ガウス過程の基礎と教師なし学習

持橋大地
統計数理研究所
daichi@ism.ac.jp

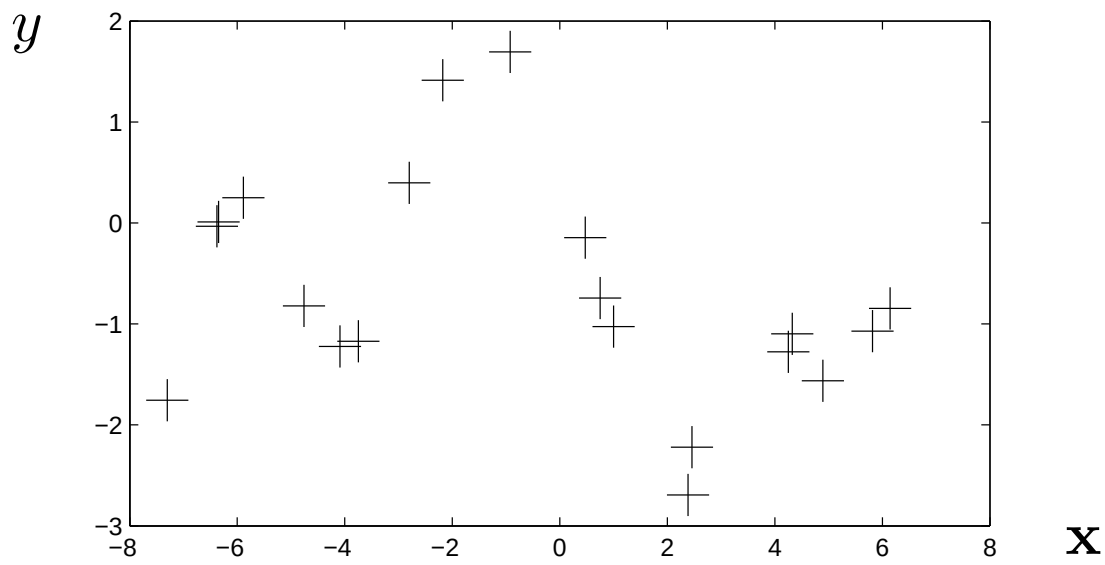
統計数理研究所公開講座資料
2015-3-3(火)

統計数理研究所
The Institute of Statistical Mathematics

Overview

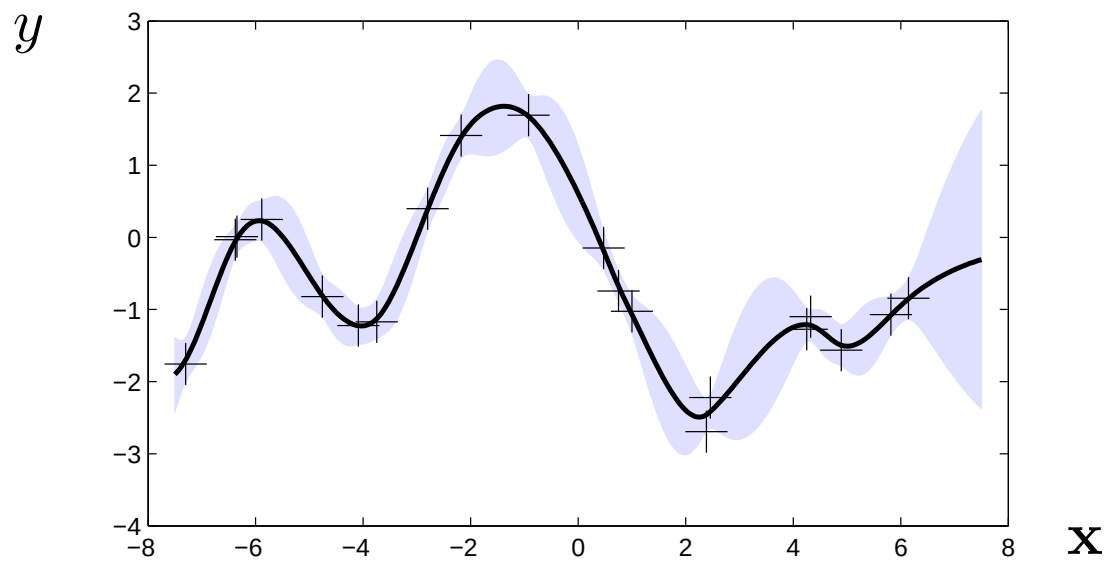
- ガウス過程 (Gaussian Process) とは
- 線形回帰から非線型回帰へ
- ガウス過程の最適化とその問題
- GPLVM とその最適化
- GPDM
- 最近のガウス過程研究

ガウス過程 (Gaussian Process) とは



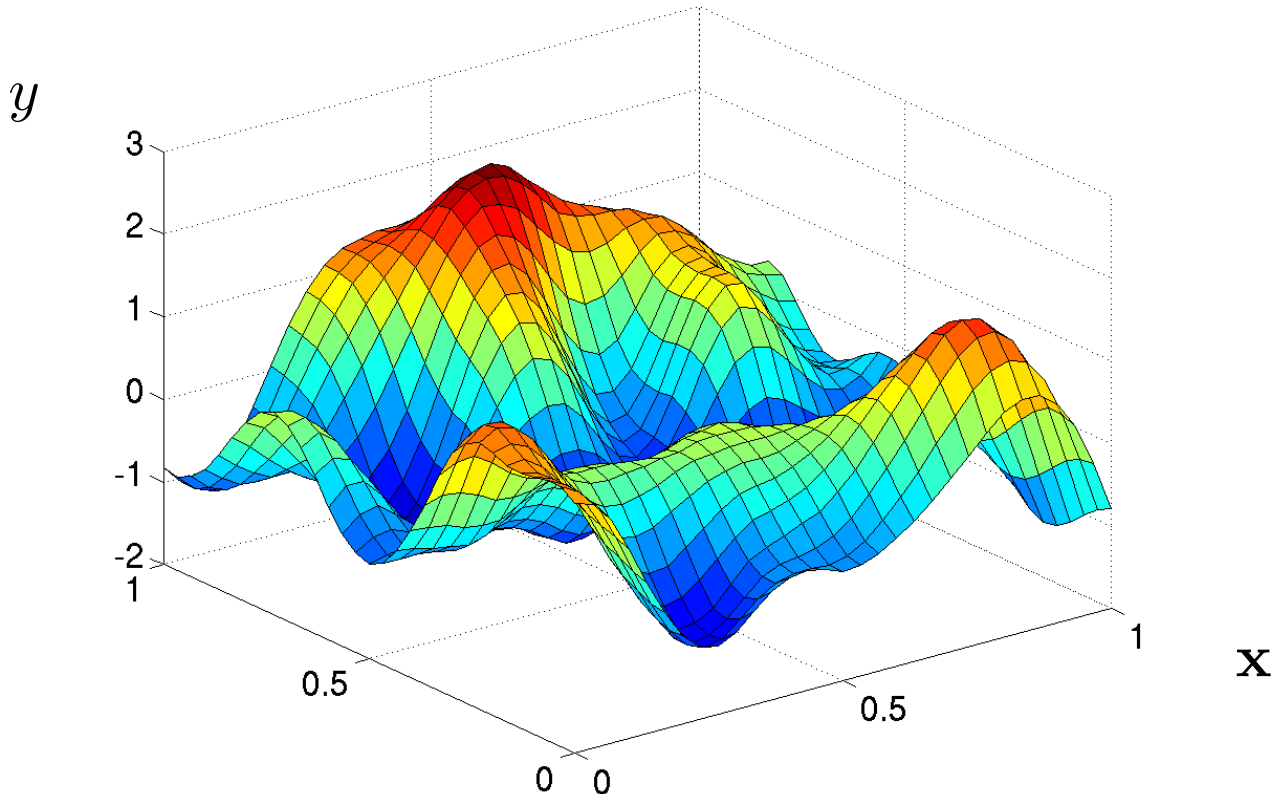
- 入力 $\mathbf{x} \to y$ を予測する回帰関数 (regressor) の確率モデル
  - データ $D = \left\{(\mathbf{x}^{(n)}, y^{(n)})\right\}_{n=1}^{N}$ が与えられた時, 新しい $\mathbf{x}^{(n+1)}$ に対する $y^{(n+1)}$ を予測
  - ランダムな関数の確率分布
  - 連続空間で動く, ベイズ的なカーネルマシン (後で)

ガウス過程 (Gaussian Process) とは



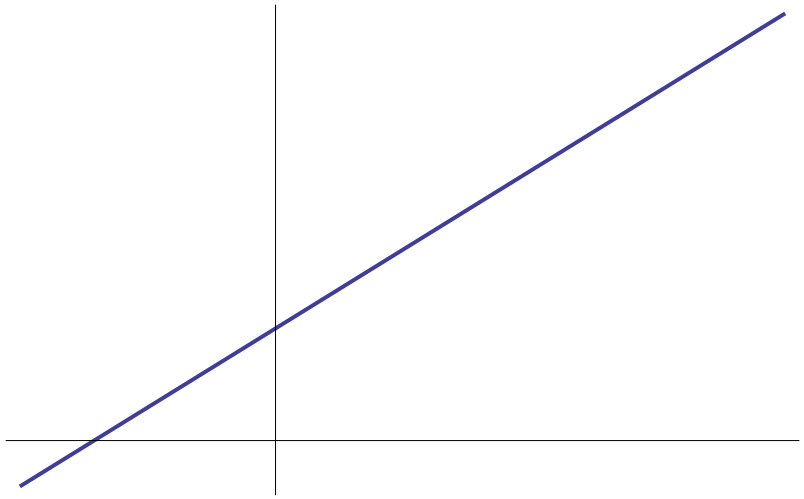
- 入力 $\mathbf{x} \to y$ を予測する回帰関数 (regressor) の確率モデル
  - データ $D = \left\{(\mathbf{x}^{(n)}, y^{(n)})\right\}_{n=1}^{N}$ が与えられた時, 新しい $\mathbf{x}^{(n+1)}$ に対する $y^{(n+1)}$ を予測
  - ランダムな関数の確率分布
  - 連続空間で動く, ベイズ的なカーネルマシン (後で)

ガウス過程 (Gaussian Process) とは



- 入力 $\mathbf{x} \to y$ を予測する回帰関数 (regressor) の確率モデル
  - データ $D = \left\{(\mathbf{x}^{(n)}, y^{(n)})\right\}_{n=1}^{N}$ が与えられた時, 新しい $\mathbf{x}^{(n+1)}$ に対する $y^{(n+1)}$ を予測
  - ランダムな関数の確率分布
  - 連続空間で動く, ベイズ的なカーネルマシン (後で)

線形モデル



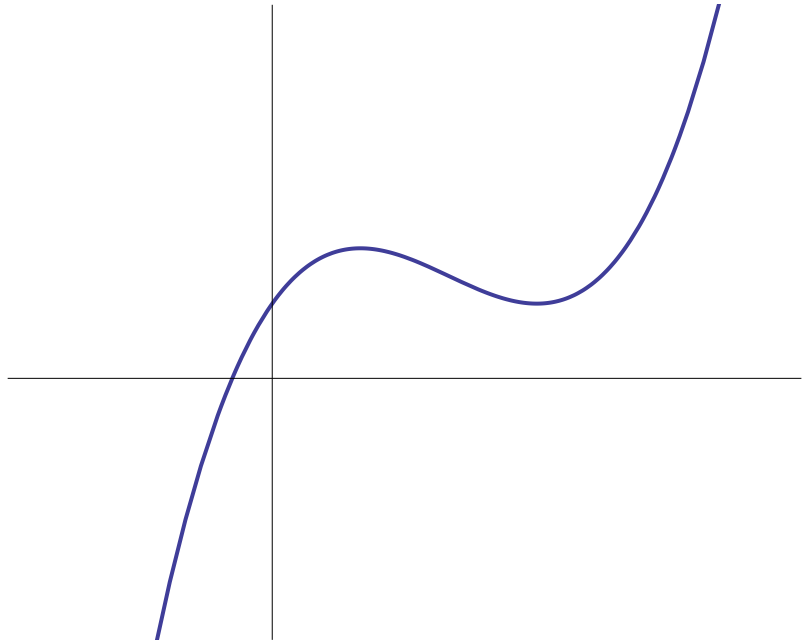
$$
\begin{aligned}
y &= w_0 + w_1 x_1 + w_2 x_2 + \epsilon \\
&= \underbrace{(w_0\ w_1\ w_2)}_{\mathbf{w}^T} \underbrace{\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}}_{\mathbf{x}} + \epsilon \\
&= \mathbf{w}^T \mathbf{x} + \epsilon
\end{aligned}
$$

