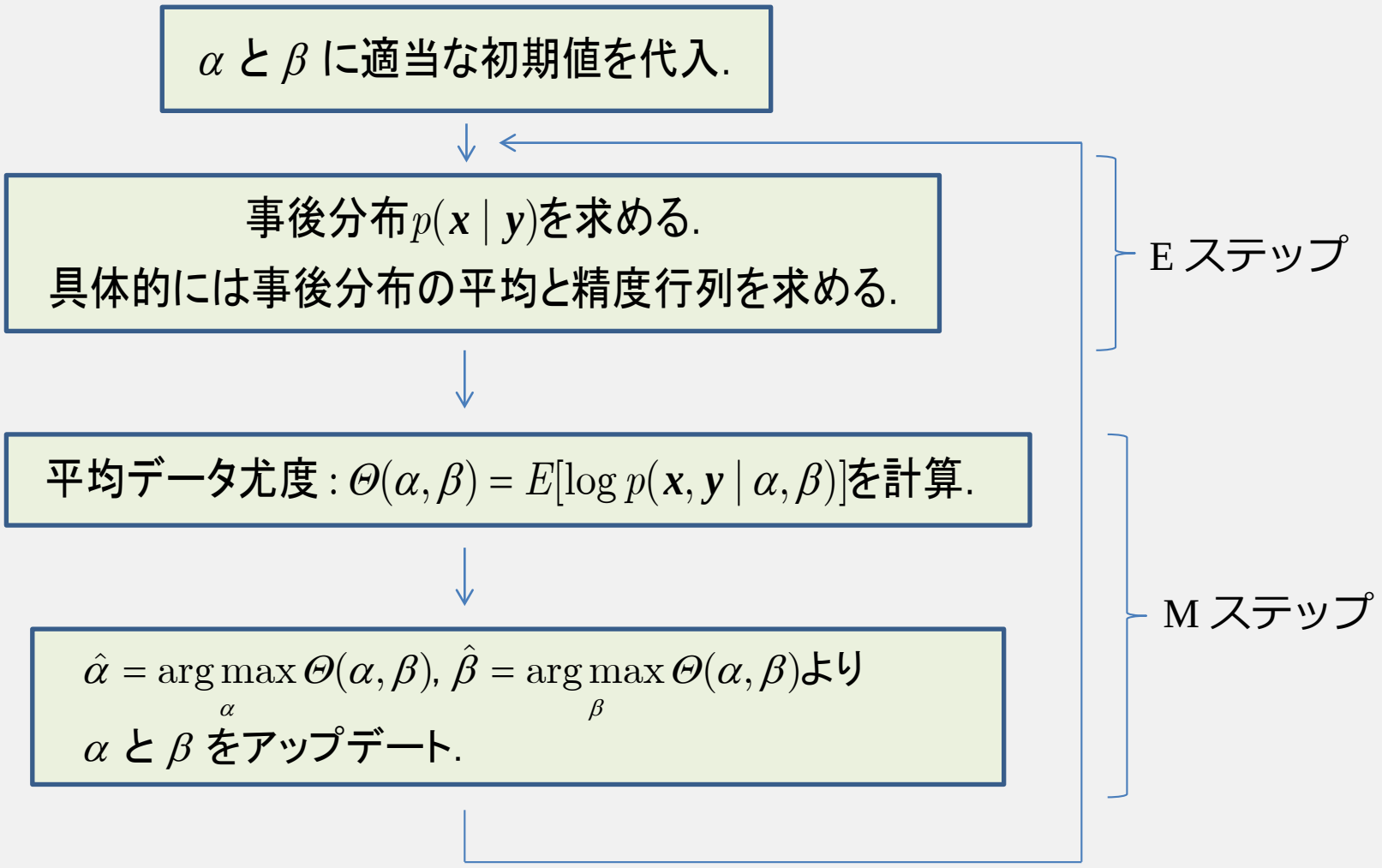
⇔

事後分布$p(\boldsymbol{x}\mid\boldsymbol{y})$の計算には
ハイパーパラメータα と β が必要.

↓

EMアルゴリズムは再帰的なアルゴリズムとなる.

Expectation maximization（EM)アルゴリズム



正規確率モデルにおける平均データ尤度の具体的計算

正規確率モデル(復習)：

事前確率分布： $p(\boldsymbol{x} \mid \alpha) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I}) = \dfrac{|\alpha|^{N/2}}{(2\pi)^{N/2}} \exp[-\dfrac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x}]$

データ確率分布： $p(\boldsymbol{y} \mid \boldsymbol{x}, \beta) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I}) = \dfrac{|\beta|^{M/2}}{(2\pi)^{M/2}} \exp[-\dfrac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})]$

⇩

事後確率分布 $p(\boldsymbol{x} \mid \boldsymbol{y}) = N(\boldsymbol{x} \mid \bar{\boldsymbol{x}}, \boldsymbol{\Gamma}^{-1})$ とすれば，ベイズ定理より

$$\boldsymbol{\Gamma} = \alpha\boldsymbol{I} + \beta\boldsymbol{H}^T\boldsymbol{H}$$

$$\bar{\boldsymbol{x}} = \beta\boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{y} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\alpha}{\beta}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

⇧

となる.

L2正則化ミニマムノルム解

正規確率モデルにおける平均データ尤度の具体的計算

完全データ尤度：

$$\log p(\boldsymbol{x}, \boldsymbol{y} \mid \alpha, \beta) = \log p(\boldsymbol{y} \mid \boldsymbol{x}, \beta) + \log p(\boldsymbol{x} \mid \alpha)$$
$$= \frac{N}{2}\log\alpha - \frac{\alpha}{2}\boldsymbol{x}^T\boldsymbol{x} + \frac{M}{2}\log\beta - \frac{\beta}{2}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})$$

$\boldsymbol{y}$ は値がわかっているが，$\boldsymbol{x}$ も含むため計算できない．

平均データ尤度：

$$\Theta(\alpha, \beta) = E[\log p(\boldsymbol{x}, \boldsymbol{y} \mid \alpha, \beta)]$$
$$= \frac{N}{2}\log\alpha - \frac{\alpha}{2}E[\boldsymbol{x}^T\boldsymbol{x}] + \frac{M}{2}\log\beta - \frac{\beta}{2}E[(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})]$$
$$= \frac{N}{2}\log\alpha - \frac{\alpha}{2}[\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}} + \mathrm{tr}(\boldsymbol{\Gamma}^{-1})] + \frac{M}{2}\log\beta - \frac{\beta}{2}[(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}})^T(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}) + \mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\Gamma}^{-1})]$$

未知量 $\boldsymbol{x}$ は含まない．代わりに，事後分布のパラメータ $\bar{\boldsymbol{x}}$ と $\boldsymbol{\Gamma}$ を含む．

ハイパーパラメータの更新式の導出

平均データ尤度：

$$\Theta(\alpha,\beta)=\frac{N}{2}\log\alpha-\frac{\alpha}{2}[\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}+\mathrm{tr}(\boldsymbol{\Gamma}^{-1})]+\frac{M}{2}\log\beta-\frac{\beta}{2}[(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})^T(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})+\mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\Gamma}^{-1})]$$

をハイパーパラメータ α と β で微分して，以下の更新式を得る．

$$\frac{\partial}{\partial\alpha}\Theta(\alpha,\beta)=\frac{N}{2\alpha}-\frac{1}{2}[\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}+\mathrm{tr}(\boldsymbol{\Gamma}^{-1})]=0\ \Rightarrow\ \alpha^{-1}=\frac{1}{N}[\bar{\boldsymbol{x}}^T\bar{\boldsymbol{x}}+\mathrm{tr}(\boldsymbol{\Gamma}^{-1})]$$

$$\frac{\partial}{\partial\beta}\Theta(\alpha,\beta)=\frac{M}{2\beta}-\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})^T(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})+\mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\Gamma}^{-1})]=0$$

$$\Rightarrow\ \beta^{-1}=\frac{1}{M}[(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})^T(\boldsymbol{y}-\boldsymbol{H}\bar{\boldsymbol{x}})+\mathrm{tr}(\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{\Gamma}^{-1})]$$

Bayesianミニマムノルム解

Eステップ: 事後分布のパラメータを更新する

$$\boldsymbol{\Gamma} = \alpha \boldsymbol{I} + \beta \boldsymbol{H}^T \boldsymbol{H}$$

$$\overline{\boldsymbol{x}} = (\boldsymbol{H}^T \boldsymbol{H} + \frac{\alpha}{\beta} \boldsymbol{I})^{-1} \boldsymbol{H}^T \boldsymbol{y}$$

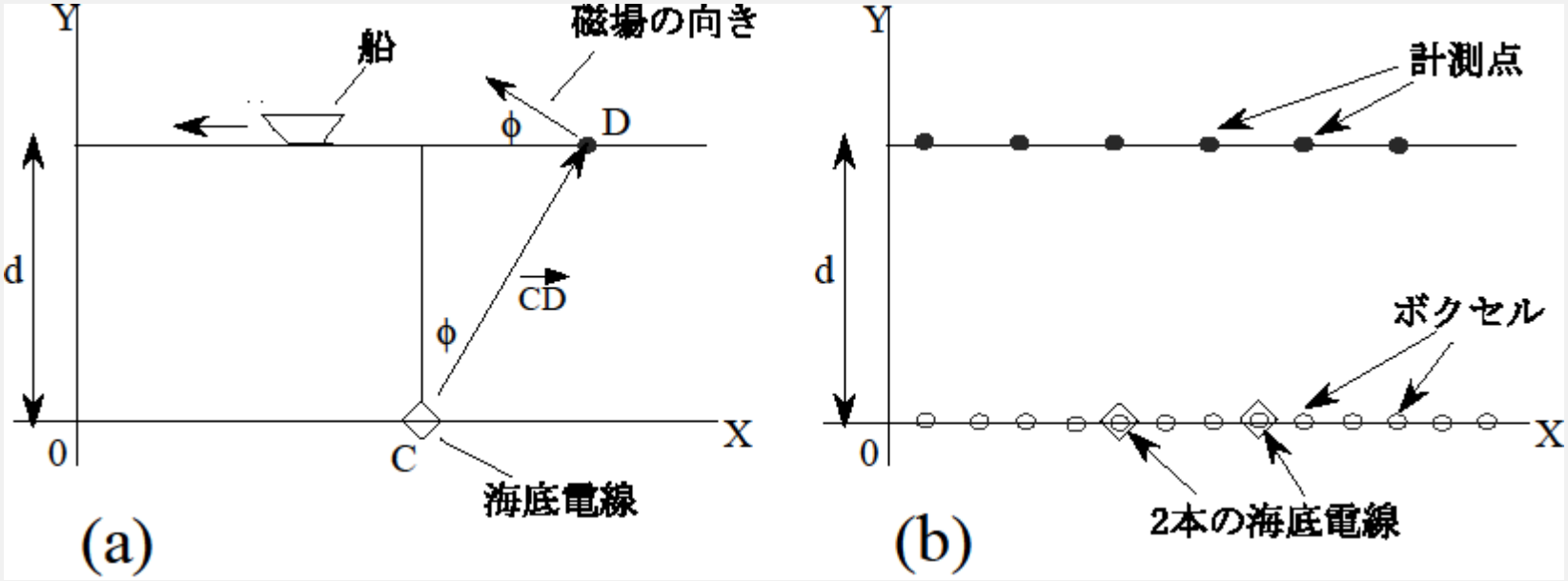
Mステップ: ハイパーパラメータを更新する

$$\hat{\alpha}^{-1} = \frac{1}{N}[\overline{\boldsymbol{x}}^T \overline{\boldsymbol{x}} + \mathrm{tr}(\boldsymbol{\Gamma}^{-1})]$$

$$\hat{\beta}^{-1} = \frac{1}{M}[\left\| \boldsymbol{y} - \boldsymbol{H}\overline{\boldsymbol{x}} \right\|^2 + \mathrm{tr}(\boldsymbol{H}^T \boldsymbol{H} \boldsymbol{\Gamma}^{-1})]$$

ベイズミニマムノルム法は，EステップとMステップを交互に繰り返す．このようにすることで，正則化定数をEMアルゴリズムを用いて与えることができる．

コンピュータシミュレーション：海底電線の位置推定



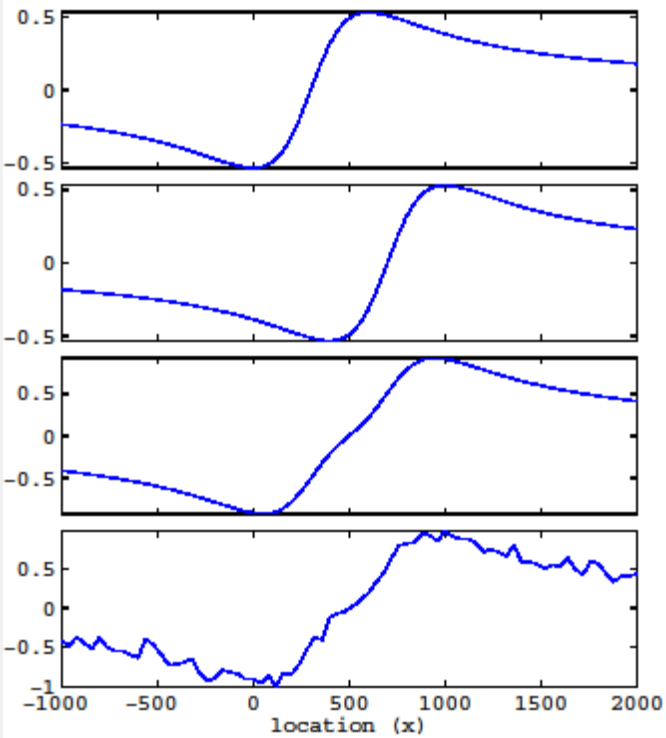
Cにある電線（電流Iが流れている）が海面のDの位置に作る磁場の法線成分:

$$B_{\perp} = \frac{I}{r}\sin\phi: \ r = \mathrm{CD}$$

x=300とx=700の位置に2本の海底電線を仮定

コンピュータシミュレーションに用いたMatlabコードは以下からダウンロード可能：
http://www.kyoritsu-pub.co.jp/bookdetail/9784320085749

