Pattern
Recognition
and
Machine
Learning

第3章　線形回帰モデル

修士1年
山田　孝太郎

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

内容

1. 線形基底関数モデル
2. バイアス-バリアンス分解
3. ベイズ線形回帰
4. ベイズモデル比較
5. エビデンス近似

はじめに

- 回帰とは？
  - D次元の入力ベクトル（観測値）とそれに対応する訓練データ集合から，新しい観測値に対応する目標値を予測するもの
- 線形回帰モデル
  - 基底関数の線形結合を回帰式とするもの

1.線形基底関数モデル

- 一般形：基底関数の線形結合

$$y(\mathbf{x},\mathbf{w}) = \omega_0 + \sum_{j=1}^{M-1} \omega_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}).$$

$$\mathbf{x} = (x_1, \ldots, x_D)$$

$$\mathbf{w} = (\omega_1, \ldots, \omega_{M-1})$$

$\phi_j(\mathbf{x})$：基底関数

$$\phi(\mathbf{x}) = (\phi_1, \ldots, \phi_{M-1})$$

- 基底関数の例

$$\phi_j(x) = x$$

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$:ガウス基底関数

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right) \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$:シグモイド基底関数

1.1 最尤推定と最少二乗法

- tを関数とガウスノイズの和であらわすと

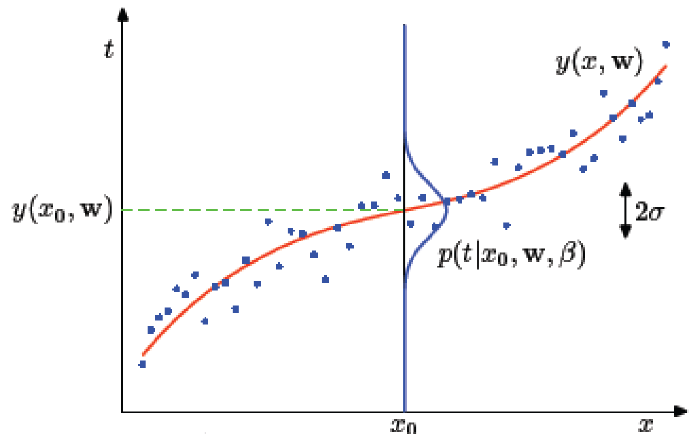
$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon$$

- つまり，tは次の分布に従う

$$p(t \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- 入力と目標値が与えられたときの尤度関数

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(\mathbf{x}, \mathbf{w}), \beta^{-1})$$



Pattern Recognition and Machine Learning

1.1 最尤推定と最少二乗法

- 尤度関数の対数をとって最小化する

$$\nabla \ln p(\mathbf{t}|\mathbf{w},\beta)=\beta\sum_{n=1}^{N}\{t-\mathbf{w}^T\phi(x_n)\}\phi(x_n)^T$$

- =0とおいてwについてとくと,

$$\mathbf{w}_{ML}=\boxed{(\Phi^T\Phi)^{-1}\Phi^T}\mathbf{t}$$

ムーア・ペンローズの
擬似逆行列

ただし $\Phi=\begin{pmatrix}\phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1)\\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2)\\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N)\end{pmatrix}$

1.2 最小二乗法の幾何学

- 幾何学的に考える

$$\mathbf{y} = \begin{pmatrix} y(\mathbf{x}_1, \mathbf{w}) \\ y(\mathbf{x}_2, \mathbf{w}) \\ \vdots \\ y(\mathbf{x}_N, \mathbf{w}) \end{pmatrix}$$

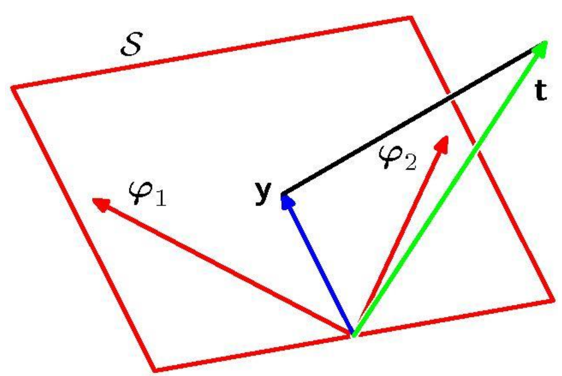
は$\varphi_j = \begin{pmatrix} \phi_j(\mathbf{x}_1) \\ \phi_j(\mathbf{x}_2) \\ \vdots \\ \phi_j(\mathbf{x}_N) \end{pmatrix}$ $j \in \{0,\dots M-1\}$で張られる線形部分空間S上にある.