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

(正規方程式)

- 簡単だが, 直線的な関係しか表せない … 「多変量解析」は今でもこれ

一般化線形モデル (GLM)



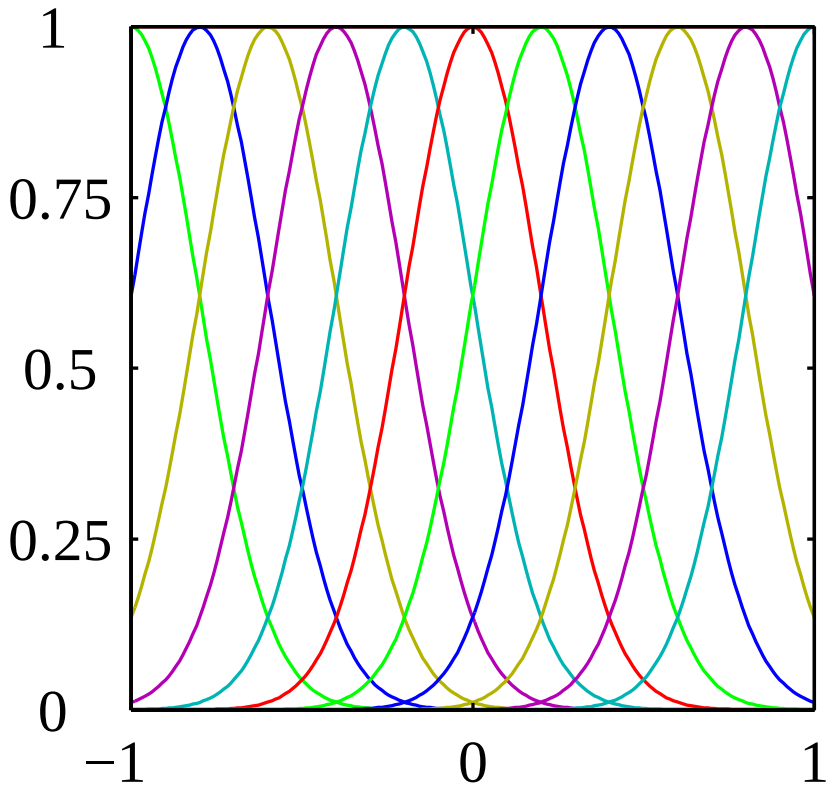
$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon \quad (1)$$

$$= \underbrace{(w_0\ w_1\ w_2\ w_3)}_{\mathbf{w}^T} \underbrace{\begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix}}_{\phi(\mathbf{x})} + \epsilon \quad (2)$$

$$= \mathbf{w}^T \phi(\mathbf{x}) + \epsilon \quad (3)$$

- 非線型な関係を表せる
- 基底関数 $\phi(\mathbf{x})$ が原点中心 $\rightarrow$ もっと複雑にしたい!

一般化線形モデル (GLM) (2)



$$\phi(\mathbf{x}) = \left(-\frac{(x-\mu_1)^2}{2\sigma^2}, \ -\frac{(x-\mu_2)^2}{2\sigma^2}, \ \cdots, -\frac{(x-\mu_K)^2}{2\sigma^2} \right) \quad (4)$$

- 基底関数により, 複雑な関数が作れる!
  - 基底関数のパラメータ $\mu = (\mu_1, \mu_2, \cdots, \mu_K)$ は固定・有限でよい?

線形回帰モデル

- 入力 $\mathbf{x}$ から出力 $y \in \mathbb{R}$ を予測する回帰関数 $y = f(\mathbf{x})$ を求めたい
  - $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ は時間や任意のベクトル
- $y = f(\mathbf{x})$ を, $\mathbf{x}$ を一般の関数 $\phi(\mathbf{x})$ で変換した上で線形モデルで表してみる

$$y = \mathbf{w}^T \phi(\mathbf{x}) \tag{5}$$

例： $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \cdots, \phi_H(\mathbf{x}))^T$

$$= (1, x_1, \cdots, x_d, x_1^2, \cdots, x_d^2)^T$$

$$\mathbf{w} = (w_0, w_1, \cdots, w_{2d})^T$$

のとき,

$$y = \mathbf{w}^T \phi(\mathbf{x})$$

$$= w_0 + w_1 x_1 + \cdots + w_d x_d + w_{d+1} x_1^2 + \cdots + w_{2d} x_d^2 \, .$$

GP の導入 (1)

- $y^{(1)} \dots y^{(N)}$ について同時に書くと, 下のように $\mathbf{y} = \mathbf{\Phi w}$ と行列形式で書ける (Φ：計画行列)

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} = \begin{pmatrix} \phi_1(\mathbf{x}^{(1)}) & \cdots & \phi_H(\mathbf{x}^{(1)}) \\ \phi_1(\mathbf{x}^{(2)}) & \cdots & \phi_H(\mathbf{x}^{(2)}) \\ \vdots & & \vdots \\ \phi_1(\mathbf{x}^{(N)}) & \cdots & \phi_H(\mathbf{x}^{(N)}) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ \vdots \\ w_H \end{pmatrix} \tag{6}$$

$$\mathbf{y} \qquad\qquad \mathbf{\Phi} \qquad\qquad \mathbf{w}$$

- 重み $\mathbf{w}$ がガウス分布 $p(\mathbf{w}) = \mathrm{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ に従っているとすると, $\mathbf{y} = \mathbf{\Phi w}$ もガウス分布に従い,
- 平均 0, 分散

$$\begin{aligned} \langle \mathbf{y}\mathbf{y}^T \rangle &= \left\langle (\mathbf{\Phi w})(\mathbf{\Phi w})^T \right\rangle = \mathbf{\Phi} \langle \mathbf{w}\mathbf{w}^T \rangle \mathbf{\Phi}^T \\ &= \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T \qquad \text{の正規分布となる} \end{aligned} \tag{7}$$

GP の導入 (2)

$$p(\mathbf{y}) = \mathrm{N}(\mathbf{y}|\mathbf{0}, \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T) \tag{8}$$

は, どんな入力 $\{\mathbf{x}_n\}_{n=1}^N$ についても成り立つ → ガウス過程の定義

- どんな入力 $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ についても, 対応する出力 $\mathbf{y} = (y_1, y_2, \cdots, y_N)$ がガウス分布に従うとき, $p(\mathbf{y})$ はガウス過程に従う　という.
  - ガウス過程 = 無限次元のガウス分布
  - ガウス分布の周辺化はまたガウス分布なので, 実際にはデータのある所だけの有限次元
- $\mathbf{K} = \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T$ の要素であるカーネル関数

$$k(\mathbf{x}, \mathbf{x}') = \alpha^{-1}\phi(\mathbf{x})^T\phi(\mathbf{x}') \tag{9}$$

だけでガウス分布が定まる
  - $k(\mathbf{x}, \mathbf{x}')$ は $\mathbf{x}$ と $\mathbf{x}'$ の距離 ; 入力 $\mathbf{x}$ が近い→出力 y が近い

GP の導入 (3)

- 実際には, 観測値にはノイズ ϵ が乗っている

$$\begin{cases} y = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon \\ \epsilon \sim \mathrm{N}(0, \beta^{-1}\mathbf{I}) \end{cases} \Longrightarrow \quad p(y|f) = \mathrm{N}(\mathbf{w}^T \phi(\mathbf{x}), \beta^{-1}\mathbf{I}) \tag{10}$$

- 途中の $f = \mathbf{w}^T \phi(\mathbf{x})$ を積分消去

$$p(y|\mathbf{x}) = \int p(y|f)p(f|\mathbf{x})df \tag{11}$$

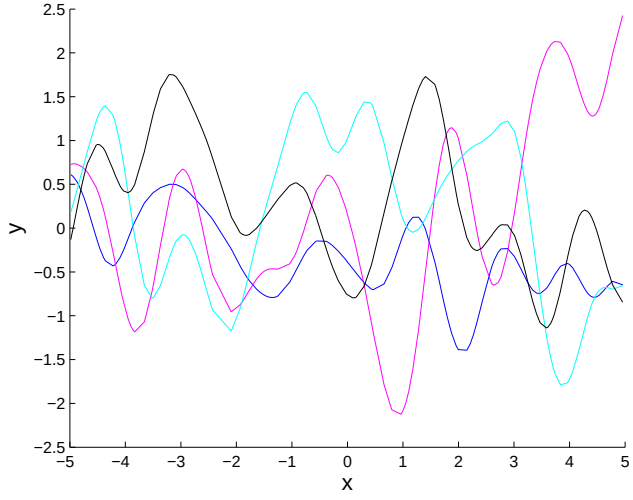
$$= \mathrm{N}(0, \mathbf{C}) \tag{12}$$

  - 二つの独立な Gaussian の畳み込みなので, $\mathbf{C}$ の要素は共分散の和:

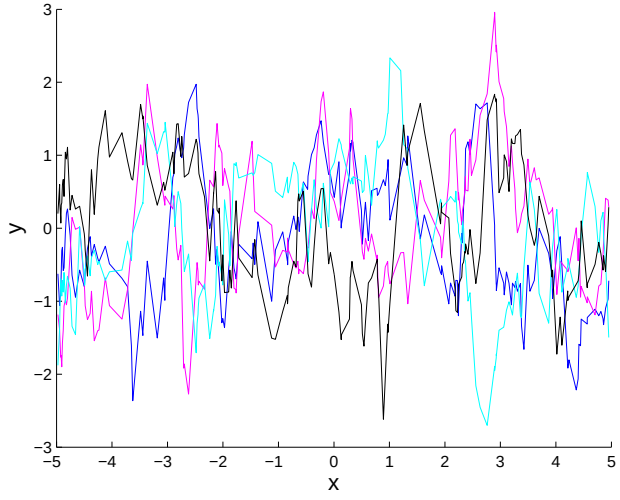
$$C(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \beta^{-1}\delta(i, j). \tag{13}$$

  - GP は, カーネル関数 $k(\mathbf{x}, \mathbf{x}')$ とハイパーパラメータ α, β だけで表すことができる.