水面上の$B_{\perp}$の分布



X=300に1本の電線が有る場合

X=700に1本の電線が有る場合

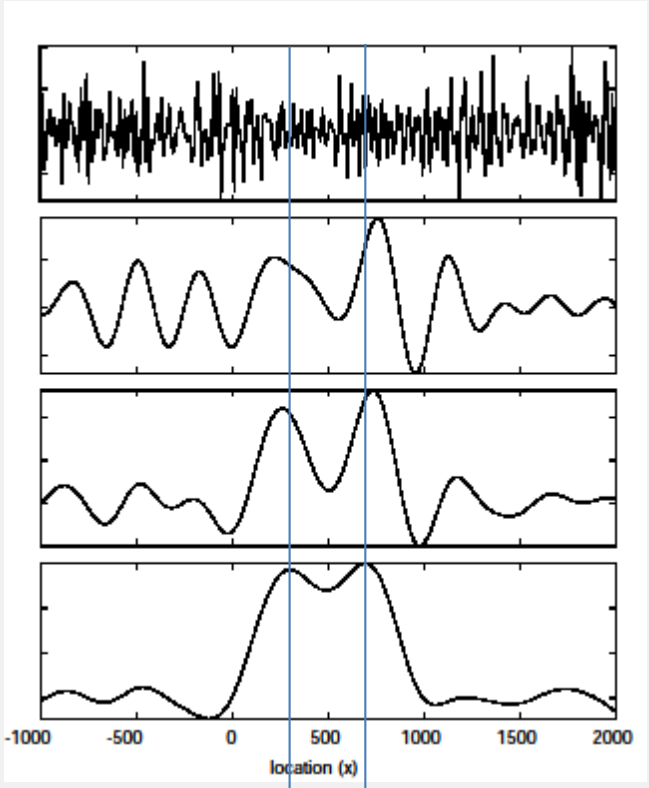
X=300とX=700に2本の電線が有る場合

X=300とX=700に2本の電線が有る場合で，更に正規ノイズを信号対雑音比10で加えた.

海底の電流分布再構成実験

$\boldsymbol{H}$の(i,j)要素$H_{i,j}$：　$H_{i,j} = \dfrac{\sin\phi_{i,j}}{r_{i,j}}$

$\hat{\boldsymbol{x}} = \boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{H}^T + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{y}$

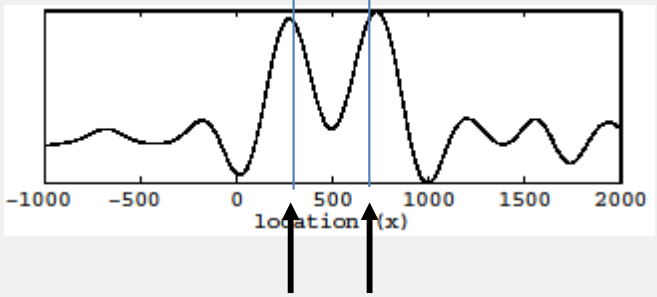


ミニマムノルム解

L2正則化ミニマムノルム解：$\gamma=10^{-4}$

L2正則化ミニマムノルム解：$\gamma=10^{-3}$

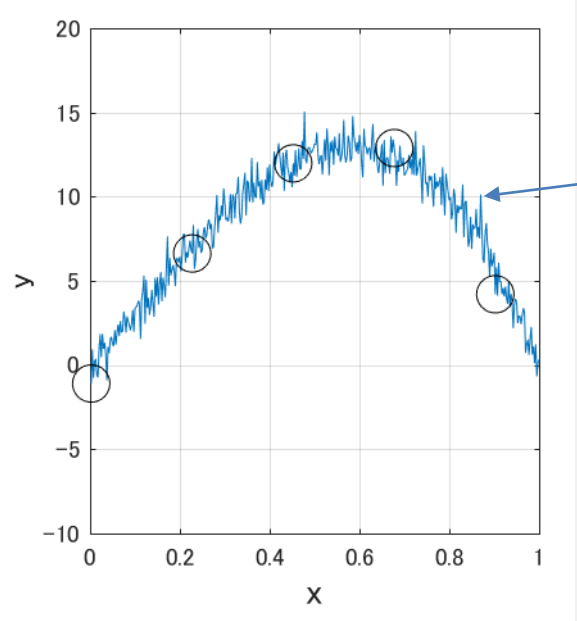
L2正則化ミニマムノルム解：$\gamma=10^{-2}$



ベイズミニマムノルム解
（正則化パラメータをEMにより導出）

２本の電線位置

線形回帰問題への応用



変数 x の[0, 1] の値域に対して，多項式

$$y = -40x^3 + 10x^2 + 30x$$

を仮定して，間隔0.002 で500 点のデータを発生し，SN 比20 で正規乱数をノイズとして加えた.

500点を使うのではなく，85点間隔の5点（丸印で示す）のデータのみが計測されたとして，この5点でこの曲線を回帰することを試みる.

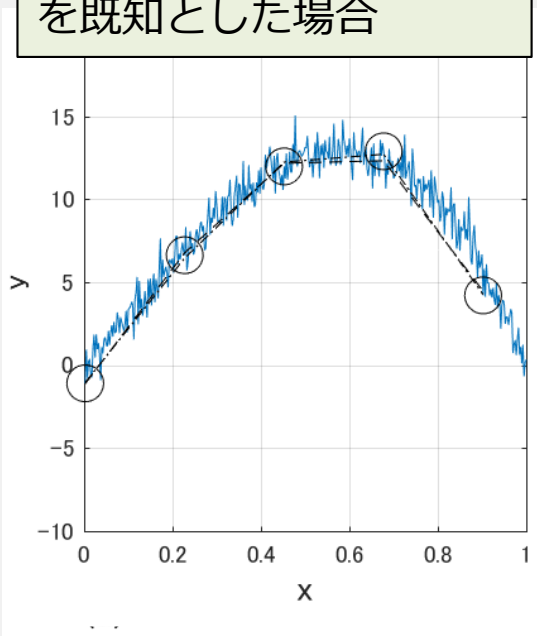
このとき，n次の多項式

$$y = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_0$$

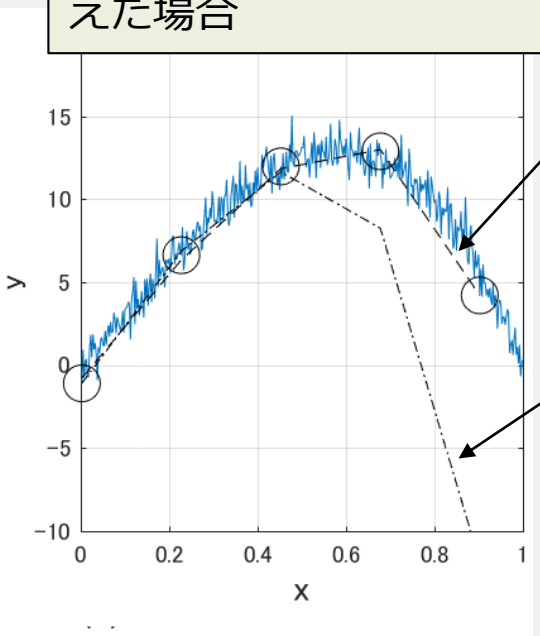
を仮定する.

線形回帰問題への応用

n=3：データの真の次数を既知とした場合



n=6：次数を大きめに与えた場合



ベイズミニマムノルム解

ミニマムノルム解

5点のデータの組： $(x_j, y_j)\ j = 1,\ldots,5$

計測ベクトル： $\boldsymbol{y} = [y_1,\ldots,y_5]^T$

解ベクトル： $\boldsymbol{x} = [c_n, c_{n-1},\ldots,c_0]^T$

システム行列

$$\boldsymbol{H} = \begin{bmatrix} x_1^n & x_1^{n-1} & \cdots & 1 \\ x_2^n & x_2^{n-1} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ x_5^n & x_5^{n-1} & \cdots & 1 \end{bmatrix}$$

EMアルゴリズムの妥当性の議論

EMアルゴリズムは，いわば，「Aを計算するためにBが必要であるが，Bを計算するにはAが必要である．」という状況において，

「まず，適当にBを与えて，Aを計算し，その結果を使って，Bを更新し，さらに，……」を繰り返す手法である．

妥当性はどのようにして示せるか？

EMの目的は周辺尤度を最大とするハイパーパラメータを求めることであるので，EMの更新にしたがい周辺尤度が増加することを示せばよい．

妥当性議論の準備ーEMアルゴリズムの汎関数を用いた導出

自由エネルギー

$$F[q(\boldsymbol{x}),\boldsymbol{\theta}] = \int d\boldsymbol{x} q(\boldsymbol{x})[\log p(\boldsymbol{x},\boldsymbol{y} \mid \boldsymbol{\theta}) - \log q(\boldsymbol{x})]$$

$q(\boldsymbol{x})$: 任意の確率分布

$\boldsymbol{\theta}$: ハイパーパラメータ

$F[q,\boldsymbol{\theta}]$を$q(\boldsymbol{x})$と$\boldsymbol{\theta}$で交互に最大化する過程を考える.

$F[q(\boldsymbol{x}),\boldsymbol{\theta}]$を最大とする確率分布$q(\boldsymbol{x})$を求める.

$$q(\boldsymbol{x}) = \arg\max_{q(\boldsymbol{x})} F[q,\boldsymbol{\theta}] \quad \text{subject to} \int_{-\infty}^{\infty} q(\boldsymbol{x}) d\boldsymbol{x} = 1$$

$$q(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \hat{\boldsymbol{\theta}})$$ の解が求まる

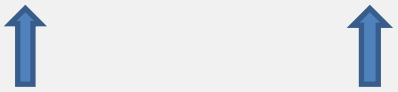
つまり，$\boldsymbol{\theta}$ をこの時点での値 $\hat{\boldsymbol{\theta}}$ に固定した事後分布が求まる

EMアルゴリズムのEステップに対応する．

$F[q(\boldsymbol{x}),\boldsymbol{\theta}]$ を最大とするハイパーパラメータ $\boldsymbol{\theta}$ を求める.

$F[q,\boldsymbol{\theta}]$は既に$q(\boldsymbol{x})$で最大化されているので，以下のように表される.

$$
\begin{aligned}
F[p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}}),\boldsymbol{\theta}] &= \int d\boldsymbol{x}\, p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}})[\log p(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{\theta}) - \log p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}})] \\
&= \int d\boldsymbol{x}\, p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}}) \log p(\boldsymbol{x},\boldsymbol{y}\mid\boldsymbol{\theta}) - \int d\boldsymbol{x}\, p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}}) \log p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}}) \\
&= \mathcal{Q}(\boldsymbol{\theta}) + H[p(\boldsymbol{x}\mid\boldsymbol{y},\hat{\boldsymbol{\theta}})]
\end{aligned}
$$



平均データ尤度　　エントロピー

エントロピーの項は$\boldsymbol{\theta}$を含まないので，自由エネルギーを最大とする$\boldsymbol{\theta}$ を求めることは，平均データ尤度を最大とする$\boldsymbol{\theta}$ を求めることに等しい.

EMアルゴリズムのMステップに対応する.

自由エネルギーを $q(\boldsymbol{x})$ と $\boldsymbol{\theta}$ で交互に最大とする仮定はEMアルゴリズムと等価である.