二乗和誤差

$$\{\mathbf{t} - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 = (\mathbf{t} - \mathbf{y})^2$$

は**t**と**y**の「距離の二乗」

例）2つのベクトルで張られる線形部分空間



最尤推定解$\mathbf{w}_{ML}$を求めることは,
線形部分空間Sにあるベクトルの中で,
最も**t**と近いベクトルを求めること.

⇒Yはtの線形部分空間Sへの正射影

1.4 正則化最小二乗法

- 過学習を防ぐため，誤差関数

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{\mathbf{t} - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$

に正則化項を加えた

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

を最小化する．

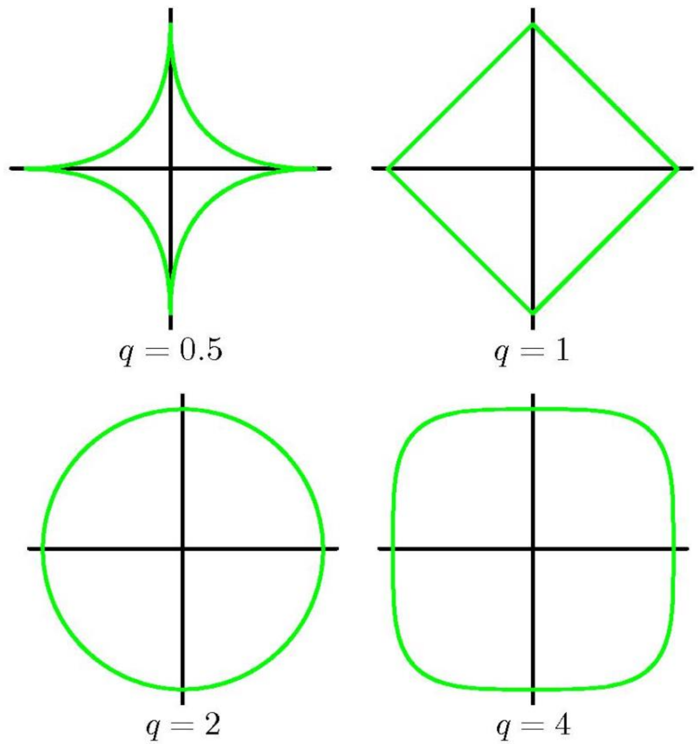
- 正則化項の例

単純形： $\frac{1}{2}\mathbf{w}^T\mathbf{w}$

一般形： $\frac{1}{2}\sum_{j=1}^{M}|w_j|^q$

q=1のときlasso

例）様々なqに対する正則化項の等高線表示



1.4 正則化最小二乗法

$$\frac{1}{2}\sum_{n=1}^{n}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$ の最小化は

$$\frac{1}{2}\sum_{n=1}^{n}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$ を，制約条件

$$\sum_{j=1}^{n}|w_j|^q \le \eta$$ の下で最小化するのと等価

- 例）2次元の場合

$$\sum_{n=1}^{2}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$

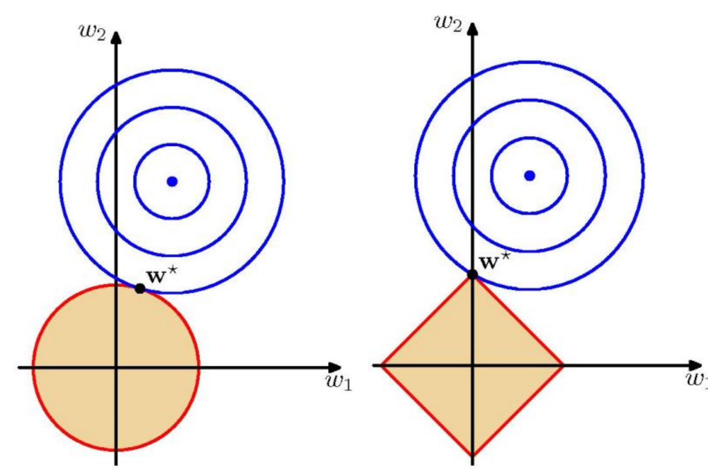
$$=\{t_1 - \mathbf{w}^T\phi(\mathbf{x}_1)\}^2 + \{t_2 - \mathbf{w}^T\phi(\mathbf{x}_2)\}^2$$

$$=\boxed{\{t_1 - (\omega_1\phi_1(\mathbf{x}_1) + \omega_2\phi_2(\mathbf{x}_1))\}^2 + \{t_2 - (\omega_1\phi_1(\mathbf{x}_2) + \omega_2\phi_2(\mathbf{x}_2))\}^2}$$