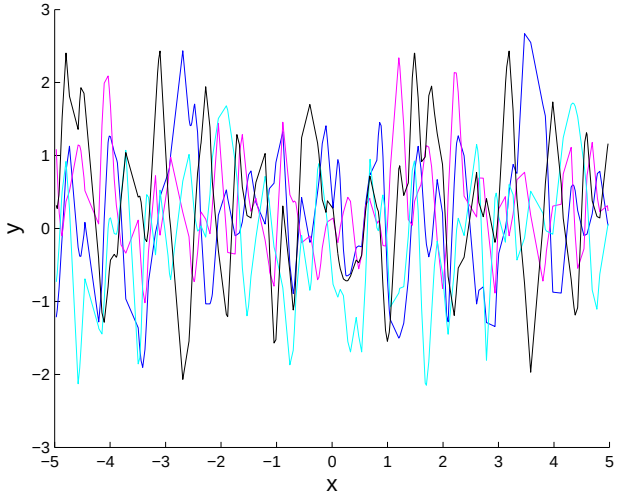
様々なカーネル



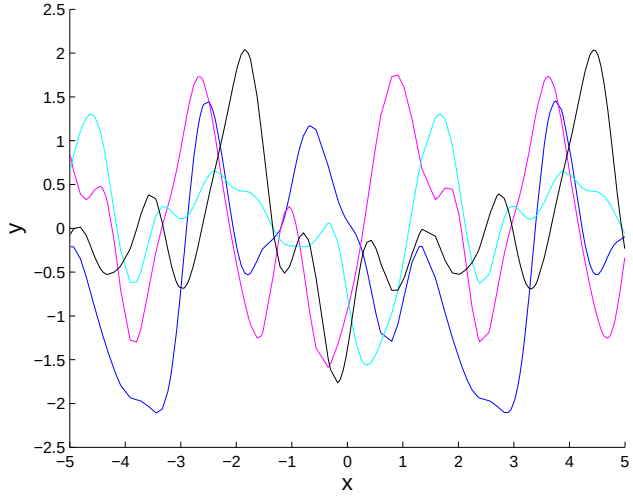
Gaussian: $\exp(-(\mathbf{x}-\mathbf{x}')^2/l)$



Exponential: $\exp(-|\mathbf{x}-\mathbf{x}'|/l)$ (OU process)



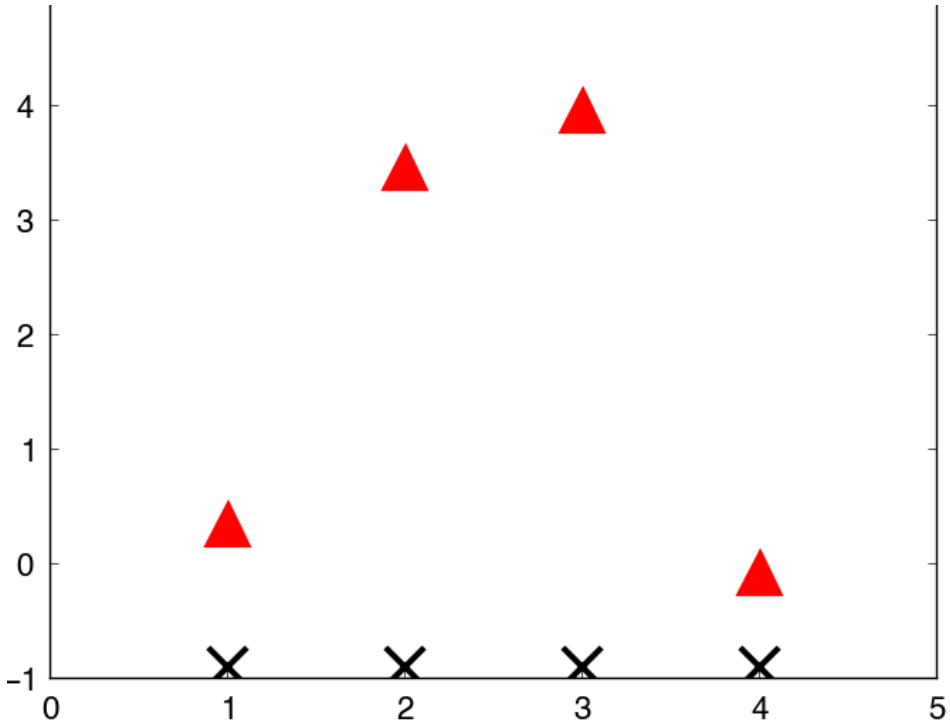
Periodic: $\exp(-2\sin^2(\frac{\mathbf{x}-\mathbf{x}'}{2})/l^2)$



Periodic(L): $\exp(-2\sin^2(\frac{\mathbf{x}-\mathbf{x}'}{2})/(10l)^2)$

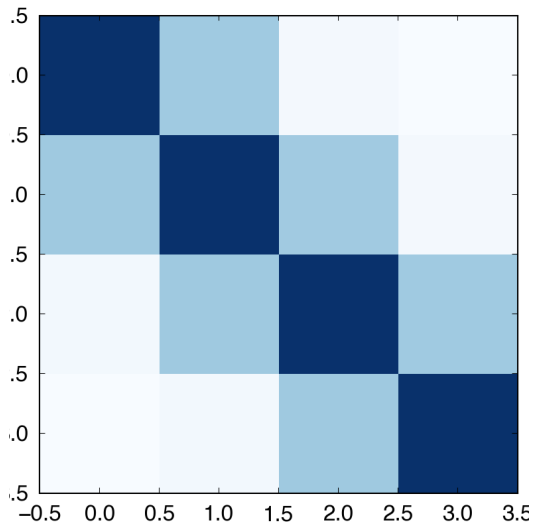
直観的理解

- Correlated Gaussian



ガウス分布からのサンプル

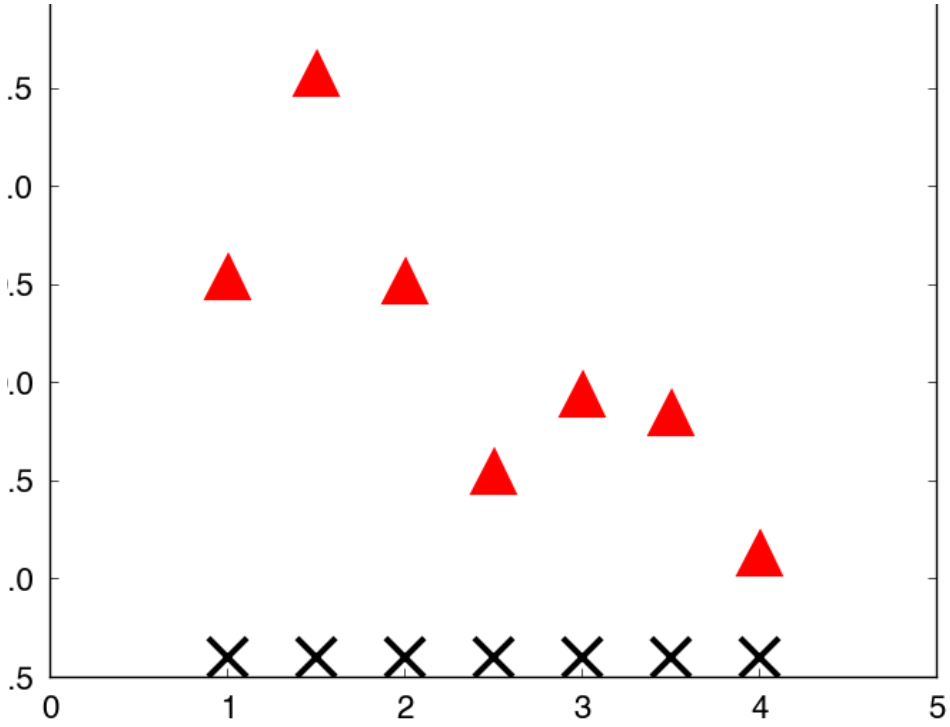
$K =$



分散・共分散行列

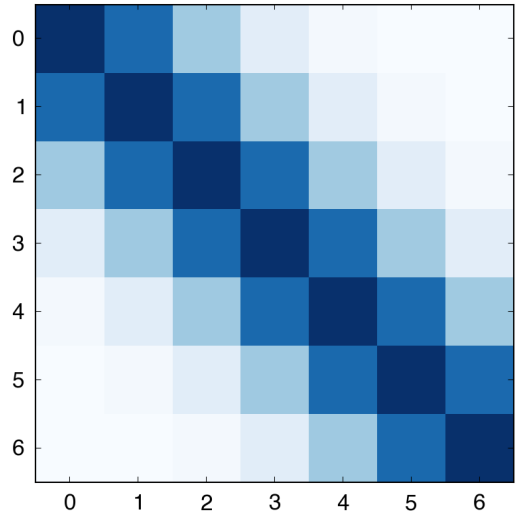
直観的理解 (2)

- Correlated Gaussian



ガウス分布からのサンプル

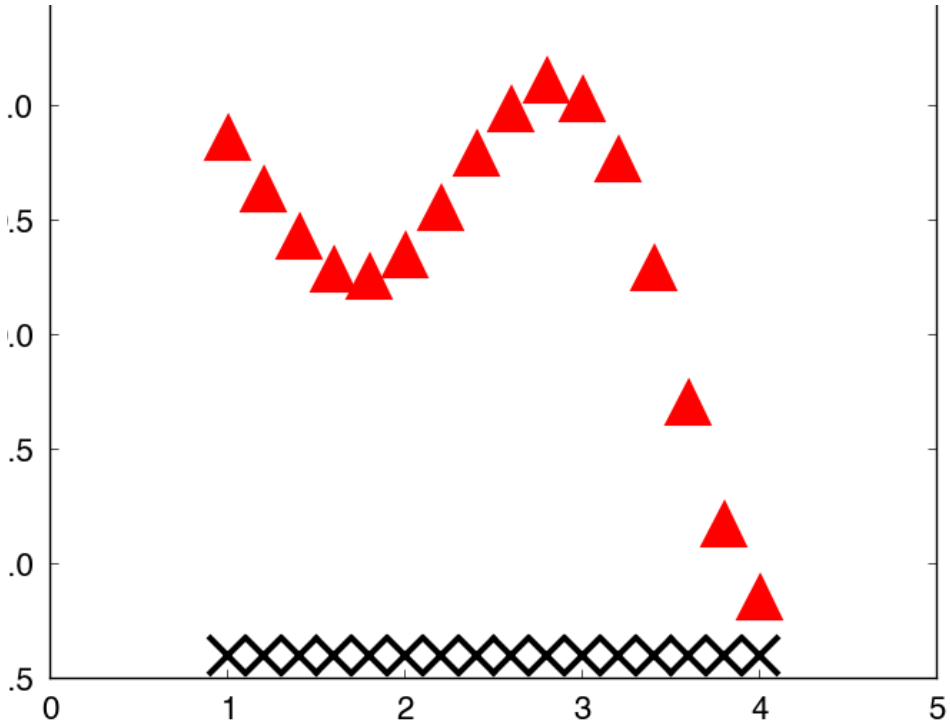
$K =$



分散・共分散行列

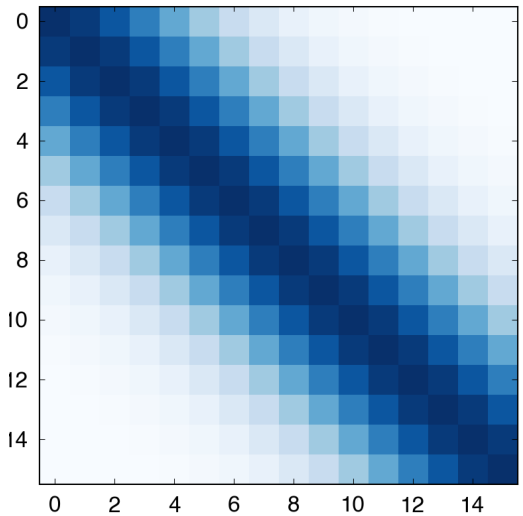
直観的理解 (3)

- Correlated Gaussian



ガウス分布からのサンプル

$K =$



分散・共分散行列

“Infinite” dimensional Gaussian

- もし, 任意の $(x_1, x_2, \cdots, x_n)$ について対応する $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ がガウス分布に従うなら, $\mathbf{y}$ はガウス過程に従うという.
  - 原理的には $(x_1, x_2, \cdots, x_n)$ 以外の次元もあるが, それらについて周辺化されている (←ガウス分布の周辺分布はガウス分布)
  - 「無限次元」のガウス分布.
- カーネル行列 K の要素 $K_{ij} = k(x_i, x_j)$ を与えるカーネル k がガウス過程のパラメータ.