EMアルゴリズムの妥当性

EMアルゴリズムによるパラメータ $\boldsymbol{\theta}$ の更新の結果，周辺尤度 $\log p(\boldsymbol{y} \mid \boldsymbol{\theta})$ が増加することを示す．

$q(\boldsymbol{x})$ を任意の確率分布として，自由エネルギーは以下のように表せる

$$F[q(\boldsymbol{x}), \boldsymbol{\theta}] = \int d\boldsymbol{x} q(\boldsymbol{x}) [\log p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) + \log p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}) - \log q(\boldsymbol{x})]$$

右辺を計算して，結局，以下を得る．

$$F[q(\boldsymbol{x}), \boldsymbol{\theta}] = \underbrace{\log p(\boldsymbol{y} \mid \boldsymbol{\theta})}_{\text{周辺尤度}} - \underbrace{K[q(\boldsymbol{x}), p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})]}_{\text{KLダイバージェンス}}$$

$K[q(\boldsymbol{x}), p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})] = -\int q(\boldsymbol{x}) \log \left[\frac{p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{x})} \right] d\boldsymbol{x}$ はKLダイバージェンスと呼ばれる

KLダイバージェンス： $K[q(\boldsymbol{x}), p(\boldsymbol{x})] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$

かならず， $K[q(\boldsymbol{x}), p(\boldsymbol{x})] \geq 0$ が成立し，等号成立は $q(\boldsymbol{x})=p(\boldsymbol{x})$ の場合のみ．

2つの確率分布の距離を表すと解釈される．

EMのk回目の更新が終わった時点で，

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$$

$$q(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k-1)})$$

$$F[\boldsymbol{\theta}^{(k)}] = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)}) - K[p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k-1)}), p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k)})]$$

EMのk+1回目のEステップが終わった時点で，

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$$

$$q(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k)})$$

$$F[\boldsymbol{\theta}^{(k)}] = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)})$$

EMのk+1回目のMステップが終わった時点で，

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{(k+1)}$$

$$q(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k)})$$

$$F[\boldsymbol{\theta}^{(k+1)}] = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k+1)}) - K[p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k)}), p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k+1)})]$$

$$\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k+1)}) - \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)}) = F[\boldsymbol{\theta}^{(k+1)}] - F[\boldsymbol{\theta}^{(k)}] + K[p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k)}), p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta}^{(k+1)})] \geq 0$$

したがって，$\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k+1)}) \geq \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)})$である

つまり，EMにより必ず周辺尤度は増加する

ベイズ因子分析：センサーアレイ計測データ

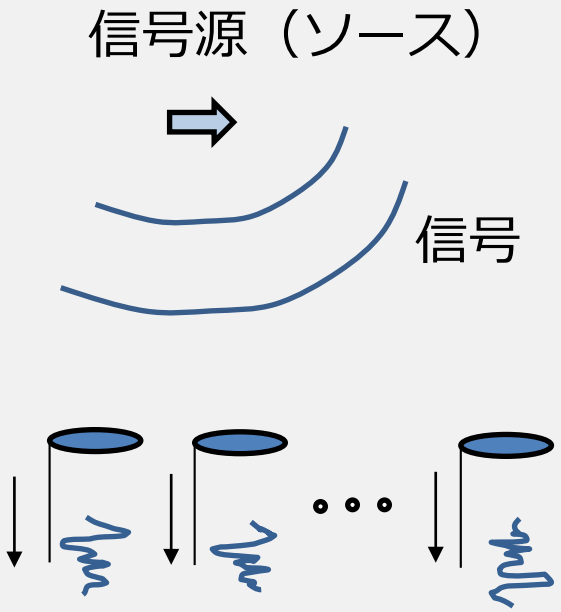
センサー信号のモデル化

センサー出力：

$$y(t) = y_s(t) + \varepsilon(t)$$

全センサーに相関の有る変動信号

各センサーに固有の変動（センサー）ノイズ



したがって，各センサーで固有な変動を分離できれば計測データからノイズを除去し，信号成分を選択的に分離できる．----denoising

ベイズ因子分析

$$\text{観測データ}: \boldsymbol{y}(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_M(t) \end{bmatrix}, \quad \text{因子活動}: \boldsymbol{u}(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_L(t) \end{bmatrix} \quad (M \gg L)$$

$\boldsymbol{y}_k = \boldsymbol{y}(t_k),\ \boldsymbol{u}_k = \boldsymbol{u}(t_k)$ と表記

因子分析モデル： $\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{u}_k + \boldsymbol{\varepsilon}$ $\boldsymbol{A}$: 混合行列

- 多チャンネルの観測データを，各チャンネルで相関のある活動 $\boldsymbol{A}\boldsymbol{u}_k$ と，各チャンネルに固有な活動 $\boldsymbol{\varepsilon}$ に分離する.
- $\boldsymbol{A}\boldsymbol{u}_k$を信号成分，$\boldsymbol{\varepsilon}$ をノイズ成分と解釈することができる.
- 信号成分は，データ数Mよりはるかに少数(L)の因子活動の和として表す.

ベイズ推定により，観測データ $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$ から，因子活動 $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$ と混合行列 $\boldsymbol{A}$ を推定する.

どのように推定するか

確率モデルを以下のように正規分布を用いて定義：

事前確率分布： $p(\boldsymbol{u}_k) = N(\boldsymbol{u}_k \mid 0, \boldsymbol{I})$ ← 因子は独立と事前分布で仮定

データ確率分布： $p(\boldsymbol{y}_k \mid \boldsymbol{u}_k) = N(\boldsymbol{y}_k \mid \boldsymbol{A}\boldsymbol{u}_k, \boldsymbol{\Lambda}^{-1})$

事後確率分布： $p(\boldsymbol{u}_k \mid \boldsymbol{y}_k) = N(\boldsymbol{u}_k \mid \bar{\boldsymbol{u}}_k, \boldsymbol{\Gamma}^{-1})$ ← 因子の相関は事後分布の共分散行列に反映される

事後確率分布のパラメータはベイズ定理より

$$\boldsymbol{\Gamma} = \boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}$$
$$\bar{\boldsymbol{u}}_k = \boldsymbol{\Gamma}^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{y}_k = (\boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{y}_k$$

ともとまる．EMアルゴリズムのEステップ更新式となる．

混合行列 $\boldsymbol{A}$ と，ノイズ精度行列 $\boldsymbol{\Lambda}$ がこの場合のハイパーパラメータである．

混合行列 $\boldsymbol{A}$ と，ノイズ精度行列 $\boldsymbol{\Lambda}$ は平均データ尤度の最大化から更新式を導く

完全データ尤度　（ハイパーパラメータを含んでいない項は無視して，…）

$$\sum_{k=1}^{K}\log p(\boldsymbol{y}_k,\boldsymbol{u}_k\mid\boldsymbol{A},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\log p(\boldsymbol{y}_k\mid\boldsymbol{u}_k,\boldsymbol{A},\boldsymbol{\Lambda})+\sum_{k=1}^{K}\log p(\boldsymbol{u}_k)$$
$$=\frac{K}{2}\log|\boldsymbol{\Lambda}|-\sum_{k=1}^{K}\frac{1}{2}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{u}_k)^T\boldsymbol{\Lambda}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{u}_k)$$

平均データ尤度

$$\Theta(\boldsymbol{A},\boldsymbol{\Lambda})=E[\sum_{k=1}^{K}\log p(\boldsymbol{y}_k,\boldsymbol{u}_k\mid\boldsymbol{A},\boldsymbol{\Lambda})]=\frac{K}{2}\log|\boldsymbol{\Lambda}|-\sum_{k=1}^{K}\frac{1}{2}E[(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{u}_k)^T\boldsymbol{\Lambda}(\boldsymbol{y}_k-\boldsymbol{A}\boldsymbol{u}_k)]$$

ハイパーパラメータのMステップ更新式：

$$\frac{\partial}{\partial\boldsymbol{A}}\Theta(\boldsymbol{A},\boldsymbol{\Lambda})=0\quad\Rightarrow\quad\hat{\boldsymbol{A}}=\boldsymbol{R}_{yu}\boldsymbol{R}_{uu}$$

$$\frac{\partial}{\partial\boldsymbol{\Lambda}}\Theta(\boldsymbol{A},\boldsymbol{\Lambda})=0\quad\Rightarrow\quad\hat{\boldsymbol{\Lambda}}^{-1}=\frac{1}{K}\mathrm{diag}\left[\boldsymbol{R}_{yy}-\boldsymbol{A}\boldsymbol{R}_{uy}\right]$$

ここで，$\boldsymbol{R}_{uu}=\sum_{k=1}^{K}\bar{\boldsymbol{u}}_k\bar{\boldsymbol{u}}_k^T+K\boldsymbol{\Gamma}^{-1}$　$\boldsymbol{R}_{uy}=\sum_{k=1}^{K}\bar{\boldsymbol{u}}_k\boldsymbol{y}_k^T=\boldsymbol{R}_{yu}^T$　$\boldsymbol{R}_{yy}=\sum_{k=1}^{K}\boldsymbol{y}_k\boldsymbol{y}_k^T$　である．

ベイズ因子分析（まとめ）：

データモデル： $\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{u}_k + \boldsymbol{\varepsilon}$

Eステップ更新式

$$\boldsymbol{\Gamma} = \boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}$$

$$\bar{\boldsymbol{u}}_k = \boldsymbol{\Gamma}^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{y}_k = \left(\boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{y}_k$$

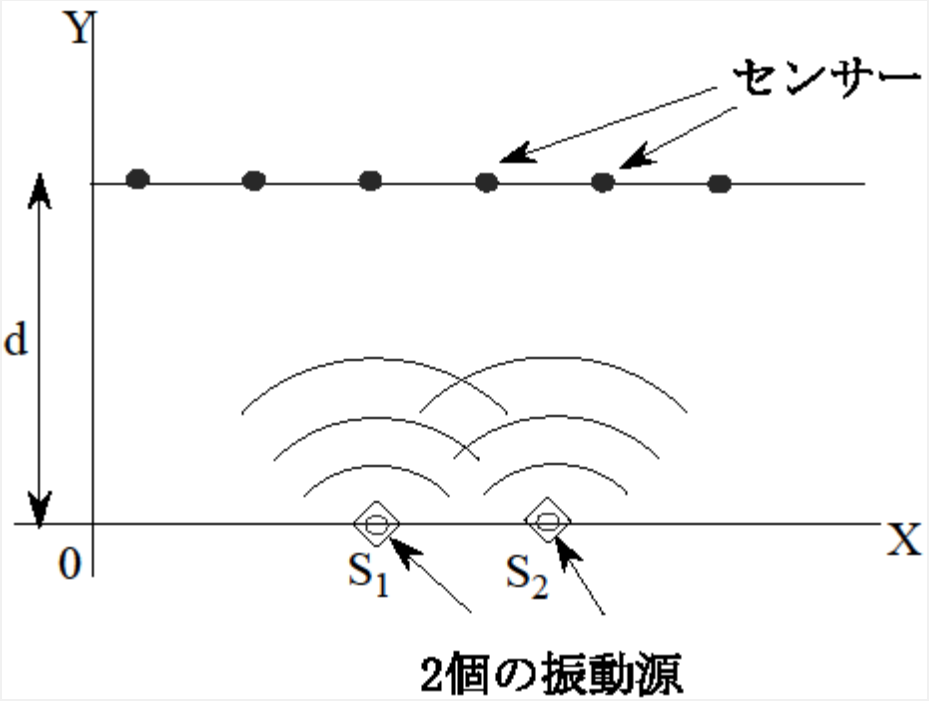
Mステップ更新式

$$\hat{\boldsymbol{A}} = \boldsymbol{R}_{yu} \boldsymbol{R}_{uu}$$

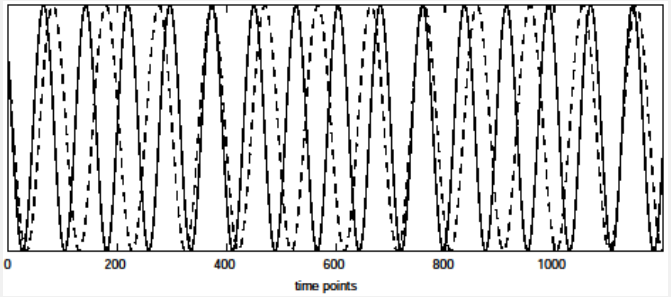
$$\hat{\boldsymbol{\Lambda}}^{-1} = \frac{1}{K} \mathrm{diag}\left[\boldsymbol{R}_{yy} - \boldsymbol{A}\boldsymbol{R}_{uy} \right]$$

EMアルゴリズム終了時点での混合行列 $\hat{\boldsymbol{A}}$ と因子活動 $\bar{\boldsymbol{u}}_k$ を用いて計算した $\hat{\boldsymbol{A}}\bar{\boldsymbol{u}}_k \quad (k = 1, \ldots K)$ が観測データに含まれる信号成分の推定解である.

コンピュータシミュレーション：センサー信号のノイズ除去



2個の振動源



振動源が発する信号

2個の振動源（音響や電磁波など）が発する信号を多数のセンサー（センサーアレイ）で同時計測する．振動は振動源からの距離の二乗で減衰すると仮定した．

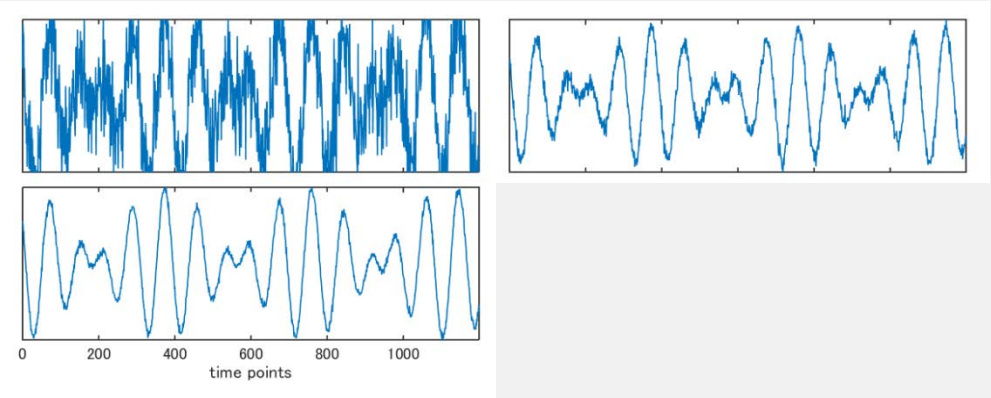
d=300として，2個のソースの位置を(300,0)および(700,0)とした．301個のセンサーによる同時計測を仮定した．