ω_1, ω_2 に関する楕円の式

q=2のとき

q=1のとき
※疎な解が得られる



2.バイアス-バリアンス分解

- 損失関数の予測値（条件付き期待値）

$$h(\mathbf{x}) = \mathbf{E}[t \mid \mathbf{x}] = \int t p(t \mid \mathbf{x}) dt$$

- 期待二乗損失

$$\mathbf{E}[L] = \int \boxed{\{y(\mathbf{x}) - h(\mathbf{x})\}^2} p(\mathbf{x}) d\mathbf{x} + \int\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

$$\Downarrow$$

この項を最小化したいが…データは有限個

- データ集合の取り方を考慮

$$\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2$$

$$= \{y(\mathbf{x};\mathcal{D}) - \mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})] + \mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2$$

$$= \{y(\mathbf{x};\mathcal{D}) - \mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})]\}^2 + \{\mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2$$

$$2\{y(\mathbf{x};\mathcal{D}) - \mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})]\}\{\mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}$$

2.バイアス-バリアンス分解

- 期待値を取ると

$$\mathrm{E}_{\mathcal{D}}[\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2]$$

$$=\underbrace{\{\mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})]-h(\mathbf{x})\}^2}_{（バイアス）^2}+\underbrace{\mathrm{E}_{\mathcal{D}}[\{[y(\mathbf{x};\mathcal{D})]-\mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})]\}^2]}_{バリアンス}$$

- **バイアス**：
  回帰関数とすべてのデータ集合の取り方に関する予測値の平均からのずれ
- **バリアンス**：
  個々のデータ集合に対する解が特定のデータ集合の選び方に関する期待値の周りでの変動の度合い

2.バイアス-バリアンス分解

- もとの損失関数に戻すと

$$\mathbf{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} + \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

$$= \boxed{\int \{\mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 d\mathbf{x}}$$

$$+ \boxed{\int \mathrm{E}_{\mathcal{D}}[\{[y(\mathbf{x};\mathcal{D})] - \mathrm{E}_{\mathcal{D}}[\mathrm{y}(\mathbf{x};\mathcal{D})]\}^2] \; d\mathbf{x}}$$

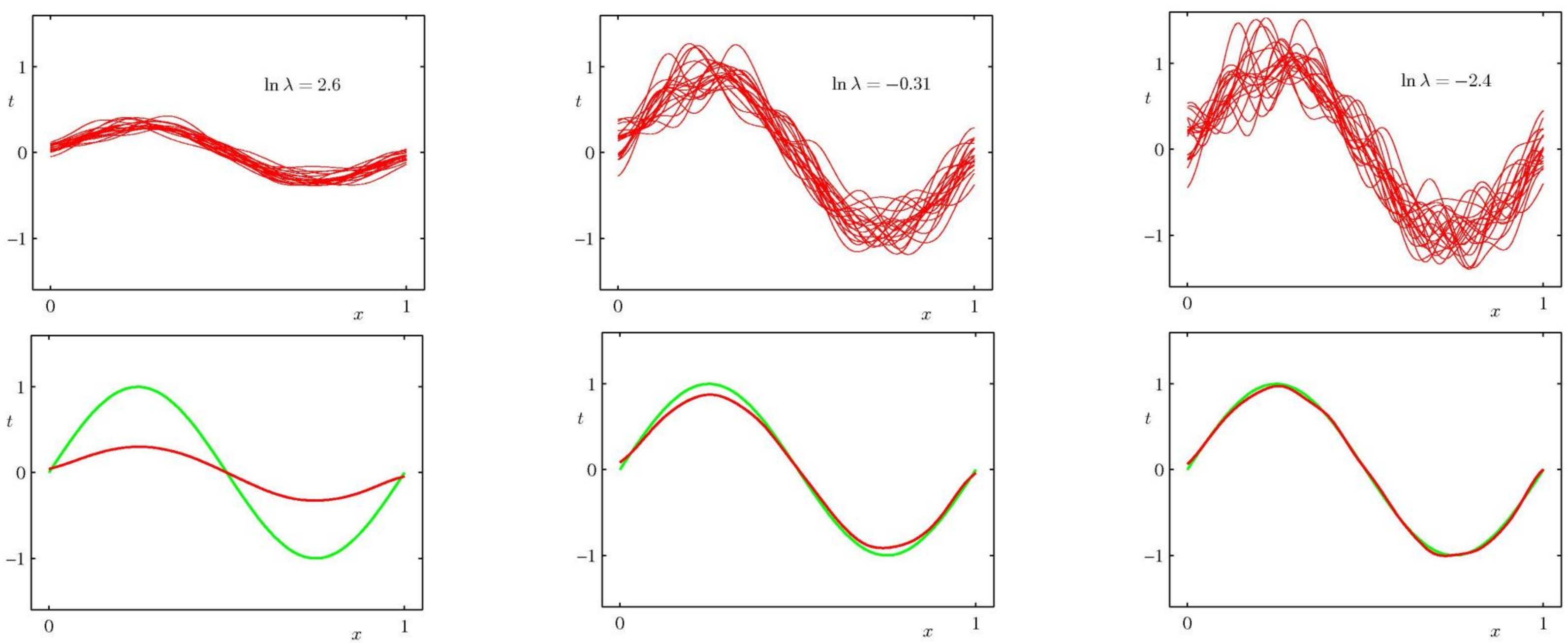
$$+ \boxed{\iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt}$$