「基底関数」の消去

- RBF 基底関数 $\phi(\mathbf{x}) = \exp((\mathbf{x} - \mathbf{h})^2/r^2)$ を考えてみる
- 1 次元の場合, h を無限個用意すると

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \sum_{h=1}^{H} \phi_h(\mathbf{x})\phi_h(\mathbf{x}') \tag{14}$$

$$\to \int_{-\infty}^{\infty} \exp\left(-\frac{(x-h)^2}{r^2}\right) \exp\left(-\frac{(x'-h)^2}{r^2}\right) dh \tag{15}$$

$$= \sqrt{\pi r^2} \exp\left(-\frac{(x-x')^2}{2r^2}\right) \equiv \theta_1 \exp\left(-\frac{(x-x')^2}{\theta_2^2}\right) \tag{16}$$

  - $(\mathbf{x}, \mathbf{x}')$ の RBF カーネルは, 無限個の RBF 基底関数を考えたことと等価.
  - カーネルのパラメータ θ_1, θ_2 は最尤推定で最適化できる

GP の予測

- 新しい入力 y^{new} とこれまでの $\mathbf{y}$ の結合分布がまた Gaussian になるので,

$$p(y^{\text{new}}|\mathbf{x}^{\text{new}}, \mathbf{X}, \mathbf{y}, \theta) \tag{17}$$

$$= \frac{p((\mathbf{y}, y^{\text{new}})|(\mathbf{X}, \mathbf{x}^{\text{new}}), \theta)}{p(\mathbf{y}|\mathbf{X}, \theta)} \tag{18}$$

$$\propto \exp\left(-\frac{1}{2}([\mathbf{y}, y^{\text{new}}] \begin{bmatrix} \mathbf{K} & \mathbf{k} \\ \mathbf{k}^T & k \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ y^{\text{new}} \end{bmatrix} - \mathbf{y}^T\mathbf{K}^{-1}\mathbf{y})\right) \tag{19}$$

$$\sim \mathrm{N}(\mathbf{k}^T\mathbf{K}^{-1}\mathbf{y}, k - \mathbf{k}^T\mathbf{K}^{-1}\mathbf{k}). \tag{20}$$

ここで

- $\mathbf{K} = [k(\mathbf{x}, \mathbf{x}')]$.
- $\mathbf{k} = (k(\mathbf{x}^{\text{new}}, \mathbf{x}_1), \cdots, k(\mathbf{x}^{\text{new}}, \mathbf{x}_N))$.

GP $\leftrightarrow$ SVR, Ridge

- 機械翻訳の自動評価, ARD カーネル関数 (Cohn+ 2013)

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_k \frac{(x_k - x_k')^2}{\sigma_k^2}\right) \tag{21}$$

Model	MAE	RMSE
μ	0.8279	0.9899
SVM	0.6889	0.8201
Linear ARD	0.7063	0.8480
Squared exp. Isotropic	0.6813	0.8146
Squared exp. ARD	**0.6680**	**0.8098**
Rational quadratic ARD	0.6773	0.8238
Matern(5,2)	0.6772	0.8124
Neural network	0.6727	0.8103

GP $\leftrightarrow$ SVR, Ridge

- 機械翻訳の自動評価, ARD カーネル関数 (Cohn+ 2013)

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_k \frac{(x_k - x_k')^2}{\sigma_k^2}\right) \tag{22}$$

Model	MAE	RMSE
μ	0.8541	1.0119
Independent SVMs	0.7967	0.9673
EasyAdapt SVM	0.7655	0.9105
Independent	0.7061	0.8534
Pooled	0.7252	0.8754
Pooled & {N}	0.7050	0.8497
Combined	**0.6966**	**0.8448**

GP>SVR の利点

確率モデルなので,

- ハイパーパラメータが第二種最尤推定で求められる
- 予測の期待値だけでなく, 分散も求まる
  - 予測がどのくらい確かか／あやふやか がわかる
- 複数の予測タスクを関係づけられる (Cohn+ 2014 etc.)
- **性能が高い**!

GP の計算量

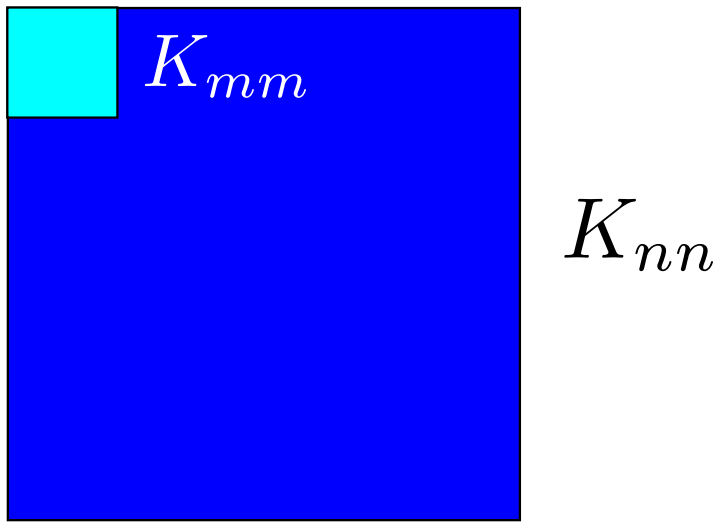
- GP の問題: 学習/推論時に X のグラム行列 K^{-1} を計算する必要$\cdots O(N^3)$ の計算量
  - $N > 1000$ を超えると, 実質的に計算不可能
- 解決:「代表的」な仮想的な m 個の入力 X_m を考え, もとの尤度を近似するように X_m を最適化 $\rightarrow O(m^2 N)$ の計算量

ナイーブな方法

- Subset of Data : データの一部でカーネル行列を計算

$$K \simeq K_{mm} \tag{23}$$

  - データのうち, ランダムな m 個のみによるカーネル行列
  - 計算量 $O(m^3)$
- 単純にデータをほとんど捨てている→分散が大きく, 低精度

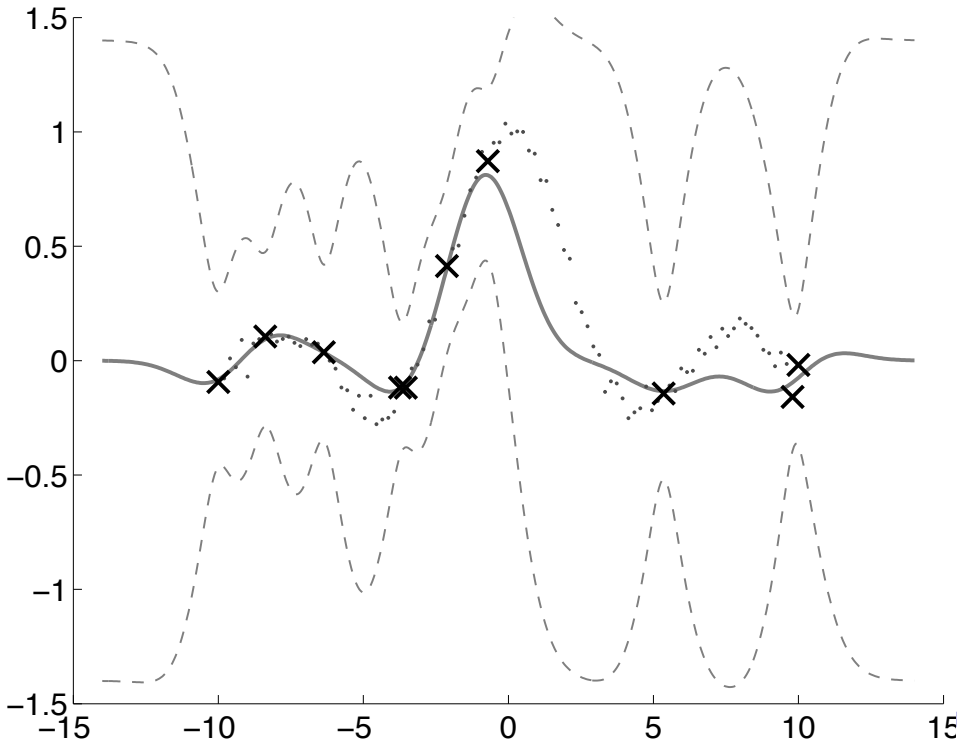


ナイーブな方法

- Subset of Data：データの一部でカーネル行列を計算

$$K \simeq K_{mm} \tag{24}$$

  - データのうち, ランダムな m 個のみによるカーネル行列
  - 計算量 $O(m^3)$
- 単純にデータをほとんど捨てている→分散が大きく, 低精度

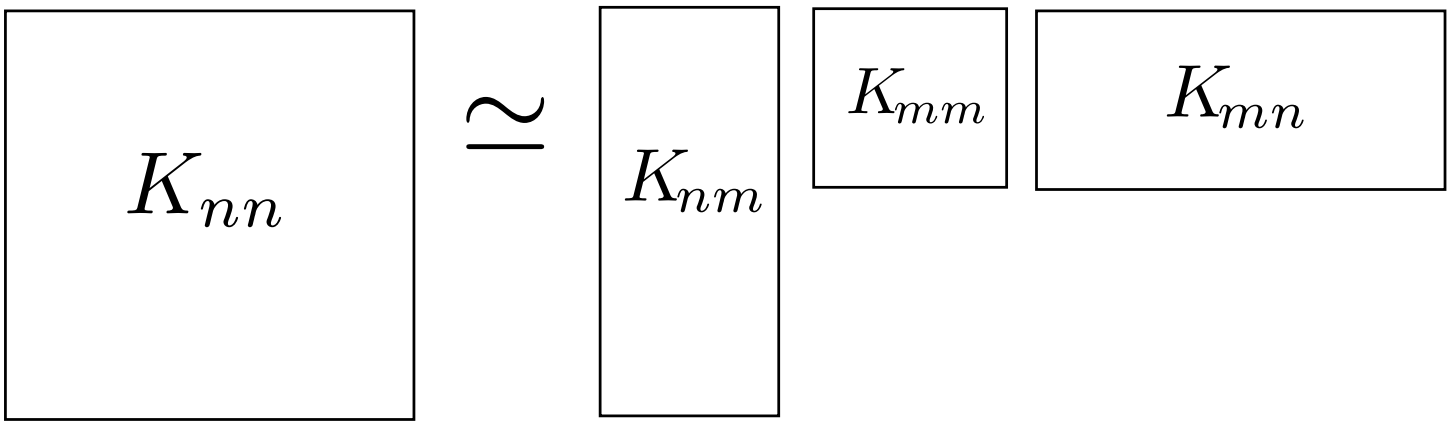


ナイーブな方法 (2)