シミュレーションの結果

センサー出力

処理結果：
ファクター数２

処理結果：
ファクター数２０



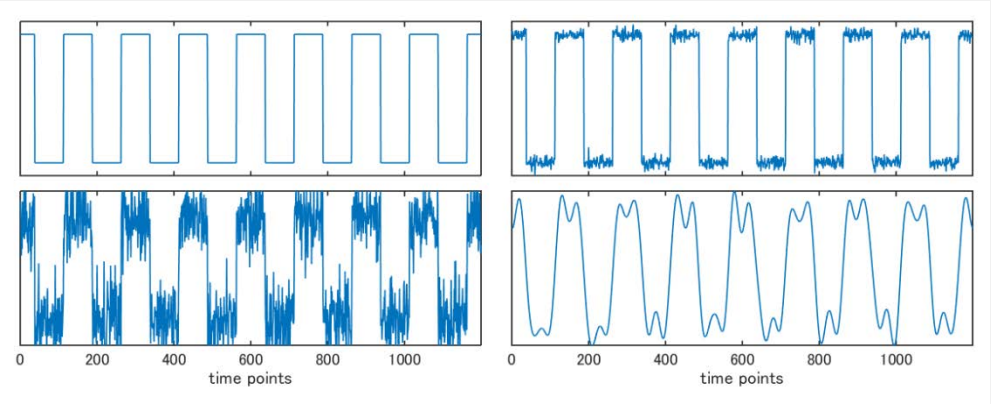
矩形波を用いたシミュレーションの結果

ソースに仮定した
波形

センサー出力

ベイズ因子分析結果

デジタルフィルター
による結果



コンピュータシミュレーションに用いたMatlabコードは以下からダウンロード可能：
http://www.kyoritsu-pub.co.jp/bookdetail/9784320085749

スパースベイズ学習：Sparse Bayesian Learning

線形方程式：$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\varepsilon}$ において，未知量 $\boldsymbol{x}$ をベイズ的に求める．

確率モデルとして以下の正規分布モデルを仮定：

事前分布： $p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \alpha^{-1}\boldsymbol{I})$

尤度： $p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$

⇒

ベイズ推定解：

$$\bar{\boldsymbol{x}} = (\boldsymbol{H}^T\boldsymbol{H} + \frac{\alpha}{\beta}\boldsymbol{I})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

=

L2正則化ミニマムノルム解

事前分布の精度行列： $\alpha\boldsymbol{I} \Rightarrow \mathrm{diag}[\alpha_1, \ldots, \alpha_2]$ とすれば，

つまり，全てのボクセルで共通の α とせずに，個々の ボクセルで異なるα_j とすれば…

↓

ミニマムノルム解とは全く異なったスパースな解が得られる

スパースベイズ学習：Sparse Bayesian Learning

確率モデル：

事前分布： $p(\boldsymbol{x}) = N(\boldsymbol{x} \mid 0, \boldsymbol{\Phi}^{-1})$ ここで $\boldsymbol{\Phi} = \mathrm{diag}(\alpha_1, \ldots, \alpha_N)$ とする.

列ベクトル $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$ と定義し，$\boldsymbol{\Phi} = \mathrm{diag}(\boldsymbol{\alpha})$ と表す.

議論を簡単にするため，尤度： $p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$ において，ノイズ精度 β は既知とする.

事後分布：$p(\boldsymbol{x} \mid \boldsymbol{y})$ も正規分布： $p(\boldsymbol{x} \mid \boldsymbol{y}) = N(\boldsymbol{x} \mid \bar{\boldsymbol{x}}, \boldsymbol{\Gamma}^{-1})$

線形正規モデルでの事後確率分布の計算の議論より，事後分布のパラメータは

$$\boldsymbol{\Gamma} = \boldsymbol{\Phi} + \beta \boldsymbol{H}^T \boldsymbol{H}$$
$$\bar{\boldsymbol{x}} = \beta \boldsymbol{\Gamma}^{-1} \boldsymbol{H}^T \boldsymbol{y} = \beta (\boldsymbol{\Phi} + \beta \boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y}$$

未知量 $\boldsymbol{x}$ を求めるには，ハイパーパラメータ $\boldsymbol{\alpha}$ を求める必要がある.

平均データ尤度を求めEMアルゴリズムを導く

事前分布： $p(\boldsymbol{x} \mid \boldsymbol{\alpha}) = N(\boldsymbol{x} \mid 0, \boldsymbol{\Phi}^{-1}) = \dfrac{\left|\boldsymbol{\Phi}\right|^{1/2}}{(2\pi)^{N/2}} \exp[-\dfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Phi}\boldsymbol{x}]$

尤度： $p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I}) = \dfrac{\left|\beta\right|^{M/2}}{(2\pi)^{N/2}} \exp[-\dfrac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})]$

完全データ尤度： $\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\alpha}) = \log p(\boldsymbol{y} \mid \boldsymbol{x}) + \log p(\boldsymbol{x} \mid \boldsymbol{\alpha})$

$$= \frac{1}{2}\log \left|\boldsymbol{\Phi}\right| - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Phi}\boldsymbol{x} - \frac{\beta}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})$$

平均データ尤度： $\Theta(\boldsymbol{\alpha}) = E[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\alpha})] = \dfrac{1}{2}\log \left|\boldsymbol{\Phi}\right| - \dfrac{1}{2}E[\boldsymbol{x}^T\boldsymbol{\Phi}\boldsymbol{x}]$

更新式： $\dfrac{\partial}{\partial \boldsymbol{\Phi}}\Theta(\boldsymbol{\alpha}) = \dfrac{1}{2}\boldsymbol{\Phi}^{-1} - \dfrac{1}{2}[\overline{\boldsymbol{x}\boldsymbol{x}}^T + \boldsymbol{\Gamma}^{-1}] = 0 \;\Rightarrow\; \boldsymbol{\Phi}^{-1} = \overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} + \boldsymbol{\Gamma}^{-1}$

を得るが，この更新式は $\boldsymbol{H}$ 行列の行数≫列数の場合に収束が非常に遅い.

EMの更新式は使えない!!

もっと収束の早い更新式が導けるか？
とにかく，それを目指して周辺尤度を直接計算する

周辺尤度は以下の式から計算する：

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \int p(\boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\alpha}) d\boldsymbol{x} = \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{\alpha}) d\boldsymbol{x}$$

具体的には：

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I})$$
$$p(\boldsymbol{x} \mid \boldsymbol{\alpha}) = N(\boldsymbol{x} \mid 0, \boldsymbol{\Phi}^{-1})$$

を用いて以下を計算する

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{\alpha}) d\boldsymbol{x} = \int N(\boldsymbol{y} \mid \boldsymbol{H}\boldsymbol{x}, \beta^{-1}\boldsymbol{I}) N(\boldsymbol{x} \mid 0, \boldsymbol{\Phi}^{-1})\ d\boldsymbol{x}$$

上式の積分計算は若干大変である．かなり長い導出の後，以下を得る．

$$\log p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = -\frac{1}{2}\log |\boldsymbol{\Sigma}_y| - \frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}_y^{-1} \boldsymbol{y}$$

ここで $\boldsymbol{\Sigma}_y = \beta^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Phi}^{-1}\boldsymbol{H}^T$ である．

したがって，コスト関数 $F(\boldsymbol{\alpha}) = \log |\boldsymbol{\Sigma}_y| + \boldsymbol{y}^T \boldsymbol{\Sigma}_y^{-1} \boldsymbol{y}$ を最小にする $\boldsymbol{\alpha}$ が $\boldsymbol{\alpha}$ の更新値である．

$F(\boldsymbol{\alpha})$を最小にする $\boldsymbol{\alpha}$ を求める

$$\frac{\partial}{\partial \alpha_j} F(\boldsymbol{\alpha}) = \frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Sigma}_y| + \frac{\partial}{\partial \alpha_j} \boldsymbol{y}^T \boldsymbol{\Sigma}_y^{-1} \boldsymbol{y}$$ を計算する.

$$\frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Sigma}_y| = -\frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Phi}| + \frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Gamma}|$$ であるので，以下を用いて

$$\frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Phi}| = \frac{\partial}{\partial \alpha_j} \sum_{i=1}^{N} \log \alpha_i = \frac{1}{\alpha_j} \qquad \frac{\partial}{\partial \alpha_j} \log |\boldsymbol{\Gamma}| = \Sigma_{j,j} \qquad (\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{-1})$$

$$\frac{\partial}{\partial \alpha_j} \boldsymbol{y}^T \boldsymbol{\Sigma}_y \boldsymbol{y} = \overline{x}_j^2$$

$$\frac{\partial}{\partial \alpha_j} F(\boldsymbol{\alpha}) = \Sigma_{j,j} - \alpha_j^{-1} + \overline{x}_j^2$$ を得る.

ここで，単純に $\alpha_j^{-1} = \Sigma_{j,j} + \overline{x}_j^2$ とすると，これはEMの更新式になってしまう.

$\Sigma_{j,j}$ も $\overline{x}_j^2$ も実は α_j を陰に含むので，どのような組み合わせで更新式を求めるかは任意性が入り込む.

$F(\boldsymbol{\alpha})$を最小にする $\boldsymbol{\alpha}$ を求める：続き

別の組み合わせを模索する.

$\frac{\partial}{\partial \alpha_j} F(\boldsymbol{\alpha}) = \Sigma_{j,j} - \alpha_j^{-1} + \overline{x}_j^2 = 0$ より, $1 - \alpha_j \Sigma_{j,j} = \alpha_j \overline{x}_j^2$ を得る.

一方，事後分布の精度行列更新式より，$\boldsymbol{I} - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi} = \beta \boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{H}$ である.

また, $[\boldsymbol{I} - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}]_{j,j} = 1 - \alpha_j \Sigma_{j,j}$ であるので,

結局，MacKayの更新式

$$\alpha_j = \frac{\left[\beta \boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{H}\right]_{j,j}}{\overline{x}_j^2}$$

を得る．MacKayの更新式は「経験的」に見出されたもの．$\boldsymbol{H}$ 行列の列数が行数に比べて大きな場合，EM更新式に比べて更新が早いことが知られている.

まとめると，以下のEMアルゴリズム類似の更新式により $\boldsymbol{x}$ の推定解 $\bar{\boldsymbol{x}}$ を求める．

Eステップ

$$\boldsymbol{\Gamma} = \boldsymbol{\Phi} + \beta \boldsymbol{H}^T \boldsymbol{H}$$

$$\bar{\boldsymbol{x}} = \beta \boldsymbol{\Gamma}^{-1} \boldsymbol{H}^T \boldsymbol{y} = \beta (\boldsymbol{\Phi} + \beta \boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{y}$$

Mステップ

$$\alpha_j = \frac{\left[\beta \boldsymbol{\Gamma}^{-1} \boldsymbol{H}^T \boldsymbol{H} \right]_{j,j}}{\bar{x}_j^2}$$

このアルゴリズムはスパースな解を与えるためスパースベイズ学習と呼ばれる

MacKayの更新式は，EM更新式よりも高速で収束し，実用的には問題は無いが，更新ごとに周辺尤度が確かに増加することの証明ができないという理論的な問題点がある．

凸関数の性質を用いた更新式の導出
分散を用いた表現

解こうとしている問題：

$\log p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = -\frac{1}{2}\log |\boldsymbol{\Sigma}_y| - \frac{1}{2}(\beta \| \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}} \|^2 + \bar{\boldsymbol{x}}^T \boldsymbol{\Phi} \bar{\boldsymbol{x}})$ の最大値を与える $\boldsymbol{\alpha}$ を見出す．

ここで $\boldsymbol{\Sigma}_y = \beta^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Phi}^{-1}\boldsymbol{H}^T$ である．

$-\frac{1}{2}\log |\boldsymbol{\Sigma}_y|$ が分散に関して凸関数となるため，精度ではなく分散を用いて書き直す．

分散 $\nu_j (= \alpha_j^{-1})$ として $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_N]^T$ を定義，また，共分散行列を $\boldsymbol{\Upsilon} (= \boldsymbol{\Phi}^{-1})$ とする．

解こうとしている問題：

$$\log p(\boldsymbol{y} \mid \boldsymbol{\nu}) = -\frac{1}{2}\log |\boldsymbol{\Sigma}_y| - \frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma}_y \boldsymbol{y} = -\frac{1}{2}\log |\boldsymbol{\Sigma}_y| - \frac{1}{2}(\beta \| \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}} \|^2 + \bar{\boldsymbol{x}}^T \boldsymbol{\Upsilon}^{-1} \bar{\boldsymbol{x}})$$

の最大値を与える $\boldsymbol{\nu}$ を見出す． ここで $\boldsymbol{\Sigma}_y = \beta^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Upsilon}\boldsymbol{H}^T$ である．

凸関数の性質を用いた更新式の導出
代用コスト関数の導出

コスト関数： $\log p(\boldsymbol{y} \mid \boldsymbol{v}) = -\frac{1}{2}\log |\boldsymbol{\Sigma}_y| - \frac{1}{2}(\beta \| \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}} \|^2 + \bar{\boldsymbol{x}}^T \boldsymbol{\Upsilon}^{-1} \bar{\boldsymbol{x}})$

この部分計算がやっかい

$\log |\boldsymbol{\Sigma}_y|$ は凹関数 $\Leftrightarrow$ 補助変数 $\boldsymbol{z}$ を用いて $\boldsymbol{z}^T \boldsymbol{v} - z_0(\boldsymbol{z}) \geq \log |\boldsymbol{\Sigma}_y|$ が常に成立

補助的変数 $\boldsymbol{z}$ を用いて，(代用)コスト関数：

$$\tilde{F}(\boldsymbol{v}, \boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z}^T \boldsymbol{v} - z_0) - \frac{1}{2}(\beta \| \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}} \|^2 + \bar{\boldsymbol{x}}^T \boldsymbol{\Upsilon}^{-1} \bar{\boldsymbol{x}})$$