$$= \boxed{(\text{バイアス})^2} + \boxed{\text{バリアンス}} + \boxed{\text{ノイズ}}$$

- バイアスとバリアンスをバランスよく小さくすることが必要

2.バイアス-バリアンス分解

- 例）$h(x) = \sin(2\pi x)$
  - サンプル25点からなる100種類のデータ集合
  - 25個のガウス関数をフィット



バイアス大，バリアンス小

バイアス小，バリアンス大

3. ベイズ線形回帰

- 最尤推定
  - モデルの複雑さはデータサイズに依存
  - 正則化項で調整
  - 過学習の可能性
- ベイズ線形回帰
  - パラメータを確率変数として扱う

3.1 パラメータの分布

- 尤度関数

$$p(\mathbf{t}\,|\,\mathbf{X},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n\,|\,\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1})$$

の指数部分は**w**の2次関数

⇒事前分布はガウス分布

$$p(\mathbf{w})=\mathcal{N}(\mathbf{w}\,|\,\mathrm{m}_0,\mathrm{S}_0)$$

- 事後分布

$$p(\mathbf{w}\,|\,\mathbf{t})=\mathcal{N}(\mathbf{w}\,|\,\mathrm{m}_N,\mathrm{S}_N)$$

$$\mathrm{m}_N=\mathrm{S}_N(\mathrm{S}_0^{-1}+\beta\Phi^T\mathbf{t})$$

$$\mathrm{S}_\mathrm{N}^{-1}=\mathrm{S}_0^{-1}\beta\Phi^T\Phi$$

3.1 パラメータの分布

- 事前分布を

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I})$$

- とすると，事後分布は次のように単純になる

$$p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathrm{m}_N, \mathrm{S}_N)$$