- Subset of Regressors (Silverman 1985) :
  m 個の基底でカーネル行列を近似

$$K \simeq K_{nm} K_{mm}^{-1} K_{mn} = K' \tag{25}$$

  - K_{nm} : $N \times m$ のグラム行列
  - 計算量 $O(m^2 N)$

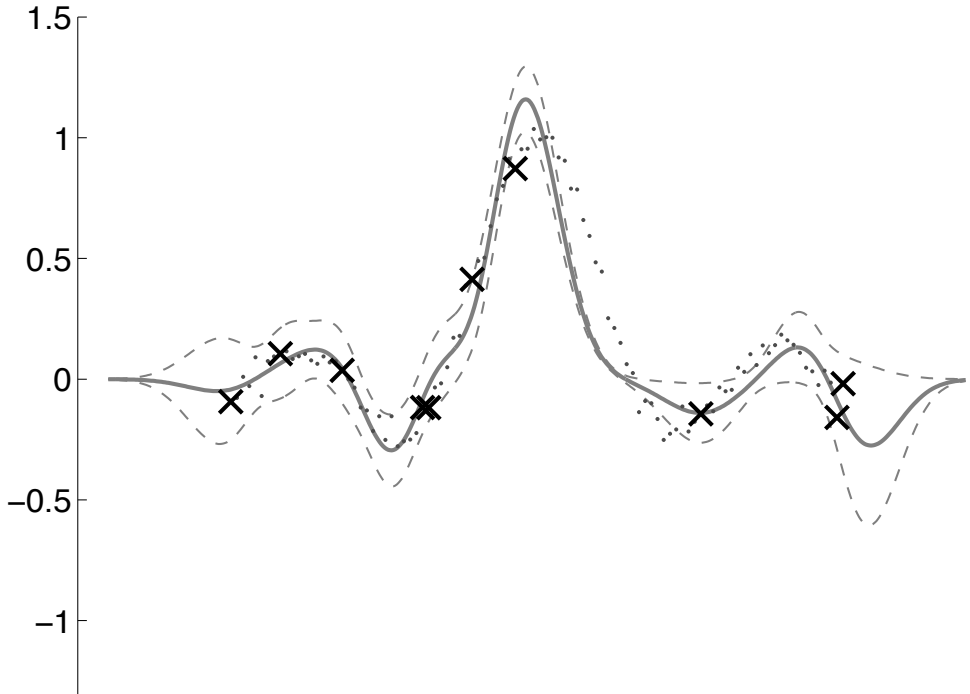


ナイーブな方法 (2)

- Subset of Regressors (Silverman 1985) :
  m 個の基底でカーネル行列を近似

$$K \simeq K_{nm} K_{mm}^{-1} K_{mn} = K' \tag{26}$$

  - K_{nm} : $N \times m$ のグラム行列
  - 計算量 $O(m^2 N)$
- 目的関数に選んだ基底が含まれているため, 過学習が起こる



ナイーブな方法の問題点

- カーネル行列 K を変更することで, モデル自体を書き換えている
  - 真のモデルと違った事前分布による学習
    (Quiñonero-Candela & Rasmussen 2005)
  - 真の分布との「距離」が必要

    $\Downarrow$

  変分ベイズ法.

変分近似 (Titsias 2009)

- 真の目的関数を変えてしまうのではなく, 真の目的関数を下から近似する
  - Jensen の不等式:

$$\log \int p(x) f(x) dx \geq \int p(x) \log f(x) dx$$

- 入力 X_m に対応する GP の値を f_m とおくと,

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m) d\mathbf{f} d\mathbf{f}_m \tag{27}$$

$$= \log \int q(\mathbf{f}, \mathbf{f}_m) \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m)}{q(\mathbf{f}, \mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{28}$$

$$\geq \int q(\mathbf{f}, \mathbf{f}_m) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m)}{q(\mathbf{f}, \mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{29}$$

  - ここで, $q(\mathbf{f}, \mathbf{f}_m)$ は下限を作るための補助分布

変分近似 (2)

- $p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m)$ だから,

$$q(\mathbf{f}, \mathbf{f}_m) = p(\mathbf{f}|\mathbf{f}_m)q(\mathbf{f}_m)$$

とおけば,

$$\log p(\mathbf{y}) \geq \int q(\mathbf{f}, \mathbf{f}_m) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m)}{q(\mathbf{f}, \mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{30}$$

$$= \int p(\mathbf{f}|\mathbf{f}_m)q(\mathbf{f}_m) \log \frac{p(\mathbf{y}|\mathbf{f})\cancel{p(\mathbf{f}|\mathbf{f}_m)}p(\mathbf{f}_m)}{\cancel{p(\mathbf{f}|\mathbf{f}_m)}q(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{31}$$

$$= \int p(\mathbf{f}|\mathbf{f}_m)q(\mathbf{f}_m) \log \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}_m)}{q(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{32}$$

$$= \int q(\mathbf{f}_m) \Big[\underbrace{\int p(\mathbf{f}|\mathbf{f}_m) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f}}_{G(\mathbf{f}_m)} + \log \frac{p(\mathbf{f}_m)}{q(\mathbf{f}_m)} \Big] d\mathbf{f}_m \tag{33}$$

変分近似 (3)

- $G(\mathbf{f}_m)$ を計算すると,

$$G(\mathbf{f}_m) = \int p(\mathbf{f}|\mathbf{f}_m) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} \tag{34}$$

$$= \int p(\mathbf{f}|\mathbf{f}_m) \left(-\frac{N}{2} \log(2\pi\sigma^2) - \frac{(\mathbf{y}-\mathbf{f})^2}{2\sigma^2} \right) d\mathbf{f} \tag{35}$$

$$= \int p(\mathbf{f}|\mathbf{f}_m) \left[-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \mathrm{tr}(y^T y - 2y^T \mathbf{f} + \mathbf{f}^T \mathbf{f}) \right] d\mathbf{f} \tag{36}$$

$$= -\frac{N}{2} \log(2\pi\sigma^2)$$

$$- \frac{1}{2\sigma^2} \left[y^T y - 2y^T \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{\alpha} + \mathrm{tr}\left(K_{nn} - K_{nm} K_{mm}^{-1} K_{mn} \right) \right]$$

$$\left(\boldsymbol{\alpha} = E[\mathbf{f}|\mathbf{f}_m] = K_{nm} K_{mm}^{-1} \mathbf{f}_m \right) \tag{37}$$

$$= \log \mathrm{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2 I) - \frac{1}{2\sigma^2} \mathrm{tr}\left(K_{nn} - K'_{nn} \right) . \tag{38}$$

変分近似 (4)

- よって,

$$\log p(\mathbf{y}) \geq \int q(\mathbf{f}_m) \left[G(\mathbf{f}_m) + \log \frac{p(\mathbf{f}_m)}{q(\mathbf{f}_m)} \right] d\mathbf{f}_m \tag{39}$$

$$= \int q(\mathbf{f}_m) \left[\log \mathrm{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2 I) - \frac{1}{2\sigma^2} \mathrm{tr}\left(K_{nn} - K'_{nn}\right) + \log \frac{p(\mathbf{f}_m)}{q(\mathbf{f}_m)} \right] d\mathbf{f}_m \tag{40}$$

$$= \int q(\mathbf{f}_m) \left[\log \frac{\mathrm{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2 I)}{q(\mathbf{f}_m)} + \log p(\mathbf{f}_m) \right] d\mathbf{f}_m - \frac{1}{2\sigma^2} \mathrm{tr}(K_{nn} - K'_{nn}) \tag{41}$$

ここで Jensen bound を逆に使って,

$$\int p(x) \log \frac{f(x)}{p(x)} dx \leq \log \int f(x) dx \tag{42}$$

変分近似 (5)

- より,

$$\text{下限} \leq \log \int \mathrm{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2 I) p(\mathbf{f}_m) d\mathbf{f}_m - \frac{1}{2\sigma^2}\mathrm{tr}(K_{nn} - K'_{nn})$$

$$(K'_{nn} = K_{nm}K_{mm}^{-1}K_{mn}) \quad (43)$$

- $\boldsymbol{\alpha} = E[\mathbf{f}|\mathbf{f}_m] = K_{nm}K_{mm}^{-1}\mathbf{f}_m$ を思い出すと, ガウス積分の公式から

$$\int \mathrm{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2 I) p(\mathbf{f}_m) d\mathbf{f}_m = \mathrm{N}(\mathbf{y}|\mathbf{0}, \sigma^2 I + K'_{nn}) \quad (44)$$

よって, 下限は

$$\log p(\mathbf{y}) \geq \log \mathrm{N}(\mathbf{y}|\mathbf{0}, \sigma^2 I + K'_{nn}) - \frac{1}{2\sigma^2}\mathrm{tr}(K_{nn} - K'_{nn}) \,. \quad (45)$$

変分近似 (6)

- 変分下限の意味

$$\log \mathrm{N}(\mathbf{y}|\mathbf{0}, \sigma^2 I + K'_{nn}) - \frac{1}{2\sigma^2}\mathrm{tr}(K_{nn} - K'_{nn})$$
$$= \log \mathrm{N}(\mathbf{y}|\mathbf{0}, \sigma^2 I + K'_{nn}) - \frac{1}{2\sigma^2}\mathrm{tr}(\mathrm{Cov}(\mathbf{f}|\mathbf{f}_m)) \qquad (46)$$

  - 第 1 項: $\mathbf{f}_m$ によるデータへのフィット
  - 第 2 項: $\mathbf{f}_m$ による真のカーネル行列 K_{nn} の近似度合い
- ナイーブな方法では, 第 1 項のみを最適化している.

GP による分類と SVM

- $y = \{+1, -1\}$ のとき, $p(y|f) = \sigma(y \cdot f)$ (logit) or $\Psi(y \cdot f)$ (probit) で分類器になる

$$\begin{aligned} \text{minimize: } & -\log p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}) \\ & = \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \sum_{i=1}^{N} \log p(y_i|f_i) \end{aligned} \tag{47}$$

- ソフトマージン SVM では $K\alpha = \mathbf{f}$ とすると,

$$\mathbf{w} = \sum_i \alpha_i \mathbf{x}_i \;\rightarrow\; |\mathbf{w}|^2 = \alpha^T K \alpha = \mathbf{f}^T K^{-1}\mathbf{f} \;\text{ ゆえ,}$$

$$\begin{aligned} \text{minimize: } & \frac{1}{2}|\mathbf{w}|^2 - C\sum_{i=1}^{N}(1 - y_i f_i)_+ \\ & = \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - C\sum_{i=1}^{N}(1 - y_i f_i)_+ . \end{aligned} \tag{48}$$

  - そっくりだが, SVM は hinge loss なのでスパース.