と定義すれば，$\log p(\boldsymbol{y} \mid \boldsymbol{v}) \geq \tilde{F}(\boldsymbol{v}, \boldsymbol{z})$ が必ず成立する．

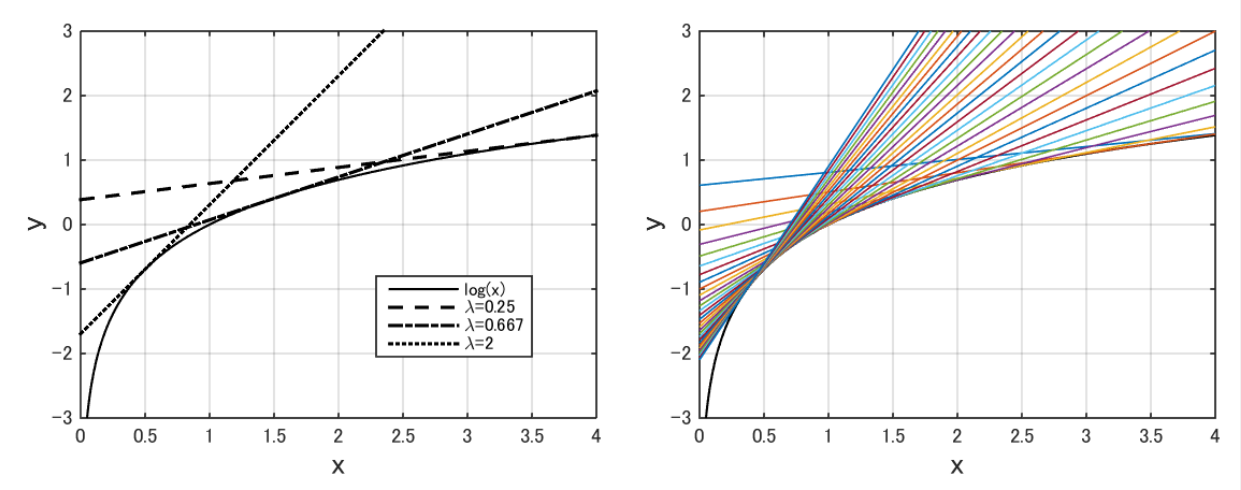
代用コスト関数 $\tilde{F}(\boldsymbol{v}, \boldsymbol{z})$ を $\boldsymbol{v}$ と $\boldsymbol{z}$ を更新しながら増加させればその $\boldsymbol{v}$ は必ず周辺尤度 $\log p(\boldsymbol{y} \mid \boldsymbol{v})$ を増加させる．

計算しづらい真のコスト関数を用いるよりも，変数を１つ増やして，計算しやすいものにした「代用」コスト関数を用いる．

1次元凹関数(Concave function)の例

$$\log\left(\frac{x_1+x_2}{2}\right) \geq \frac{\log(x_1)+\log(x_2)}{2}$$ が成立 $\Rightarrow y=\log(x)$ は凹関数

このとき，$\lambda x-\lambda_0(\lambda) \geq \log(x)$　を満たす直線 $y=\lambda x-\lambda_0(\lambda)$ が必ず存在する．



多次元の場合

$\log|\boldsymbol{\Sigma}_y|$ は凹関数 ⟶ $\boldsymbol{z}^T\boldsymbol{v}-z_0(\boldsymbol{z}) \geq \log|\boldsymbol{\Sigma}_y|$ を満たす平面 $\boldsymbol{z}^T\boldsymbol{v}-z_0(\boldsymbol{z})$ が必ず存在する．

更新式の導出

(代用)コスト関数：

$$\tilde{F}(\boldsymbol{v},\boldsymbol{z}) = -\frac{1}{2}(\boldsymbol{z}^T\boldsymbol{v} - z_0(\boldsymbol{z})) - \frac{1}{2}(\beta \,||\, \boldsymbol{y} - \boldsymbol{H}\overline{\boldsymbol{x}} \,||^2 + \overline{\boldsymbol{x}}^T\boldsymbol{\Upsilon}^{-1}\overline{\boldsymbol{x}})$$

を増加させるパラメータ $\boldsymbol{v}$ と $\boldsymbol{z}$ を求める.

ν_jの更新式

$$\hat{\nu}_j = \arg\max_{\nu_j} \tilde{F}(\boldsymbol{v},\boldsymbol{z}) = \arg\min_{\nu_j}[\boldsymbol{z}^T\boldsymbol{v} + \overline{\boldsymbol{x}}^T\boldsymbol{\Upsilon}^{-1}\overline{\boldsymbol{x}}] = \arg\min_{\nu_j}\sum_{j=1}^{N}[z_j\nu_j + \frac{\overline{x}_j^2}{\nu_j}] = \frac{|\,\overline{x}_j\,|}{\sqrt{z_j}}$$

z_jの更新式

$$\hat{\boldsymbol{z}} = \arg\max_{\boldsymbol{z}} \tilde{F}(\boldsymbol{v},\boldsymbol{z}) = \arg\min_{\boldsymbol{z}}(\boldsymbol{z}^T\boldsymbol{v} - z_0(\boldsymbol{z}))$$**である.**

しかし, $\boldsymbol{z}$ は $\boldsymbol{z}^T\boldsymbol{v} - z_0 \geq \log|\boldsymbol{\Sigma}_y|$ を満たさなければならない.

$\boldsymbol{z}^T\boldsymbol{v} - z_0(\boldsymbol{z}) \geq \log|\boldsymbol{\Sigma}_y|$ を満たしながら，$\boldsymbol{z}^T\boldsymbol{v} - z_0(\boldsymbol{z})$ を最小にする $\boldsymbol{z}$ はどうやって求める？

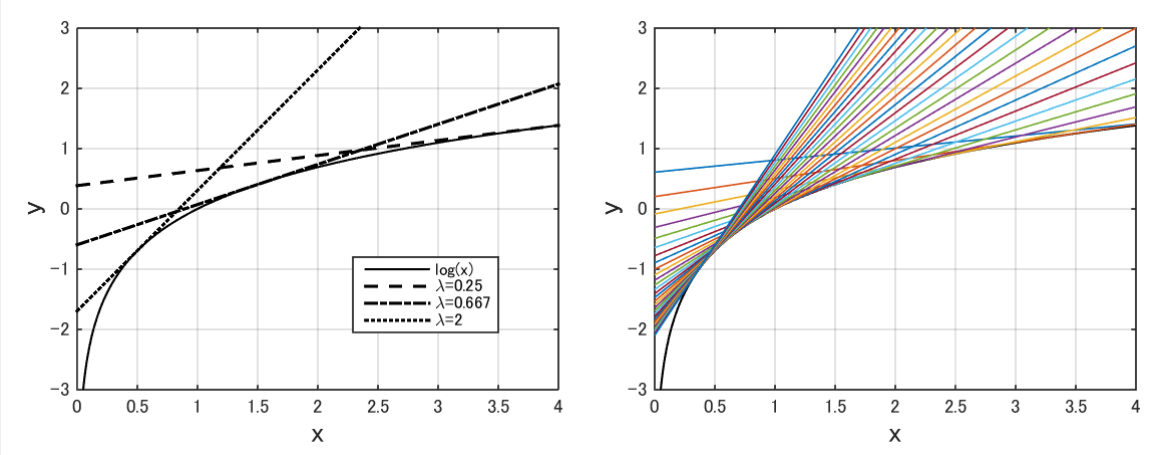
1次元凹関数(Concave function)の例

$\lambda x - \lambda_0(\lambda) \geq \log(x)$ を満たす直線のなかで，$y = \lambda x - \lambda_0(\lambda)$ を最小にする，つまり，

$\log(x)$ に最も近い直線は $y = \log(x)$ の接線である．

したがって，$\lambda x - \lambda_0 \geq \log(x)$ の関係式を満たし，かつ$\lambda x - \lambda_0$を最小とするλ は，各x における$\log(x)$の接線の傾きで与えられる．

つまり，$\hat{\lambda} = \dfrac{\partial}{\partial x} \log(x) = \dfrac{1}{x}$ である．



$\boldsymbol{z}^T \boldsymbol{v} - z_0(\boldsymbol{z}) \geq \log |\boldsymbol{\Sigma}_y|$ を満たしながら，$\boldsymbol{z}^T \boldsymbol{v} - z_0(\boldsymbol{z})$ を最小にする $\boldsymbol{z}$ は

$\hat{z}_j = \dfrac{\partial}{\partial v_j} \log |\Sigma_y|$ として求まる．

更新式の導出

z_j の更新式は以下のように $\log|\Sigma_y|$ の接平面の傾きからもとまる.

$$\hat{z}_j = \frac{\partial}{\partial \nu_j}\log|\Sigma_y| = \mathrm{tr}[\Sigma_y^{-1}\frac{\partial}{\partial \nu_j}\Sigma_y] = \mathrm{tr}[\Sigma_y^{-1}\frac{\partial}{\partial \nu_j}(\beta^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Upsilon}\boldsymbol{H}^T)]$$

$$= \mathrm{tr}[\Sigma_y^{-1}\boldsymbol{H}\frac{\partial}{\partial \nu_j}\boldsymbol{\Upsilon}\boldsymbol{H}^T] = \boldsymbol{h}_j^T\Sigma_y^{-1}\boldsymbol{h}_j$$

まとめると

ν_jの更新式

$$\hat{\nu}_j = \arg\max_{\nu_j}\tilde{F}(\boldsymbol{\nu},\boldsymbol{z}) = \arg\min_{\nu_j}[\boldsymbol{z}^T\boldsymbol{\nu} + \bar{\boldsymbol{x}}^T\boldsymbol{\Upsilon}^{-1}\bar{\boldsymbol{x}}] = \arg\min_{\nu_j}\sum_{j=1}^{N}[z_j\nu_j + \frac{\bar{x}_j^2}{\nu_j}] = \frac{|\bar{x}_j|}{\sqrt{z_j}}$$

z_jの更新式

$$\hat{z}_j = \boldsymbol{h}_j^T\Sigma_y^{-1}\boldsymbol{h}_j = \boldsymbol{h}_j^T(\beta^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{\Upsilon}\boldsymbol{H}^T)^{-1}\boldsymbol{h}_j$$

$\bar{\boldsymbol{x}}$の更新式(Eステップ)

$$\bar{\boldsymbol{x}} = \beta\boldsymbol{\Gamma}^{-1}\boldsymbol{H}^T\boldsymbol{y} = \beta(\boldsymbol{\Upsilon}^{-1} + \beta\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{y}$$

$\boldsymbol{z}, \boldsymbol{\nu}, \bar{\boldsymbol{x}}$ を以上の更新式により更新する.

各更新方式の特徴

- EMアルゴリズムによる更新式：

必ず周辺尤度を増加させる．ただし，推定がill-posedな場合で収束がきわめて遅いことが（経験的に）知られている．

- Mackayの更新式：

劣決定な場合でも収束は，EM更新式に比べ早い．ただ，周辺尤度を必ず増加させる証明はない．（実用的には十分である．）

- 凸関数の性質を用いた更新式：

劣決定な場合でも収束は，EM更新式に比べ早い．また，必ず周辺尤度を増加させることを証明できる．

なぜスパースな解が得られるのかーコスト関数の導出

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}} \left[\log |\boldsymbol{\Sigma}_y| + \beta \| \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}} \|^2 + \bar{\boldsymbol{x}}^T \boldsymbol{\Upsilon}^{-1} \bar{\boldsymbol{x}} \right]$$

$$\bar{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left[\beta \| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \|^2 + \boldsymbol{x}^T \hat{\boldsymbol{\Upsilon}}^{-1} \boldsymbol{x} \right] \quad \text{ここで} \hat{\boldsymbol{\Upsilon}} = [\boldsymbol{\Upsilon}]_{\nu=\hat{\nu}}$$

} を書き直す

$F_T = \beta \| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \|^2 + \log |\boldsymbol{\Sigma}_y| + \boldsymbol{x}^T \boldsymbol{\Upsilon}^{-1} \boldsymbol{x}$ と定義すると以下のように表せる

$$\hat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}} (\min_{\boldsymbol{x}} F_T)$$

$$\bar{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} (\min_{\boldsymbol{v}} F_T)$$

$\bar{\boldsymbol{x}}$ を導出するためのコスト関数は,

$$\min_{\boldsymbol{v}} F_T = \beta \| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \|^2 + \min_{\boldsymbol{v}} \left[\log |\boldsymbol{\Sigma}_y| + \boldsymbol{x}^T \boldsymbol{\Upsilon}^{-1} \boldsymbol{x} \right]$$

であるので，結局，以下のように書ける.

$$\bar{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} F : \; F = \beta \| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \|^2 + \phi(\boldsymbol{x})$$

ここで，制約条件 $\phi(\boldsymbol{x})$ は

$$\phi(\boldsymbol{x}) = \min_{\boldsymbol{v}} \left[\log |\boldsymbol{\Sigma}_y| + \boldsymbol{x}^T \boldsymbol{\Upsilon}^{-1} \boldsymbol{x} \right] = \min_{\boldsymbol{v}} \left[\sum_{j=1}^{N} \frac{x_j^2}{\nu_j} + \log |\boldsymbol{\Sigma}_y| \right]$$

である.