$$\mathrm{m}_N = \beta \mathrm{S}_N \Phi^T \mathbf{t}$$

$$\mathrm{S}_\mathrm{N}^{-1} = \alpha I + \beta \Phi^T \Phi$$

3.1 パラメータの分布

- 例）線形基底関数モデル $y(x,\mathbf{w}) = w_0 + w_1 x$

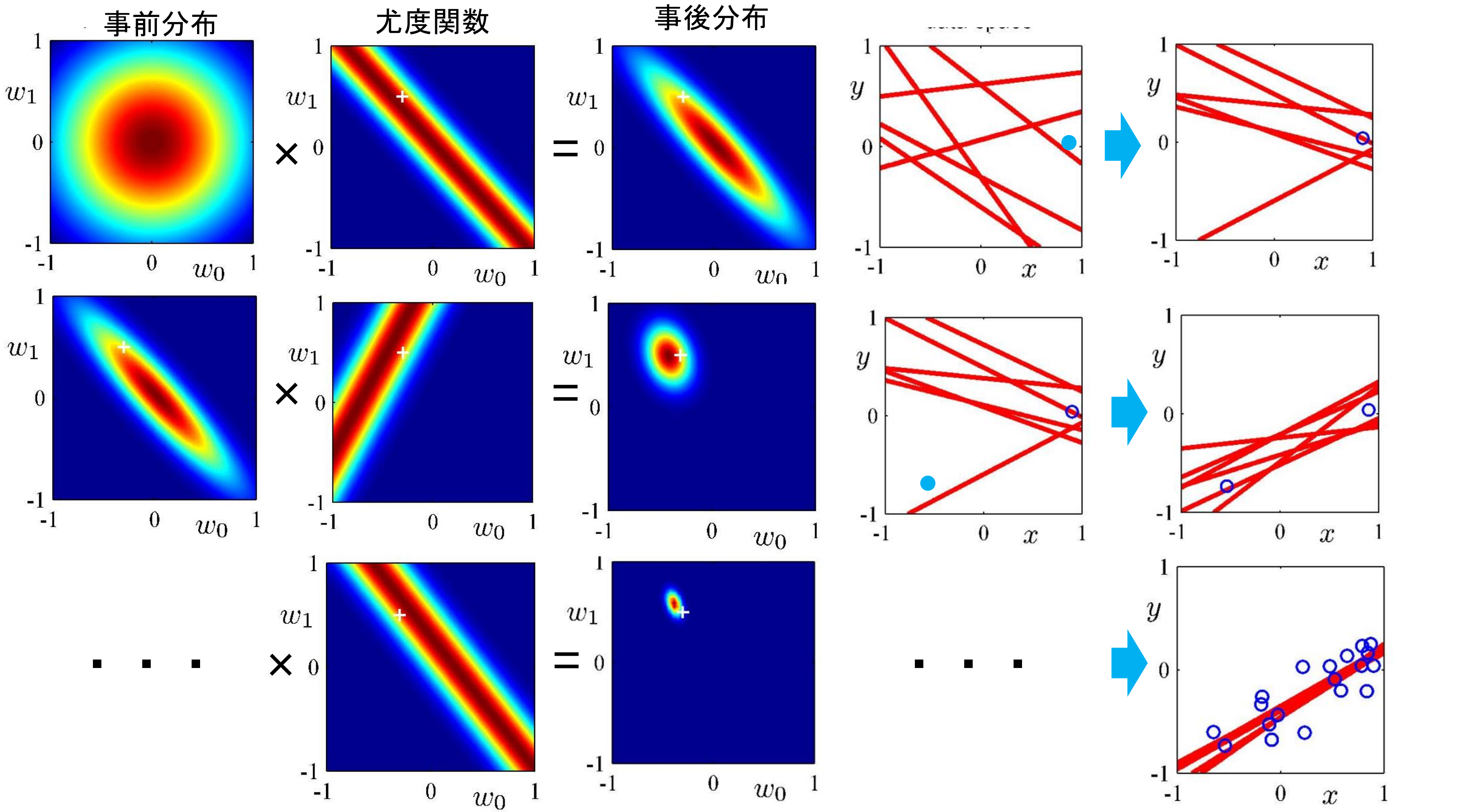
  関数

  $$f(x,\mathbf{a}) = a_0 + a_1 x \quad (a_0 = -0.3, a_1 = 0.5)$$

  を復元する.

1. 初期値を適当に（復元する関数周辺で）取り出す
2. 初期値から尤度関数を求める
3. 尤度関数と事前分布をかけて，パラメータの事後分布を求める
4. パラメータの事後分布から適当に取り出し，関数を推定する.
5. データ点を再度取り出す
6. 2～5を繰り返す

3.1 パラメータの分布



3.2 予測分布

- 予測分布:tを予測したい

$$p(t|\mathbf{t},\alpha,\beta)=\int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w}$$

ただし $p(t|\mathbf{w},\beta)=\mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1})$

$$p(\mathbf{w}|\mathbf{t},\alpha,\beta)=\mathcal{N}(\mathbf{w}|\mathrm{m}_N,\mathrm{S}_N)$$

- 結局

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta)=\mathcal{N}(\mathrm{t}|\mathbf{m}_{\mathrm{N}}^{\mathrm{T}}\phi(\mathbf{x}),\sigma_N^2(\mathbf{x}))$$