Loss functions

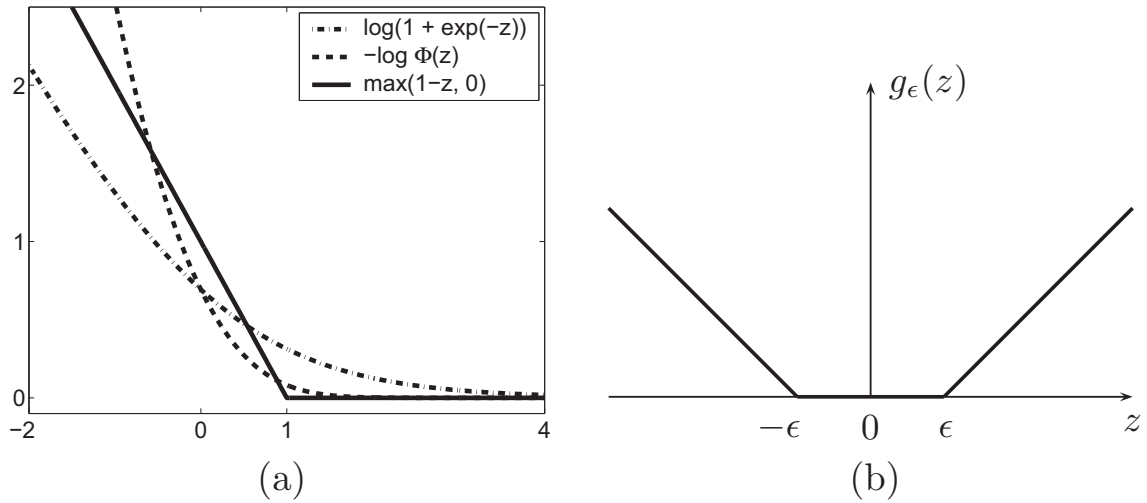
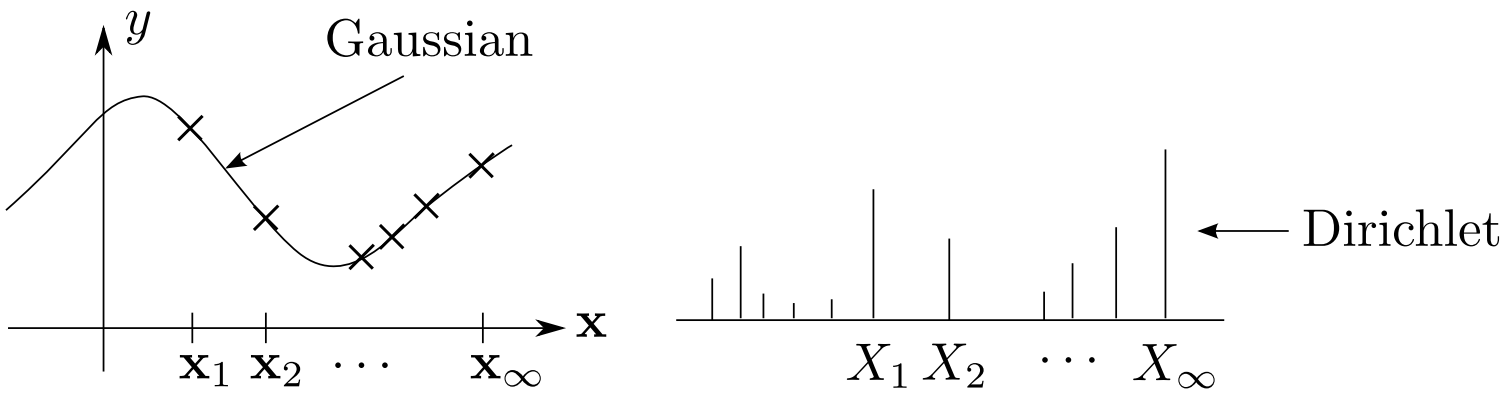


Figure 6.3: (a) A comparison of the hinge error, g_λ and g_Φ. (b) The ϵ-insensitive error function used in SVR.

- カーネル法でお馴染みの議論
  - SVM と ME の関係についても, 同様な関係が成り立つ
- 注意: 分類したいだけなら, GP classifier は回りくどすぎる (他の理由があるなら有効)

DP との関係

- Gaussian process と Dirichlet process の定義はそっくり
  [偶然ではない]
  - **GP:** どんな $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\infty)$ をとってきても, 対応する $(y_1, y_2, \cdots, y_\infty)$ がガウス分布に従う
  - **DP:** 空間のどんな離散化 $(X_1, X_2, \cdots, X_\infty)$ についても, 対応する離散分布がディリクレ分布 $\mathrm{Dir}(\alpha(X_1), \alpha(X_2), \cdots, \alpha(X_\infty))$ に従う

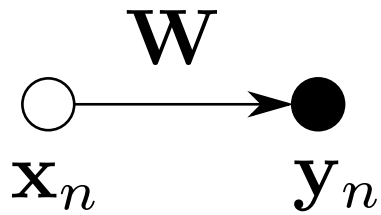


- どちらも, 無限次元の smoother になっている

ガウス過程の教師なし学習

確率的主成分分析

- Probabilistic PCA (Tipping & Bishop 1999)



$$\begin{cases} \mathbf{y}_n = \mathbf{W}\mathbf{x}_n + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \quad \sim \mathrm{N}(0, \sigma^2 I) \end{cases} \tag{49}$$

- よって,

$$L = \log p(\mathbf{y}_n) = \log \mathrm{N}(\mathbf{W}\mathbf{x}_n, \sigma^2 I) \tag{50}$$

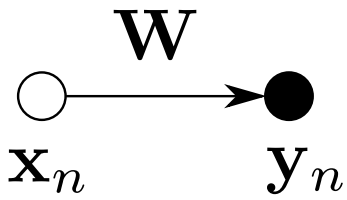
$$= -\frac{N}{2}\left(\log 2\pi + \log |C| + \mathrm{tr}(C^{-1}S)\right) \tag{51}$$

ここで,

$$C = \mathbf{W}\mathbf{W}^T + \sigma^2 I \tag{52}$$

$$S = \frac{1}{N}\mathbf{Y}\mathbf{Y}^T . \tag{53}$$

確率的主成分分析 (2)



- $\dfrac{\partial L}{\partial \mathbf{W}} = 0$ より, 尤度 L を最大にする $\mathbf{W}$ の最尤推定値は

$$\rightarrow \ \hat{\mathbf{W}} \equiv U_q(\Lambda_q - \sigma^2 I)^{\frac{1}{2}} \quad (\sigma^2 = 0 \text{ のとき } U_q \Lambda^{\frac{1}{2}}) \tag{54}$$

  - Λ_q, U_q : $\mathbf{Y}\mathbf{Y}^T$ の最大 q 個の固有値・固有ベクトルを並べた行列

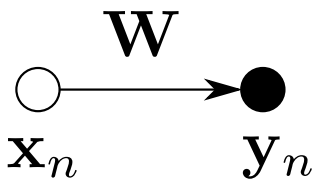
- $\sigma^2 = 0$ で通常の主成分分析と一致

Gaussian Process Latent Variable Models (GPLVM)

- Probabilistic PCA (Tipping&Bishop 1999):

$$p(\mathbf{y}_n|\mathbf{W},\beta) = \int p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{W},\beta)p(\mathbf{x}_n)d\mathbf{x}_n \quad (55)$$

$$p(\mathbf{Y}|\mathbf{W},\beta) = \prod_n p(\mathbf{y}_n|\mathbf{W},\beta) \quad \rightarrow \mathbf{W} \text{ を最適化}$$



- GPLVM (Lawrence, NIPS 2003): $\mathbf{W}$ の方に prior を与えて積分消去

$$p(\mathbf{W}) = \prod_{i=1}^{D} \mathrm{N}(\mathbf{w}_i|0,\alpha^{-1}\mathbf{I}) \quad (56)$$

$$p(\mathbf{Y}|\mathbf{X},\beta) = \int p(\mathbf{Y}|\mathbf{X},\beta)p(\mathbf{W})d\mathbf{W} \quad (57)$$

$$= \frac{1}{(2\pi)^{DN/2}|K|^{D/2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)\right) \quad (58)$$

GPLVM (2): PPCA の Dual

$$\log p(\mathbf{Y}|\mathbf{X},\beta) = -\frac{DN}{2}\log(2\pi) - \frac{D}{2}\log|K| - \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T) \tag{59}$$

$$\mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I} \tag{60}$$

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \tag{61}$$

- $\mathbf{X}$ に関して微分すると,

$$\frac{\partial L}{\partial \mathbf{X}} = \alpha\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{X} - \alpha D\mathbf{K}^{-1}\mathbf{X} = 0 \tag{62}$$

$$\Longleftrightarrow \quad \mathbf{X} = \frac{1}{D}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1}\mathbf{X} \quad \Rightarrow \quad \mathbf{X} \simeq U_Q L V^T \tag{63}$$

  - U_Q $(N \times Q)$: $\mathbf{Y}\mathbf{Y}^T$ の Q 個の最大固有値 $\lambda_1 \cdots \lambda_Q$ に対応する固有ベクトル
  - $L = \mathrm{diag}(l_1, \cdots, l_Q);\ l_i = 1/\sqrt{\frac{\lambda_i}{\alpha D} - \frac{1}{\alpha\beta}}$

GPLVM (3) : Kernel 化

$$\log p(\mathbf{Y}|\mathbf{X},\beta) = -\frac{DN}{2}\log(2\pi) - \frac{D}{2}\log|K| - \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T)$$

$$\mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}, \tag{64}$$

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T \tag{65}$$

- 自然にカーネル化されている $\Longrightarrow$ 任意のカーネル $\mathbf{K}$ を導入

$$k(\mathbf{x}_n, \mathbf{x}_m) = \alpha\exp\left(-\frac{\gamma}{2}(\mathbf{x}_n - \mathbf{x}_m)^2\right) + \delta(n,m)\beta^{-1} \tag{66}$$

- $\dfrac{\partial L}{\partial \mathbf{K}} = \mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-1} - D\mathbf{K}^{-1}$
  - $\dfrac{\partial L}{\partial x_{n,j}} = \dfrac{\partial L}{\partial \mathbf{K}}\dfrac{\partial \mathbf{K}}{\partial x_{n,j}}$ を適用して微分
  - Scaled Conjugate Gradient で解ける

    GPLVM in MATLAB: http://www.cs.man.ac.uk/~neill/gplvm/

GPLVM (4): 例

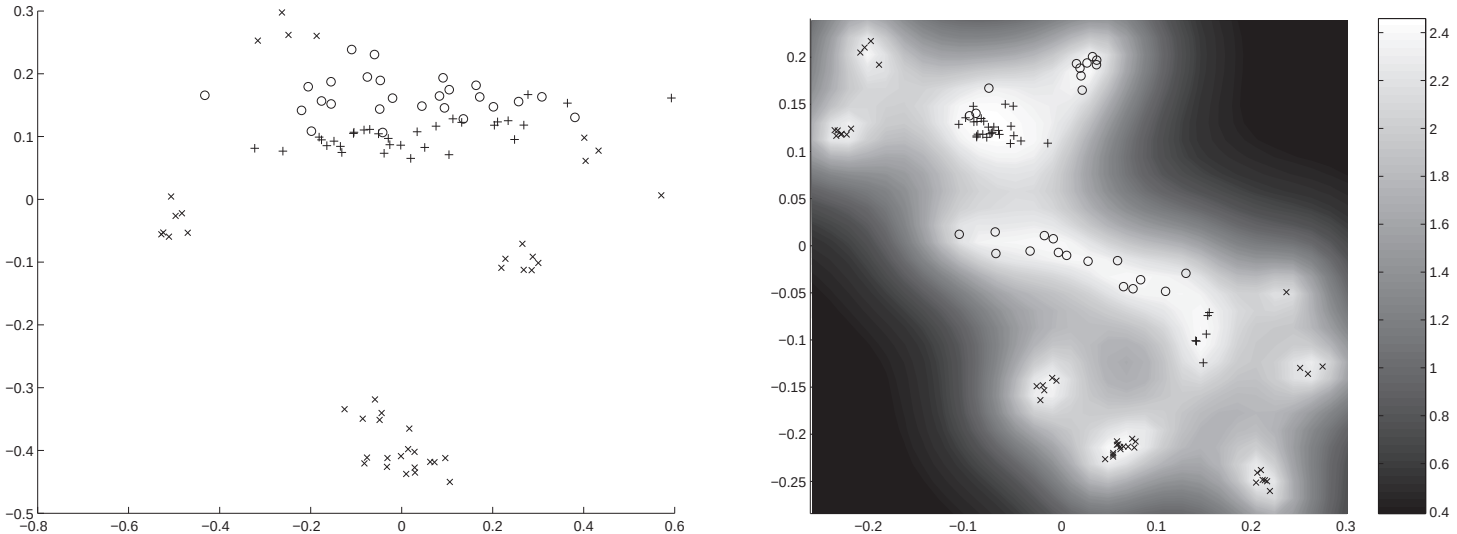
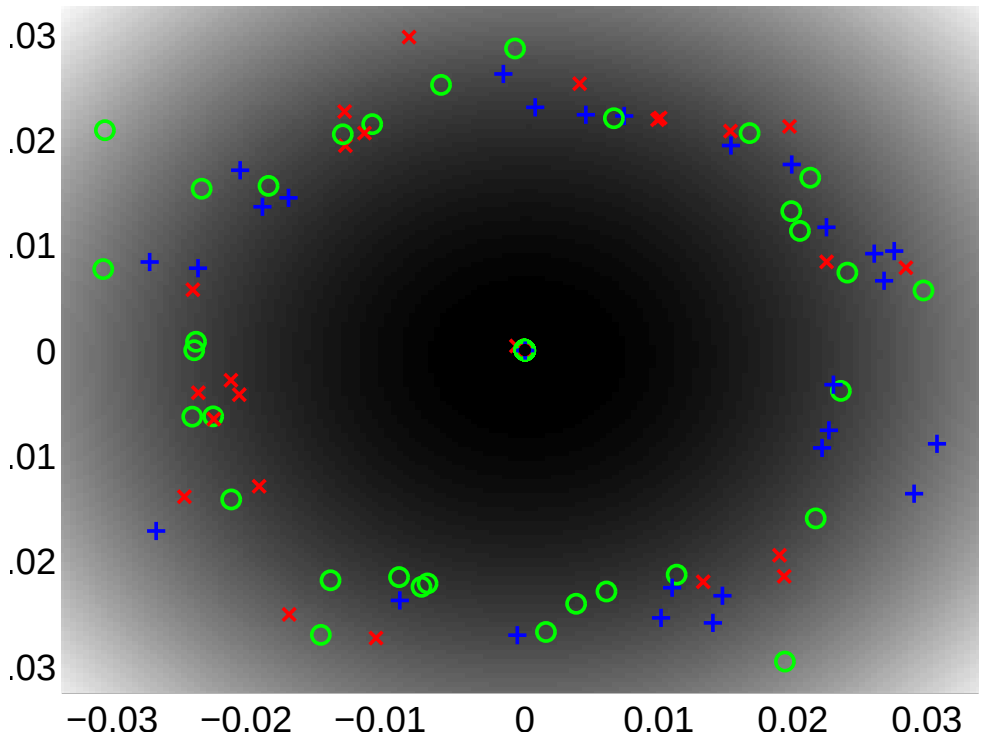