制約項$\phi(\boldsymbol{x})$の計算

話を少し簡単にするため，行列$\boldsymbol{H}$の列ベクトルに対して，$\boldsymbol{h}_i^T \boldsymbol{h}_j = I_{i,j}$を仮定する．

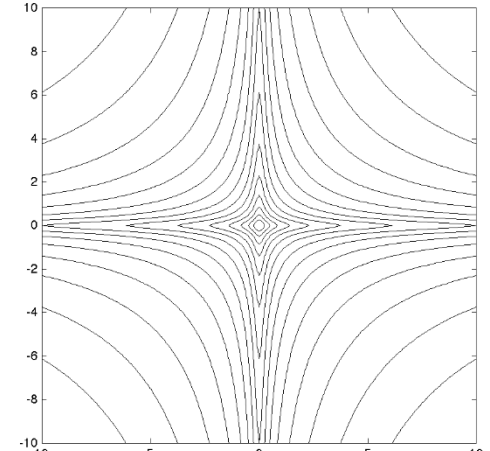
すると，以下を得る．

$$\phi(\boldsymbol{x}) = \sum_{j=1}^{N} \varphi(x_j)$$

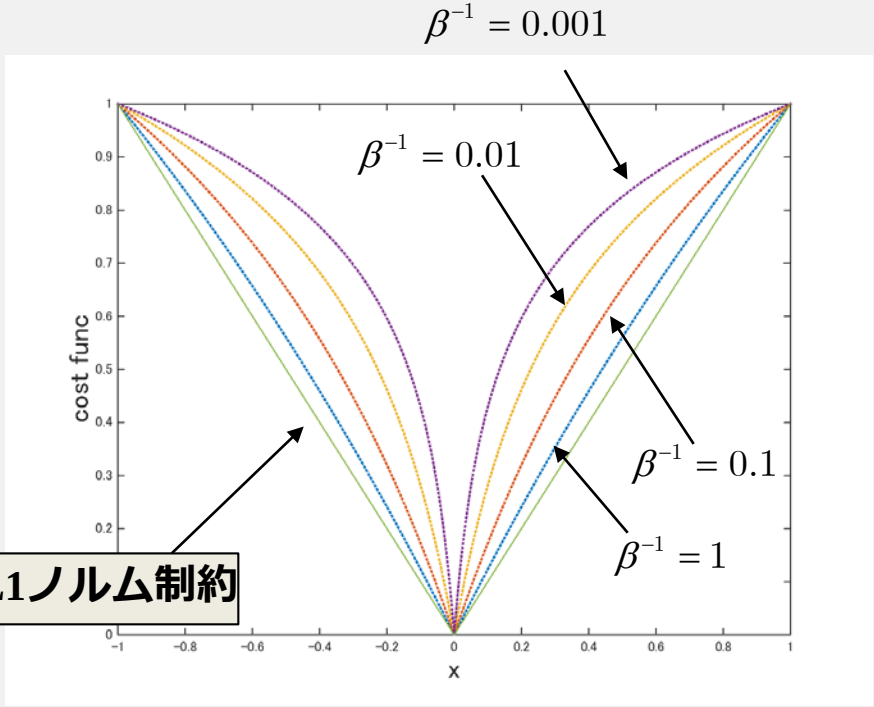
$$\varphi(x_j) = \min_{\nu_j} \left[\frac{x_j^2}{\nu_j} + \log(\beta^{-1} + \nu_j) \right]$$

制約項の１Dプロット．

ノイズ分散が小さくなるに従い制約項はより急峻になる．反対に，ノイズ分散が大きくなると，制約項はL1ノルム制約に近づく．

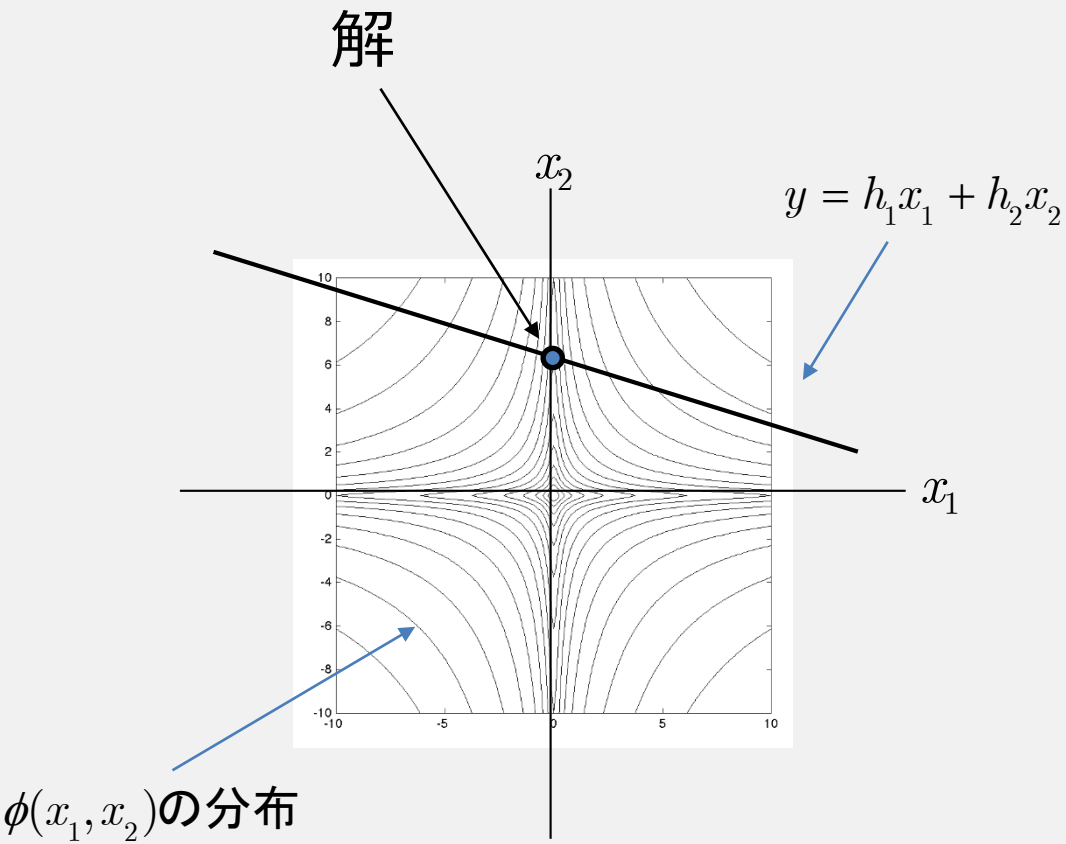


$\beta^{-1} = 0.001$の場合の2Dプロット

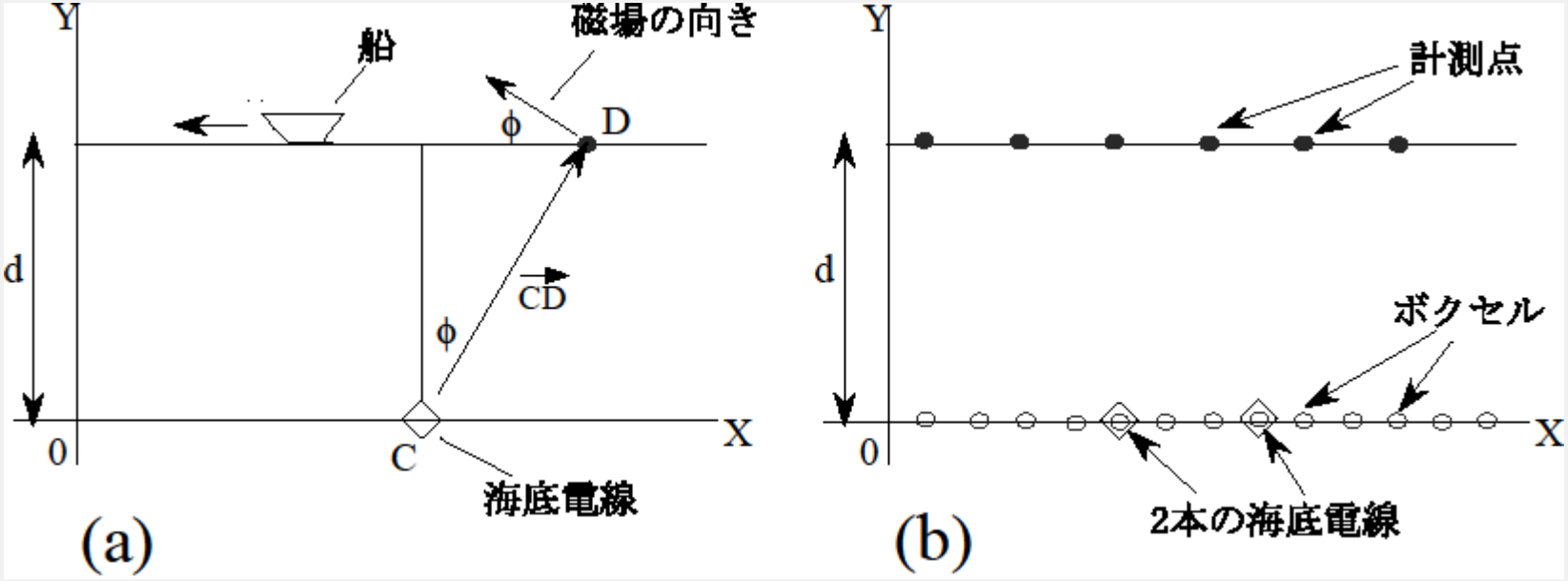


2次元問題で考えてみる：スパースベイズの解

最適化：$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \phi(x_1, x_2)$ subject to $y = h_1 x_1 + h_2 x_2$



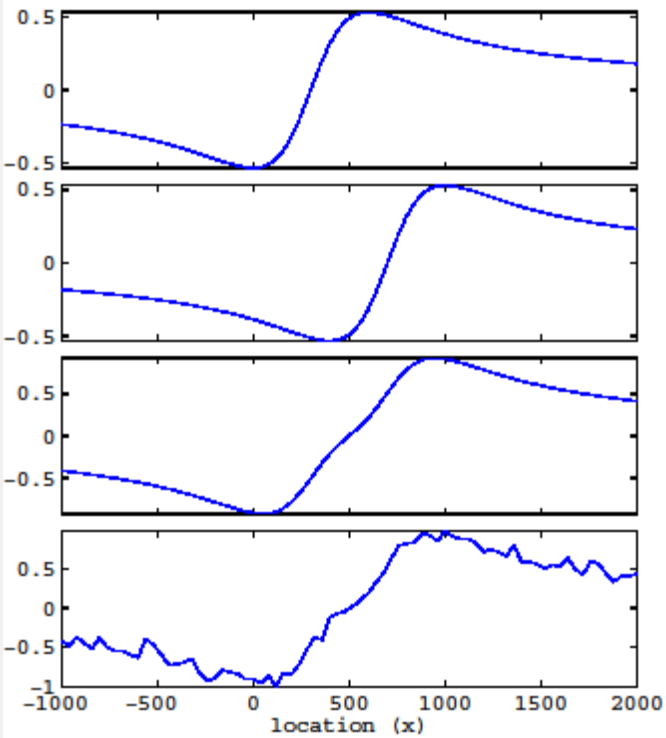
コンピュータシミュレーション：海底電線の位置推定



Cにある電線(電流I)が海面のDの位置に作る磁場の法線成分$B_{\perp} = \frac{I}{r}\sin\phi$： $r = \mathrm{CD}$

コンピュータシミュレーションに用いたMatlabコードは以下からダウンロード可能：
http://www.kyoritsu-pub.co.jp/bookdetail/9784320085749

水面上の$B_{\perp}$の分布



X=300に1本の電線が有る場合

X=700に1本の電線が有る場合

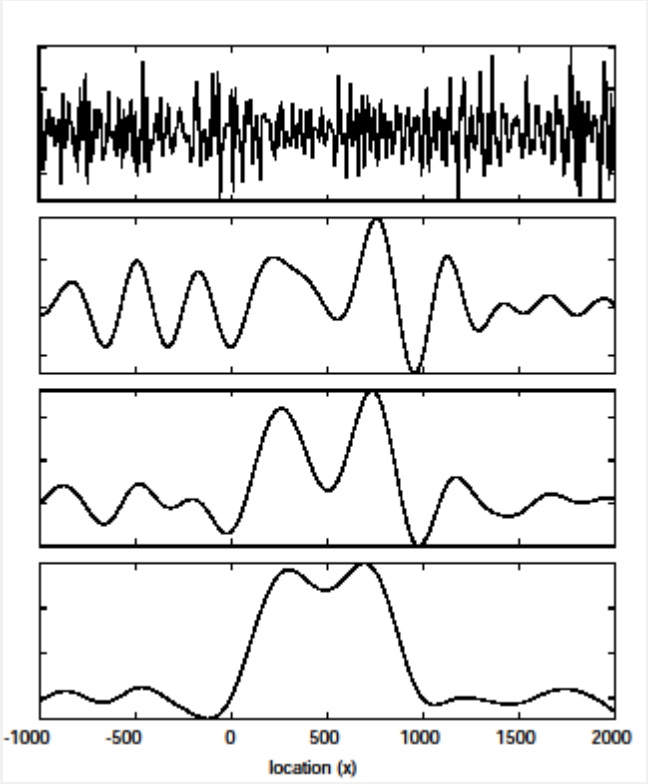
X=300とX=700に2本の電線が有る場合

X=300とX=700に2本の電線が有る場合で，更に正規ノイズを信号対雑音比10で加えた.