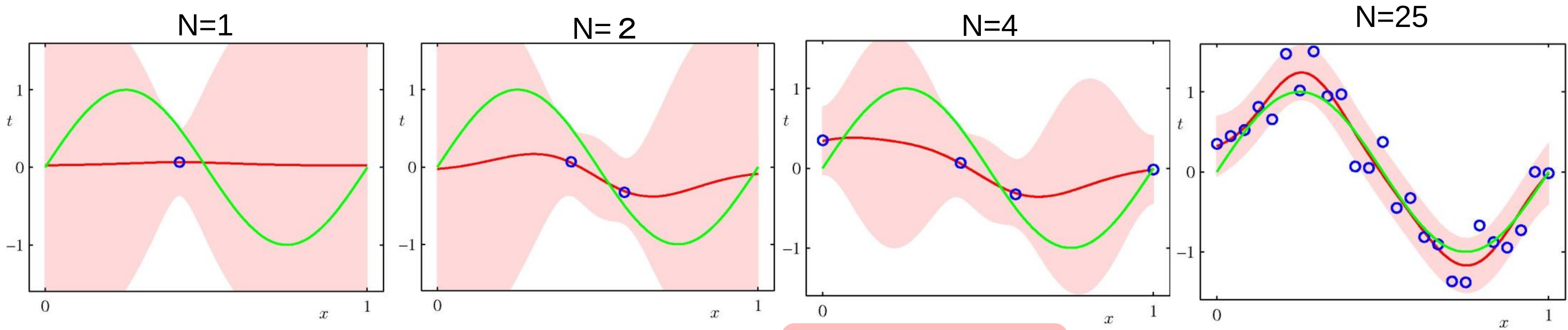
ただし $\sigma_N^2(\mathbf{x})=\boxed{\frac{1}{\beta}}+\boxed{\phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x})}$

Wに関する不確かさ

データに含まれる
ノイズ

3.2 予測分布

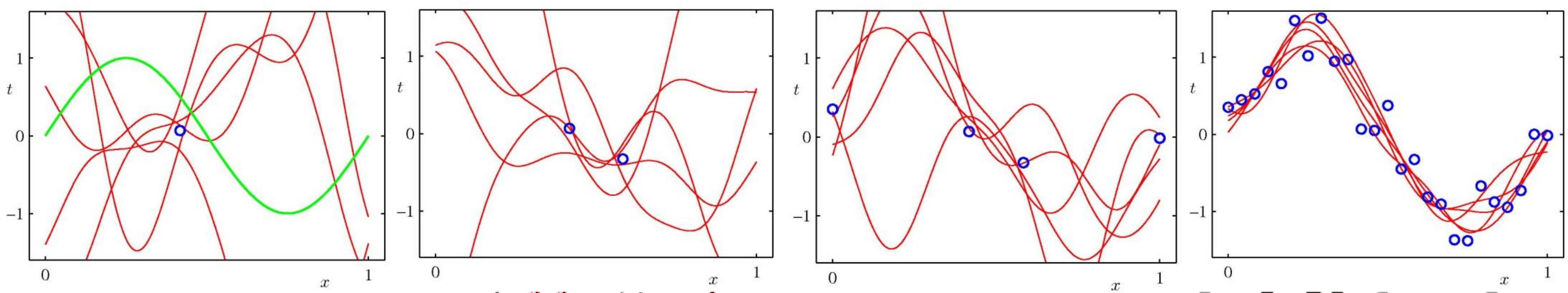
- 例）ガウス基底関数結合モデルの $\sin(2\pi x)$ へのあてはめ



ガウス予測分布の平均

+-標準偏差

wの事後分布から選んでプロットしたy(x, **w**)



3.3 等価カーネル

- 訓練データの目標値だけから予測する
- 線形基底関数モデルに対して
  事後分布の平均解を導入

$$y(\mathbf{x}, \mathrm{m}_N) = \mathrm{m}_N^T \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t} = \sum_{n=1}^{N} \beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(x_n) t_n$$

$$\mathrm{m}_N = \beta \mathrm{S}_N \Phi^T \mathbf{t}$$

$$\mathrm{S}_{\mathrm{N}}^{-1} = \alpha I + \beta \Phi^T \Phi$$

- つまり，訓練データの目標値t_nの線形結合

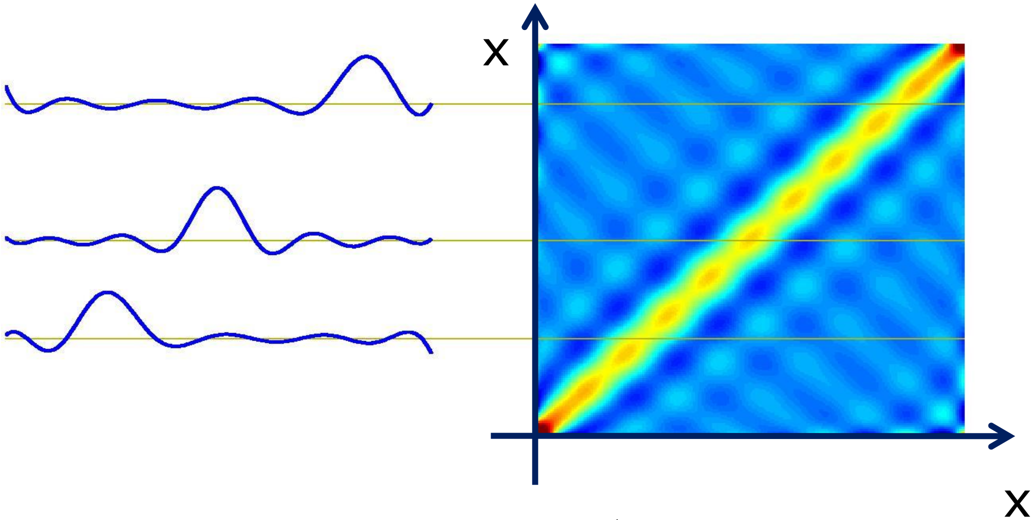
$$\boxed{y(\mathbf{x}, \mathrm{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n}$$