Figure 1: Visualisation of the Oil data with (a) PCA (a linear GPLVM) and (b) A GPLVM which uses an RBF kernel. Crosses, circles and plus signs represent stratifi ed, annular and homogeneous fbws respectively. The greyscales in plot (b) indicate the precision with which the manifold was expressed in data-space for that latent point. The optimised parameters of the kernel were $\gamma = 150$, $\alpha = 0.403$ and $\beta = 316$.

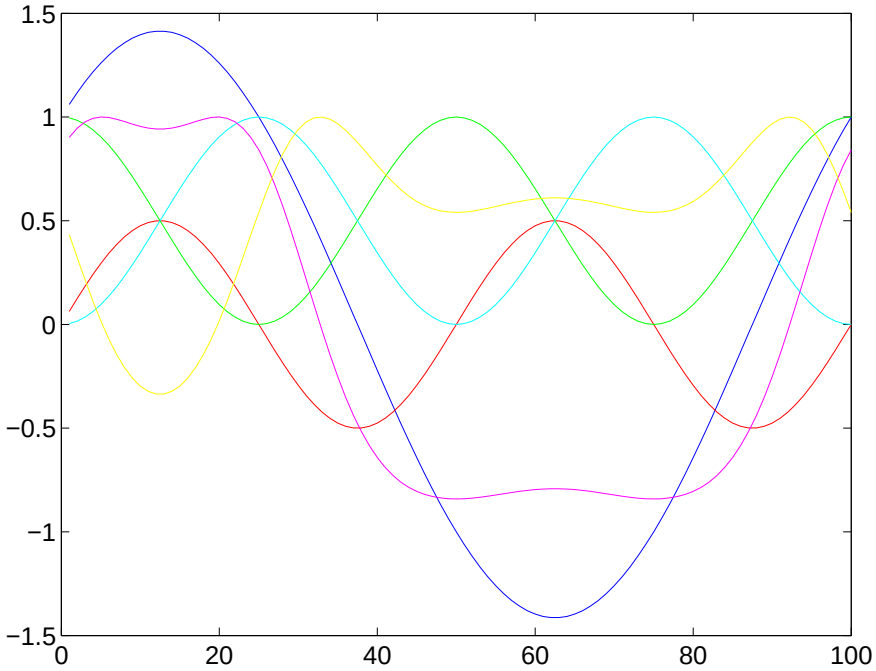
- 線形の PPCA(左) より, GP-LVM(右) の方が分離性能が高い
  - ベイズなので, Confidence の分布が同時に得られる
- 計算量が問題 ($O(N^3)$): active set (サポートベクターみたいなもの) を選んで, データをスパース化して最適化

GPLVM (4): Caveat

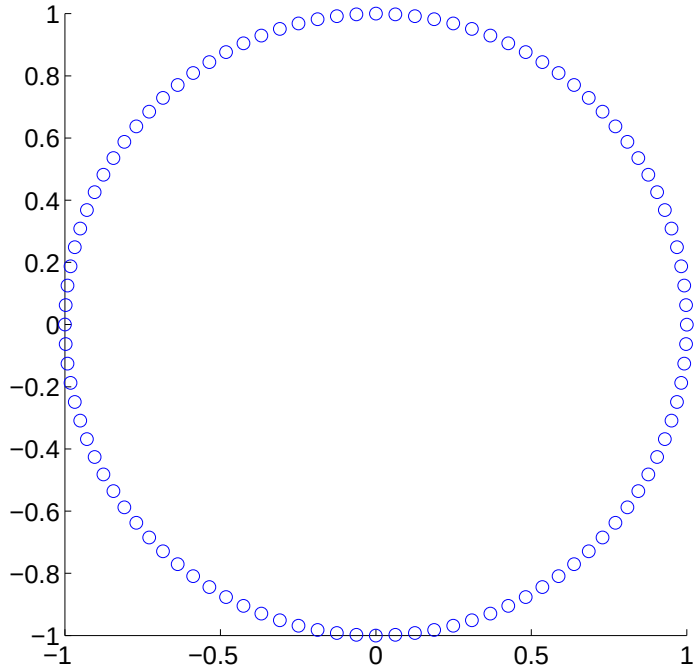
- PCA ではなく, ランダムに初期化した場合の結果
  - Neil Lawrence のコードで, `1e-2*randn(N,dims)` で初期化
  - Scaled conjugate gradient で最適化