海底の電流分布再構成実験

$\boldsymbol{H}$の(i,j)要素$H_{i,j}$：　$H_{i,j} = \dfrac{\sin\phi_{i,j}}{r_{i,j}}$

$\hat{\boldsymbol{x}} = \boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{H}^T + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{y}$

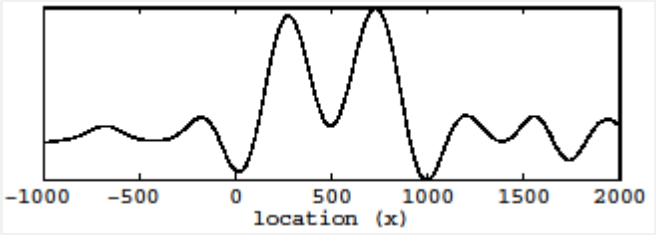


ミニマムノルム解

L2正則化ミニマムノルム解：$\gamma=10^{-4}$

L2正則化ミニマムノルム解：$\gamma=10^{-3}$

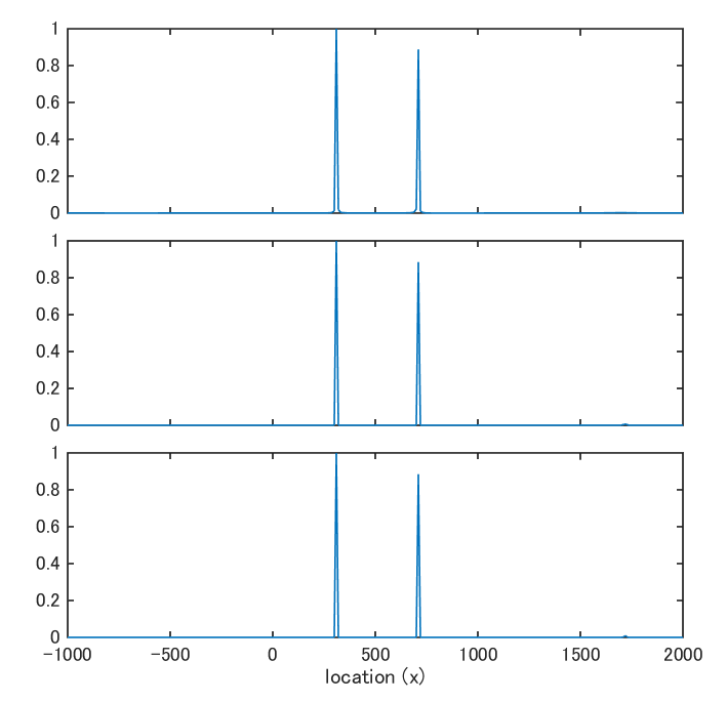
L2正則化ミニマムノルム解：$\gamma=10^{-2}$



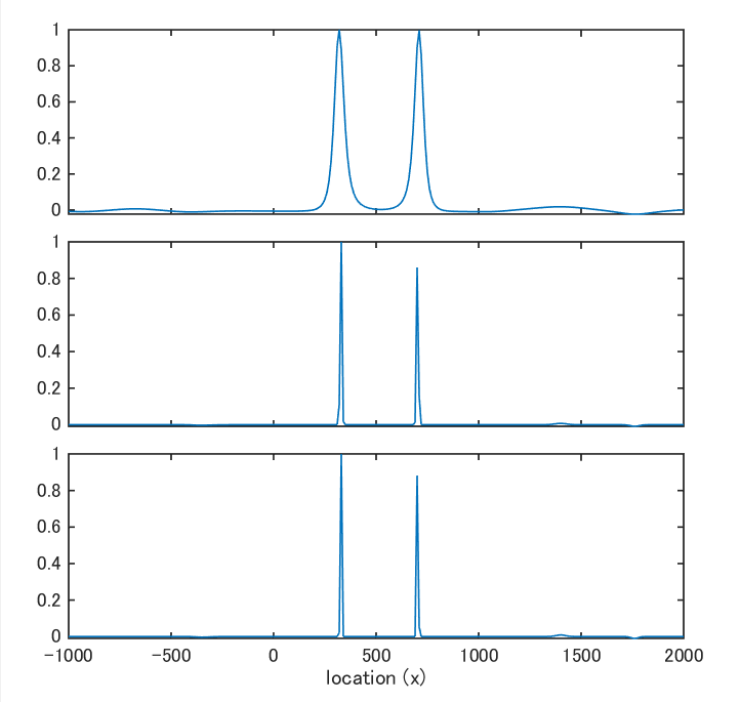
ベイズミニマムノルム解
（正則化パラメータをEMにより導出）

スパース・ベイズ法

反復計算２００回



反復計算10回



上段：EM更新式
中断：MacKay更新式
下段：凸関数の性質を用いた更新式

変分ベイズ法

未知量 $\boldsymbol{x}$ と，ハイパーパラメータ $\boldsymbol{\theta}$ の両方を求めたい.

ベイズ原理主義的な考え方：

両方の未知量に対する結合事後分布：

$$p(\boldsymbol{x},\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

を求めたいが，求められない.

次善の策：

$\hat{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{y} \mid \boldsymbol{\theta})$ を計算する

周辺尤度を最大とするハイーパーパラメータを求める.

なんとか，結合事後分布を近似的にでも求められないか？

もう1度…　自由エネルギーの最大化について．

$$F[q(\boldsymbol{x})] = \int d\boldsymbol{x} q(\boldsymbol{x})[\log p(\boldsymbol{x}, \boldsymbol{y}) - \log q(\boldsymbol{x})]$$ （ハイパーパラメータの表記は省略して）

任意の確率分布 $q(\boldsymbol{x})$ で定義される．汎関数 $F[q(\boldsymbol{x})]$ を最大とする $q(\boldsymbol{x})$を考える．

$$q(\boldsymbol{x}) = \arg\max_{q(\boldsymbol{x})} F[q(\boldsymbol{x})] \text{ subject to } \int_{-\infty}^{\infty} q(\boldsymbol{x}) d\boldsymbol{x} = 1$$

$q(\boldsymbol{x})$=$p(\boldsymbol{x} \mid \boldsymbol{y})$の解が求まる

つまり，自由エネルギーと呼ばれる汎関数を最大とする確率分布として事後分布が求まる．

ベイズ定理を用いずに事後分布を求めることができる．

自由エネルギーの最大化が事後分布を求める手段として使える!!

自由エネルギーについてさらに補足すると

$q(\boldsymbol{x})$ を任意の確率分布として，自由エネルギーは以下のように表せる

$$F[q(\boldsymbol{x})] = \log p(\boldsymbol{y}) - K[q(\boldsymbol{x}) \,||\, p(\boldsymbol{x} \mid \boldsymbol{y})]$$

KLダイバージェンス： $K[q(\boldsymbol{x}), p(\boldsymbol{x})] = \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$

かならず， $K[q(\boldsymbol{x}), p(\boldsymbol{x})] \geq 0$ が成立し，等号成立は $q(\boldsymbol{x})=p(\boldsymbol{x})$ の場合のみ．

2つの確率分布の距離を表すと解釈される．

したがって， $\underset{q(\boldsymbol{x})}{\operatorname{argmax}} F[q(\boldsymbol{x})]$ の解は， $\underset{q(\boldsymbol{x})}{\arg\min} K[q(\boldsymbol{x}) \,||\, p(\boldsymbol{x} \mid \boldsymbol{y})]$ の解，

すなわち，事後分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$を最もよく近似する $q(\boldsymbol{x})$ が得られる．

自由エネルギーを用いた「近似的」結合事後分布の導出

自由エネルギーを結合事後分布に対して定義する

$$F[q(\boldsymbol{x},\boldsymbol{\theta})] = \int d\boldsymbol{x}d\boldsymbol{\theta}q(\boldsymbol{x},\boldsymbol{\theta})[\log p(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y}) - \log q(\boldsymbol{x},\boldsymbol{\theta})]$$

この $F[q(\boldsymbol{x},\boldsymbol{\theta})]$ を最大とする $q(\boldsymbol{x},\boldsymbol{\theta})$ が求める結合事後分布である.

この汎関数の最大化を計算するため，未知量 $\boldsymbol{x}$ とハイパーパラメータ $\boldsymbol{\theta}$ の事後分布が独立であると仮定する

$$p(\boldsymbol{x},\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

これを変分近似と呼ぶ.

$p(\boldsymbol{x} \mid \boldsymbol{y}),\ p(\boldsymbol{\theta} \mid \boldsymbol{y})$ を求めるために，$q(\boldsymbol{x},\boldsymbol{\theta})$ についても $q(\boldsymbol{x},\boldsymbol{\theta}) = q(\boldsymbol{x})q(\boldsymbol{\theta})$ とすれば

自由エネルギーは以下のように表される.

$$F[q(\boldsymbol{x}),q(\boldsymbol{\theta})] = \int d\boldsymbol{x}d\boldsymbol{\theta}q(\boldsymbol{x})q(\boldsymbol{\theta})[\log p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \log q(\boldsymbol{x}) - \log q(\boldsymbol{\theta})]$$

自由エネルギーを用いた「近似的」結合事後分布の導出

自由エネルギー：

$$F[q(\boldsymbol{x}),q(\boldsymbol{\theta})]=\int d\boldsymbol{x}d\boldsymbol{\theta}q(\boldsymbol{x})q(\boldsymbol{\theta})[\log p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})-\log q(\boldsymbol{x})-\log q(\boldsymbol{\theta})]$$

この $F[q(\boldsymbol{x}),q(\boldsymbol{\theta})]$ を同時に最大化する $q(\boldsymbol{x})$ と $q(\boldsymbol{\theta})$ を $\hat{p}(\boldsymbol{x}\mid\boldsymbol{y})$, $\hat{p}(\boldsymbol{\theta}\mid\boldsymbol{y})$ と書けば，これらは

$$\hat{p}(\boldsymbol{x}\mid\boldsymbol{y}),\hat{p}(\boldsymbol{\theta}\mid\boldsymbol{y})=\underset{q(\boldsymbol{x}),q(\boldsymbol{\theta})}{\arg\min}\,K[p(\boldsymbol{x},\boldsymbol{\theta}\mid\boldsymbol{y})\,||\,q(\boldsymbol{x})q(\boldsymbol{\theta})]$$

を満たす.

すなわち，積 $\hat{p}(\boldsymbol{x}\mid\boldsymbol{y})\hat{p}(\boldsymbol{\theta}\mid\boldsymbol{y})$ は結合事後分布 $p(\boldsymbol{x},\boldsymbol{\theta}\mid\boldsymbol{y})$ を最もよく近似する（積で表された）確率分布である.

最大化の計算ーVBEMアルゴリズム

$$\hat{p}(\boldsymbol{x} \mid \boldsymbol{y}),\ \hat{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \underset{q(\boldsymbol{x}),q(\boldsymbol{\theta})}{\arg\max}\, F[q(\boldsymbol{x}), q(\boldsymbol{\theta})] \ \text{ subject to } \int q(\boldsymbol{x}) d\boldsymbol{x} = 1 \ \text{ and } \ \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$$

の解を求める．少々長い導出の後，以下を得る

$$\log \hat{p}(\boldsymbol{x} \mid \boldsymbol{y}) = \int d\boldsymbol{\theta} \hat{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})] \qquad (1)$$

$$\log \hat{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \int d\boldsymbol{x} \hat{p}(\boldsymbol{x} \mid \boldsymbol{y}) \log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = E_{\boldsymbol{x}}[\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})] \qquad (2)$$

すなわち，

$\hat{p}(\boldsymbol{x} \mid \boldsymbol{y})$ を求めるには $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$についての平均データ尤度を計算する．

$\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ を求めるには $\hat{p}(\boldsymbol{x} \mid \boldsymbol{y})$についての平均データ尤度を計算する．

式（１）（２）を交互に繰り返しながら再帰的に事後分布を求める．
VBEM(variational Bayes EM) アルゴリズムと呼ばれる．

変分ベイズ因子分析

モデル：$\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{u}_k + \boldsymbol{\varepsilon}$

から観測データを各チャンネルで共通な活動と，チャンネルに固有な活動 に分離する．

このとき，混合行列 $\boldsymbol{A}$ の事後分布も求める． ⇒ モデルオーダー（因子数）をアルゴリズムが決めることができる．

混合行列： $\boldsymbol{A} = \begin{bmatrix} A_{1,1} & \cdots & A_{1,L} \\ A_{2,1} & \cdots & A_{2,L} \\ \vdots & & \vdots \\ A_{M,1} & \cdots & A_{M,L} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_M^T \end{bmatrix}$

として，事前分布： $p(\boldsymbol{a}_j) = N(\boldsymbol{a}_j \mid 0, (\lambda_j \boldsymbol{\alpha})^{-1})$ を仮定． （精度 $\boldsymbol{\alpha}$ は対角行列）

事後分布は，やはり正規分布： $p(\boldsymbol{a}_j \mid \boldsymbol{y}) = N(\boldsymbol{a}_j \mid \bar{\boldsymbol{a}}_j, (\lambda_j \boldsymbol{\psi})^{-1})$ で与えられる．

変分ベイズ因子分析では，因子数 L は真の因子数 L_0 より大きめに設定する．

$\bar{\boldsymbol{a}}_j = [\bar{A}_{j,1}, \ldots, \bar{A}_{j,L_0}, \bar{A}_{j,L_0+1}, \ldots, \bar{A}_{j,L}]$ 存在しない因子に対する因子行列の要素．

非常に小さな値に推定される

変分ベイズ因子分析

データモデル: $\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{u}_k + \boldsymbol{\varepsilon}$

$\boldsymbol{A}\boldsymbol{u}_k$ を信号成分，$\boldsymbol{\varepsilon}$ をノイズ成分と解釈し，信号成分を分離する．