$$k(\mathbf{x}, \mathbf{x}') = \beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}')$$

平滑化行列または等価カーネル

3.3 等価カーネル

- ガウス基底関数に対するk(x,x')をプロット



⇒xに近いx'を大きく重みづけ

3.4 ベイズモデル比較

- データ集合 $\mathcal{D}$ 上のモデル集合 $\{\mathcal{M}_i\}(i=1,\ldots,L)$ からモデル選択をベイズ的に行う

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$$

- モデルエビデンス

$$p(\mathcal{D} \mid \mathcal{M}_i)$$

モデルでデータがどれぐらい説明できているかを表す.

- ベイズ因子

$$\frac{p(\mathcal{D} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_j)}$$

3.4 ベイズモデル比較

- モデルエビデンスは確率の加法・乗法定理により

$$p(\mathcal{D}\,|\,\mathcal{M}_i) = \int p(\mathcal{D}\,|\,\mathbf{w},\mathcal{M}_i)\,p(\mathbf{w}\,|\,\mathcal{M}_i)\,d\mathbf{w}$$

となる.
⇒パラメータを事前分布から適当にサンプリングしたときにデータ集合 $\mathcal{D}$が生成される確率

3.4 ベイズモデル比較

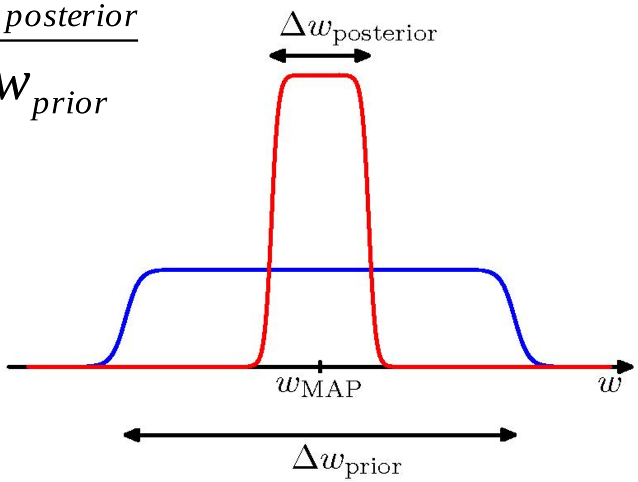
- 例）パラメータ1つのモデル

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid w)\, p(w)\, dw$$

事後分布：最頻値付近で尖って，幅 $\Delta w_{posterior}$

事前確率：平坦で，幅 Δw_{prior}

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid w)\, p(w)\, dw \approx p(\mathcal{D} \mid w_{MAP}) \frac{\Delta w_{posterior}}{\Delta w_{prior}}$$



対数をとると

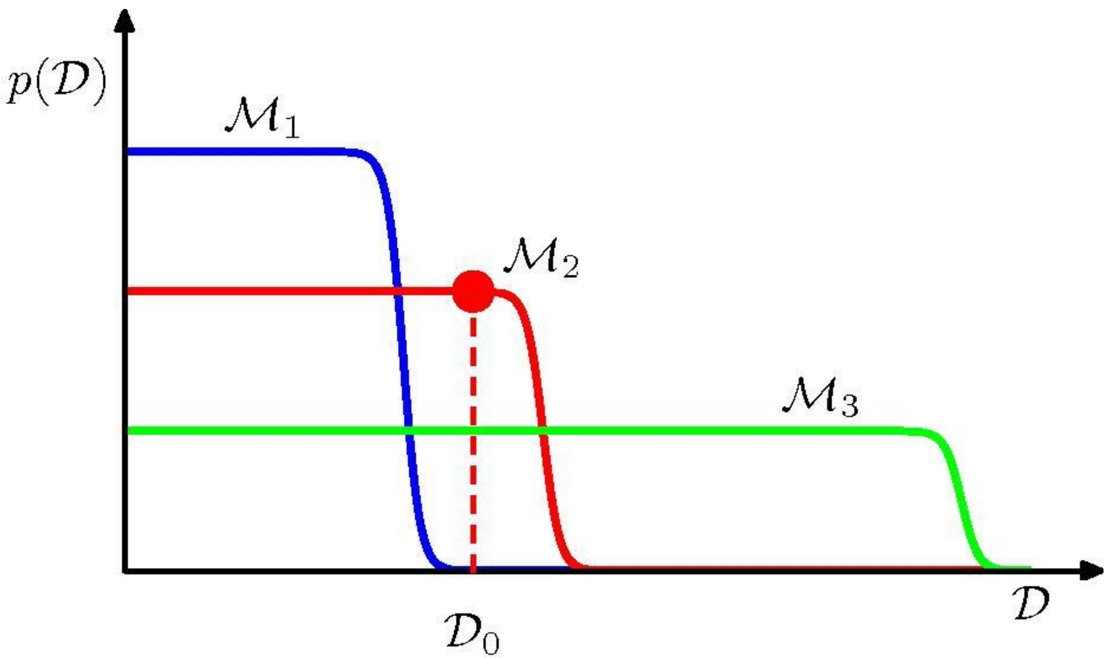
$$\ln p(\mathcal{D}) \approx \boxed{\ln p(\mathcal{D} \mid w_{MAP})} + \boxed{\ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)}$$

データへの
フィッティング度

ペナルティ項

3.4 ベイズモデル比較

- 3つのモデルの比較.
- 複雑さは $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ の順で大きくなる



$\mathcal{M}_1$：単純なモデル
生成できるデータ集合の範囲が狭く，データにフィットできない.

$\mathcal{M}_3$：複雑なモデル
得られるデータは広範囲だが，割り当てられる確率は低い

3.5 エビデンス近似

- パラメータwの分布を決める超パラメータα,βについても事前分布を考える

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w},\beta)\,p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,p(\alpha,\beta|\mathbf{t})\,d\mathbf{w}\,d\alpha\,d\beta$$

$$p(t|\mathbf{x},\mathbf{w},\beta) = \boxed{\mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1})}\qquad p(\mathbf{w}|\mathbf{t}) = \boxed{\mathcal{N}(\mathbf{w},\alpha,\beta|\mathrm{m}_N,\mathrm{S}_N)}$$

$$\mathrm{m}_N = \beta \mathrm{S}_N \Phi^T \mathbf{t}$$

$$\mathrm{S}_\mathrm{N}^{-1} = \alpha I + \beta \Phi^T \Phi$$

$$p(\alpha,\beta|\mathbf{t}) \propto \boxed{p(\mathbf{t}|\alpha,\beta)}\,p(\alpha,\beta)$$

周辺尤度関数

- 周辺尤度関数を最大化することが目標

5.1 エビデンス関数の評価

- 周辺尤度関数をwに関する積分で表現

$$p(\mathbf{t}|\alpha,\beta)=\int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$$

- これまでの結果より

$$\boxed{p(\mathbf{t}|\alpha,\beta)=\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int\exp\{-E(\mathbf{w})\}d\mathbf{w}}$$

$$E(\mathbf{w})=\beta E_D(\mathbf{w})+\alpha E_W(\mathbf{w})=\frac{\beta}{2}\|\mathbf{t}-\Phi\mathbf{w}\|^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$=E(\mathrm{m}_N)+\frac{1}{2}(\mathbf{w}-\mathrm{m}_N)^T\mathbf{A}(\mathbf{w}-\mathrm{m}_N)$$ ←平方完成

$$\mathbf{A}=\alpha\mathbf{I}+\beta\Phi^T\Phi$$

$$E(\mathrm{m}_N)=\frac{\beta}{2}\|\mathbf{t}-\Phi\mathrm{m}_N\|^2+\frac{\alpha}{2}\mathrm{m}_N{}^T\mathrm{m}_N$$

5.2 エビデンス関数の最大化

- 周辺尤度の対数をとると

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathrm{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi)$$

- これを最大化するα,βの値は

$$\alpha = \frac{\gamma}{\mathrm{m}_N^T\mathrm{m}_N} \qquad \frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \mathrm{m}_N^T\phi(\mathbf{x}_N)\}^2$$

$$\gamma = \sum_i \frac{\lambda_i}{\beta+\lambda_i}$$

λ_iは$\beta\Phi^T\Phi$の固有値