ベイズ因子分析と異なり，因子数Lの決め方に（結果が）あまり依存しない．

事前分布：

$$p(\boldsymbol{u}) = \prod_{k=1}^{K} N(\boldsymbol{u}_k \mid 0, \boldsymbol{I})$$

$$p(\boldsymbol{A}) = \prod_{j=1}^{M} N(\boldsymbol{a}_j \mid 0, (\lambda_j \boldsymbol{\alpha})^{-1})$$ ⟸ ハイパーパラメータ $\boldsymbol{A}$ に対して事前分布を仮定

データ尤度：

$$p(\boldsymbol{y} \mid \boldsymbol{A}, \boldsymbol{u}) = \prod_{k=1}^{K} N(\boldsymbol{y}_k \mid \boldsymbol{A}\boldsymbol{u}_k, \boldsymbol{\Lambda}^{-1})$$

事後分布：

$$p(\boldsymbol{u} \mid \boldsymbol{y}) = \prod_{k=1}^{K} N(\boldsymbol{u}_k \mid \bar{\boldsymbol{u}}_k, \boldsymbol{\Gamma}^{-1})$$

$$p(\boldsymbol{A} \mid \boldsymbol{y}) = \prod_{j=1}^{M} N(\boldsymbol{a}_j \mid \bar{\boldsymbol{a}}_j, (\lambda_j \boldsymbol{\psi})^{-1})$$ ⟸ ハイパーパラメータ $\boldsymbol{A}$ に対する事後分布をこのように置く

事後分布に対する変分近似： $p(\boldsymbol{u}, \boldsymbol{A} \mid \boldsymbol{y}) = p(\boldsymbol{u} \mid \boldsymbol{y}) p(\boldsymbol{A} \mid \boldsymbol{y})$ を仮定．

VBEMアルゴリズム

完全データ尤度： $\log p(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{A}) = \log p(\boldsymbol{y} \mid \boldsymbol{A}, \boldsymbol{u}) + \log p(\boldsymbol{u}) + \log p(\boldsymbol{A})$

Eステップ

$\log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y}) = E_{\boldsymbol{A}}[\log p(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{A})]$ から事後分布 $\hat{p}(\boldsymbol{u} \mid \boldsymbol{y})$ を推定

$$\log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y}) = E_{\boldsymbol{A}}[\log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{A}) + \log p(\boldsymbol{u})]$$

$$= \sum_{k=1}^{K} E_{\boldsymbol{A}}[\log p(\boldsymbol{y}_k \mid \boldsymbol{u}_k, \boldsymbol{A}) + \log p(\boldsymbol{u}_k)]$$

であるので，事後分布を $p(\boldsymbol{u}_k \mid \boldsymbol{y}_k) = N(\boldsymbol{u}_k \mid \bar{\boldsymbol{u}}_k, \boldsymbol{\Gamma}^{-1})$ とおけば，若干の計算の後

以下のEステップ更新式を得る：

$$\boldsymbol{\Gamma} = \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \bar{\boldsymbol{A}} + M \boldsymbol{\psi}^{-1} + \boldsymbol{I}$$

$$\bar{\boldsymbol{u}}_k = \boldsymbol{\Gamma}^{-1} \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \boldsymbol{y}_k$$

VBEMアルゴリズム

Mステップ

$\log \hat{p}(\boldsymbol{A} \mid \boldsymbol{y}) = E_{\boldsymbol{u}}[\log p(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{A})]$ から，事後分布 $\hat{p}(\boldsymbol{A} \mid \boldsymbol{y})$ を推定

$$\begin{aligned}\log \hat{p}(\boldsymbol{A} \mid \boldsymbol{y}) &= E_{\boldsymbol{u}}[\log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{A})] + \log p(\boldsymbol{A}) \\ &= \sum_{k=1}^{K} E_{\boldsymbol{u}}[-\frac{1}{2}(\boldsymbol{y}_k - \boldsymbol{A}\boldsymbol{u}_k)^T \boldsymbol{\Lambda}(\boldsymbol{y}_k - \boldsymbol{A}\boldsymbol{u}_k)] + \log p(\boldsymbol{A})\end{aligned}$$

を用いて，事後分布は，$p(\boldsymbol{A} \mid \boldsymbol{y}) = \prod_{j=1}^{M} N(\boldsymbol{a}_j \mid \bar{\boldsymbol{a}}_j, (\lambda_j \boldsymbol{\psi})^{-1})$ と置けば，

若干の計算の後，以下のMステップ更新式を得る

$$\boldsymbol{\psi} = \boldsymbol{R}_{uu} + \boldsymbol{\alpha}$$
$$\bar{\boldsymbol{A}} = \boldsymbol{R}_{yu} \boldsymbol{\psi}^{-1}$$

ただし，

$$\boldsymbol{R}_{uu} = \sum_{k=1}^{K} \bar{\boldsymbol{u}}_k \bar{\boldsymbol{u}}_k^T + K\boldsymbol{\Gamma}^{-1}$$
$$\boldsymbol{R}_{uy} = \sum_{k=1}^{K} \bar{\boldsymbol{u}}_k \boldsymbol{y}_k^T = \boldsymbol{R}_{yu}^T$$

を得る．

VBEMの更新式にはパラメータ $\boldsymbol{\alpha}, \boldsymbol{\Lambda}$ が含まれてしまうので
これらも決めてやらなければならない．

$\boldsymbol{\alpha}, \boldsymbol{\Lambda}$ は自由エネルギーを最大とするこれらのパラメータの値として求める.

自由エネルギー

$$
\begin{aligned}
F(\boldsymbol{\alpha}, \boldsymbol{\Lambda}) &= E_{\boldsymbol{A}}[E_{\boldsymbol{u}}[\log p(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{A}) - \log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y}) - \log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y})]] \\
&= E_{\boldsymbol{A}}[E_{\boldsymbol{u}}[\log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{A}) + \log p(\boldsymbol{u}) + \log p(\boldsymbol{A}) \\
&\qquad - \log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y}) - \log \hat{p}(\boldsymbol{u} \mid \boldsymbol{y})]]
\end{aligned}
$$

$\boldsymbol{\alpha}$ の推定

自由エネルギー： $F(\boldsymbol{\alpha}, \boldsymbol{\Lambda}) = E_{\boldsymbol{A}}[\log p(\boldsymbol{A})]$ と表されるので，若干の計算の後，

更新式： $\boldsymbol{\alpha} = \mathrm{diag}[\frac{1}{M} \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \bar{\boldsymbol{A}} + \boldsymbol{\psi}^{-1}]$ を得る.

$\boldsymbol{\Lambda}$ の推定

自由エネルギー：$F(\boldsymbol{\alpha}, \boldsymbol{\Lambda}) = E_{\boldsymbol{A}}[\log p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{A}) + \log p(\boldsymbol{A})]$ であるので，若干の計算の後，

更新式： $\boldsymbol{\Lambda}^{-1} = \frac{1}{K} \mathrm{diag}[\boldsymbol{R}_{yy} - \bar{\boldsymbol{A}} \boldsymbol{R}_{uy}]$ を得る.

ベイズ因子分析と変分ベイズ因子分析：比較

Eステップ更新式：

ベイズ因子分析

$$\boldsymbol{\Gamma} = \boldsymbol{I} + \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{A}$$

$$\bar{\boldsymbol{u}}_k = \boldsymbol{\Gamma}^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda} \boldsymbol{y}_k$$

変分ベイズ因子分析

$$\boldsymbol{\Gamma} = \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \bar{\boldsymbol{A}} + M\boldsymbol{\psi}^{-1} + \boldsymbol{I}$$

$$\bar{\boldsymbol{u}}_k = \boldsymbol{\Gamma}^{-1} \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \boldsymbol{y}_k$$

Mステップ更新式：

ベイズ因子分析

$$\boldsymbol{A} = \boldsymbol{R}_{yu} \boldsymbol{R}_{uu}^{-1}$$

$$\boldsymbol{\Lambda}^{-1} = \frac{1}{K} \mathrm{diag}\left[\boldsymbol{R}_{yy} - \boldsymbol{A}\boldsymbol{R}_{uy} \right]$$

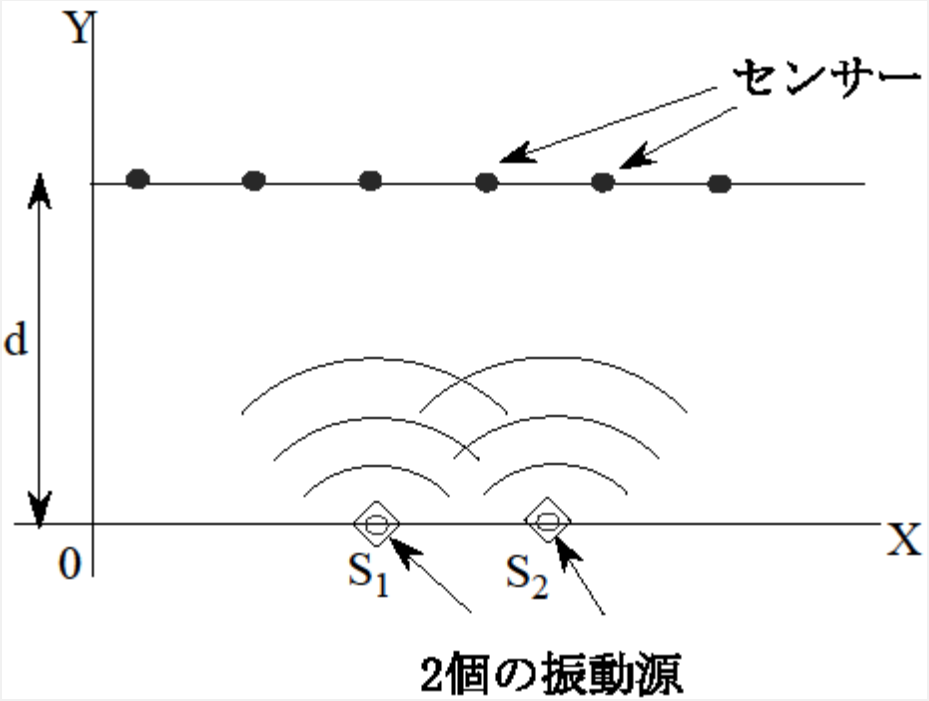
変分ベイズ因子分析

$$\bar{\boldsymbol{A}} = \boldsymbol{R}_{yu} \boldsymbol{\Psi}^{-1} = \boldsymbol{R}_{yu} \left(\boldsymbol{R}_{uu} + \boldsymbol{\alpha} \right)^{-1}$$

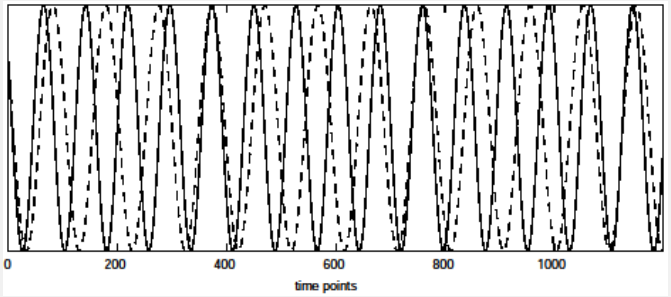
$$\boldsymbol{\Lambda}^{-1} = \frac{1}{K} \mathrm{diag}[\boldsymbol{R}_{yy} - \bar{\boldsymbol{A}}\boldsymbol{R}_{uy}]$$

$$\boldsymbol{\alpha} = \mathrm{diag}[\frac{1}{M} \bar{\boldsymbol{A}}^T \boldsymbol{\Lambda} \bar{\boldsymbol{A}} + \boldsymbol{\Psi}^{-1}]$$

コンピュータシミュレーション：センサー信号のノイズ除去



2個の振動源



振動源が発する信号

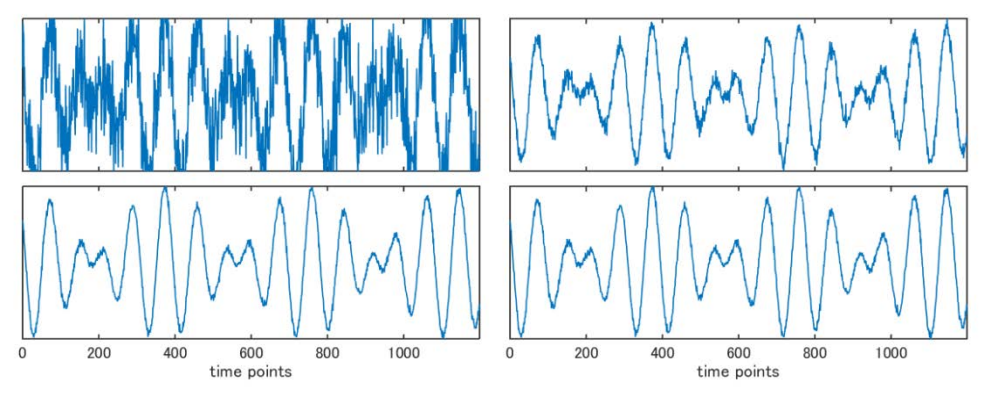
2個の振動源（音響や電磁波など）が発する信号を多数のセンサー（センサーアレイ）で同時計測する．振動は振動源からの距離の二乗で減衰すると仮定した．

d=300として，2個のソースの位置を(300,0)および(700,0)とした．301個のセンサーによる同時計測を仮定した．

シミュレーションの結果

センサー出力
（処理前）

処理結果：
ファクター数２



処理結果：
ファクター数20

変分ベイズ法
ファクター数20