GPLVM (5): 時系列データ

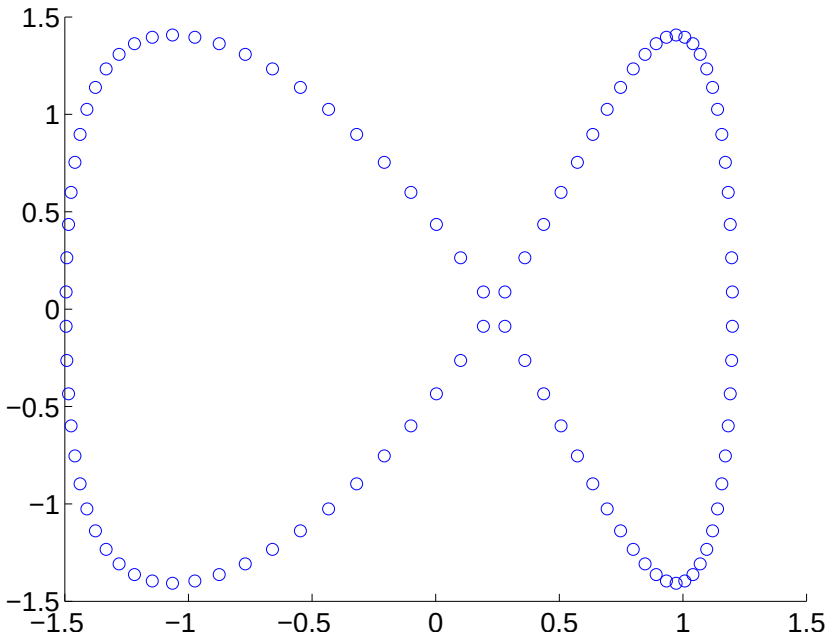


入力



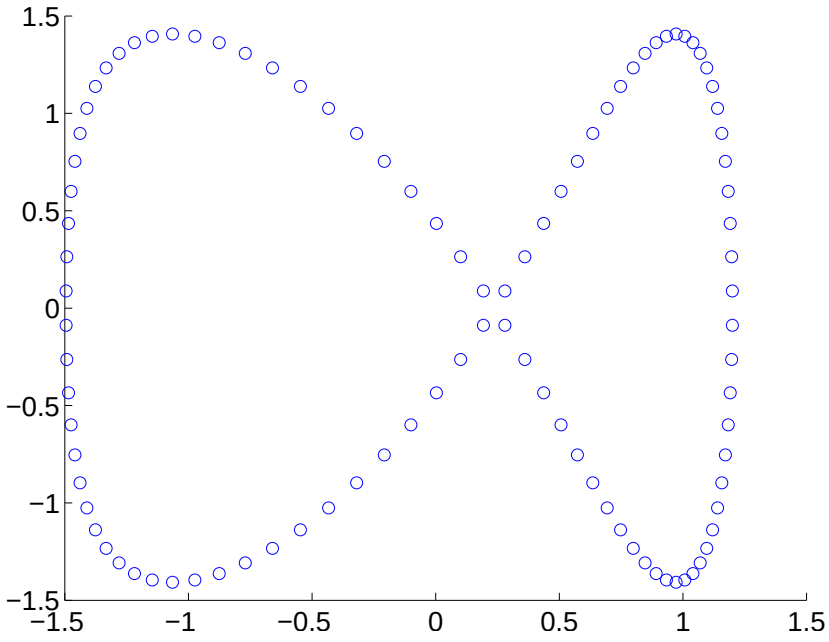
正解

GPLVM (6): 最適化結果

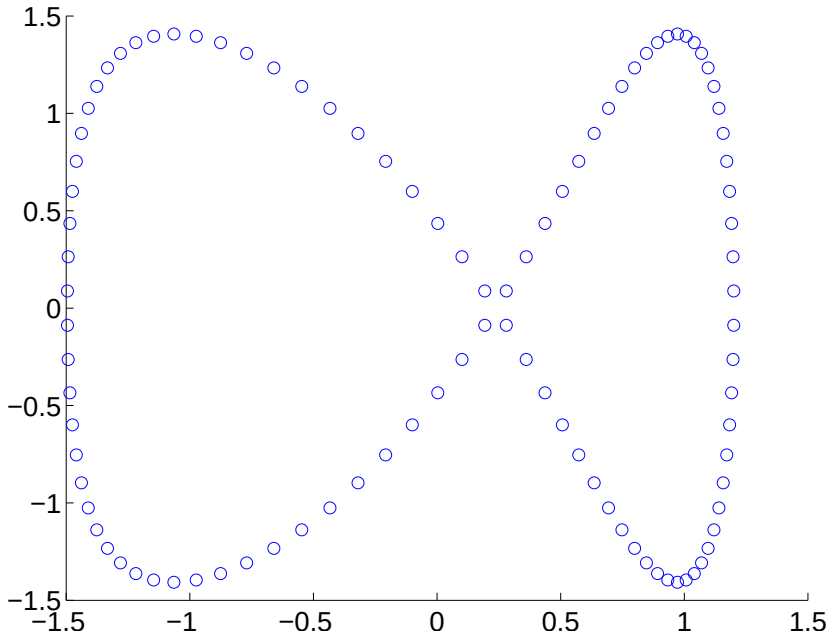


計算結果

GPLVM (6): 最適化結果

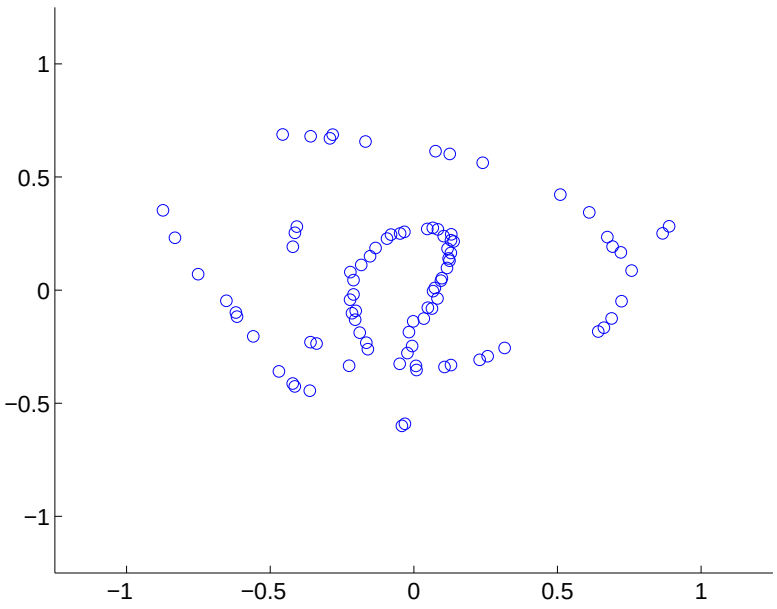


計算結果

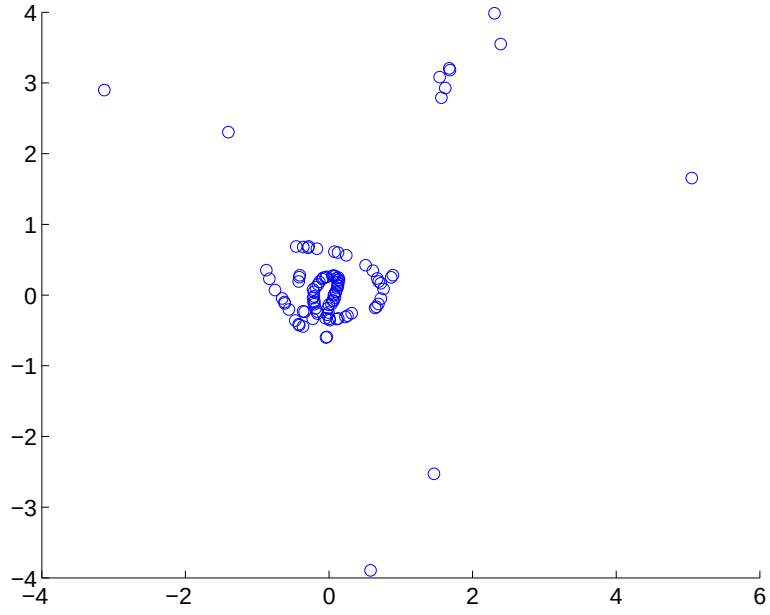


PCA による初期化

GPLVM (7): MCMC による計算



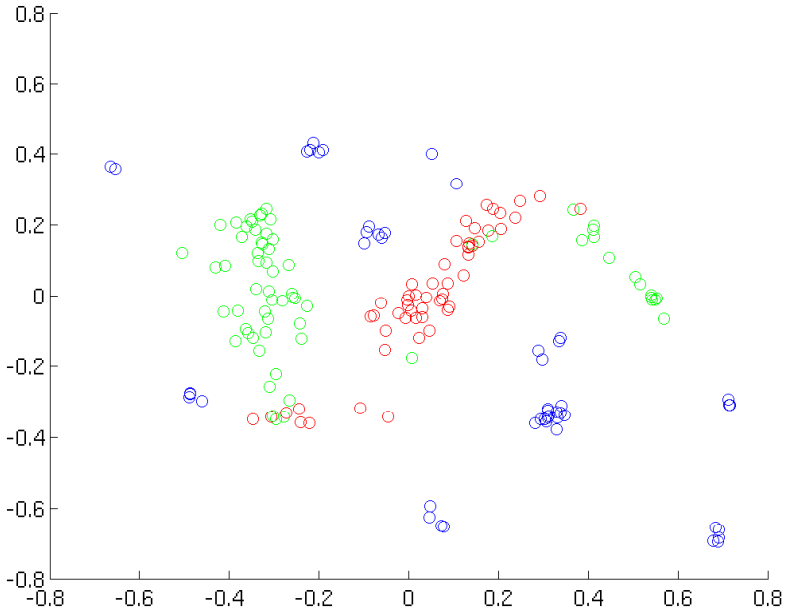
Local



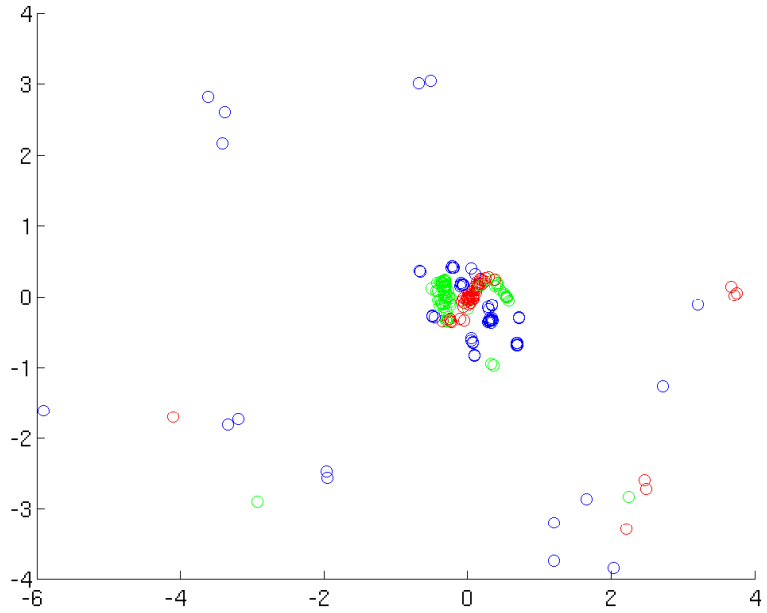
Global

- MCMC で注意深く最適化 (ステップ=0.2, 400 iteration)
- 点はすべて 0 で初期化
- 点を「繋ぐ」ような制約はない ($\rightarrow$ GPDM)

GPLVM (8): MCMC による計算 (Oil Flow)



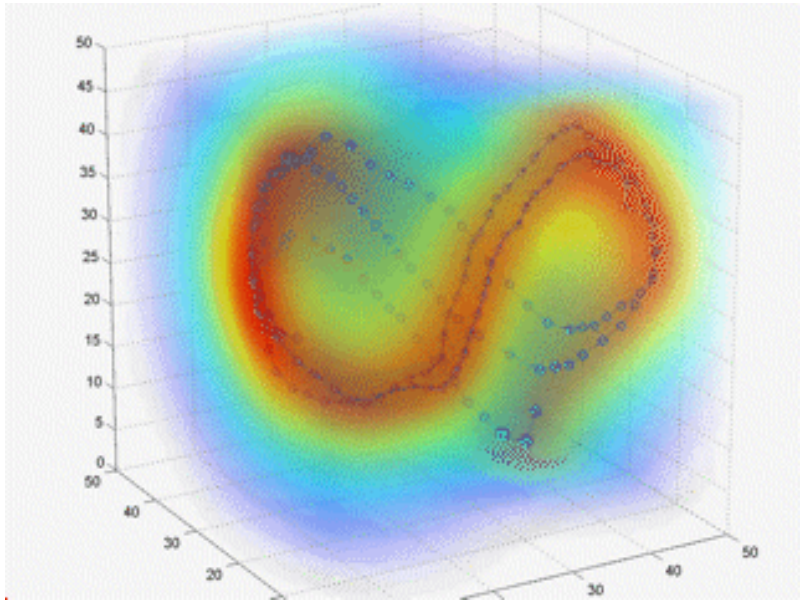
Local



Global

- MCMC で計算すると, X にクラスタが現れる
- ただし, X に外れ値が発生

Gaussian Process Dynamical Model (Hertzmann 2005)



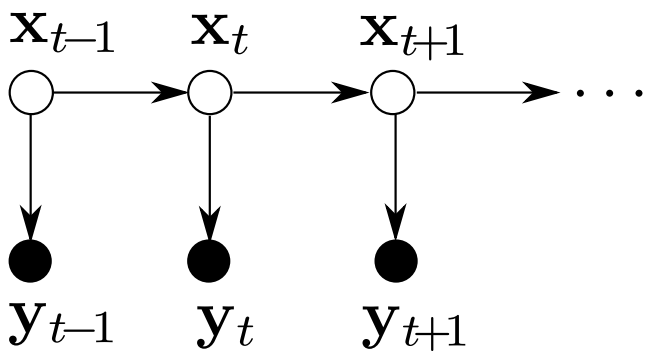
http://www.dgp.toronto.edu/~jmwang/gpdm/

- GPLVM では, 潜在変数 $\mathbf{x}_n$ に分布がなかった

  $\Downarrow$

- $\mathbf{x}_n$ が (GP で) 時間発展するモデル.
  - 人間の動き (角度ベクトル) 等の時系列データ.
  - 自然言語の時系列データ?

GPDM (2): Formulation (1)



$$\begin{cases} \mathbf{x}_t &= f(\mathbf{x}_{t-1};\mathbf{A}) + \epsilon_{x,t} \\ \mathbf{y}_t &= g(\mathbf{x}_t;\mathbf{B}) + \epsilon_{y,t} \end{cases}, \quad \begin{aligned} f &\sim \mathrm{GP}(0,\mathbf{K}_\mathbf{x}) \quad (67) \\ g &\sim \mathrm{GP}(0,\mathbf{K}_\mathbf{y}) \quad (68) \end{aligned}$$

- 結合確率 $p(\mathbf{Y},\mathbf{X}|\alpha,\beta) = p(\mathbf{Y}|\mathbf{X},\beta)p(\mathbf{X}|\alpha)$ を考える.
- 第 1 項

$$p(\mathbf{Y}|\mathbf{X},\beta) = \frac{|\mathbf{W}|^N}{(2\pi)^{ND/2}|\mathbf{K}_Y|^{D/2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T)\right) \quad (69)$$

  は GPLVM と基本的に同じ.
  - $\mathbf{K}_Y$ は (この論文では) 普通の RBF カーネル

GPDM (3): Formulation (2)

- 第 2 項は Markov 時系列

$$p(\mathbf{X}|\alpha) = p(\mathbf{x}_1) \int \prod_{t=2}^{N} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{A}, \alpha) \underbrace{p(\mathbf{A}|\alpha)}_{\text{Gaussian}} d\mathbf{A} \tag{70}$$

$$= p(\mathbf{x}_1) \frac{1}{(2\pi)^{d(N-1)/2} |\mathbf{K}_X|^d} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_X^{-1}\mathbf{X}_-\mathbf{X}_-^T)\right) \tag{71}$$

  - $\mathbf{X}_- = [\mathbf{x}_2, \cdots, \mathbf{x}_N]^T$ とおいた
  - $\mathbf{K}_X$ は $\mathbf{x}_1 \cdots \mathbf{x}_{N-1}$ の RBF+線形カーネル

$$k(\mathbf{x}, \mathbf{x}') = \alpha_1 \exp\left(-\frac{\alpha_2}{2}||\mathbf{x} - \mathbf{x}'||^2\right) + \alpha_3 \mathbf{x}^T\mathbf{x} + \alpha_4^{-1}\delta(\mathbf{x}, \mathbf{x}') \,. \tag{72}$$

GPDM (4): Formulation(3)

$$p(\mathbf{Y}, \mathbf{X}, \alpha, \beta) = p(\mathbf{Y}|\mathbf{X}, \beta)p(\mathbf{X}|\alpha)p(\alpha)p(\beta) \tag{73}$$

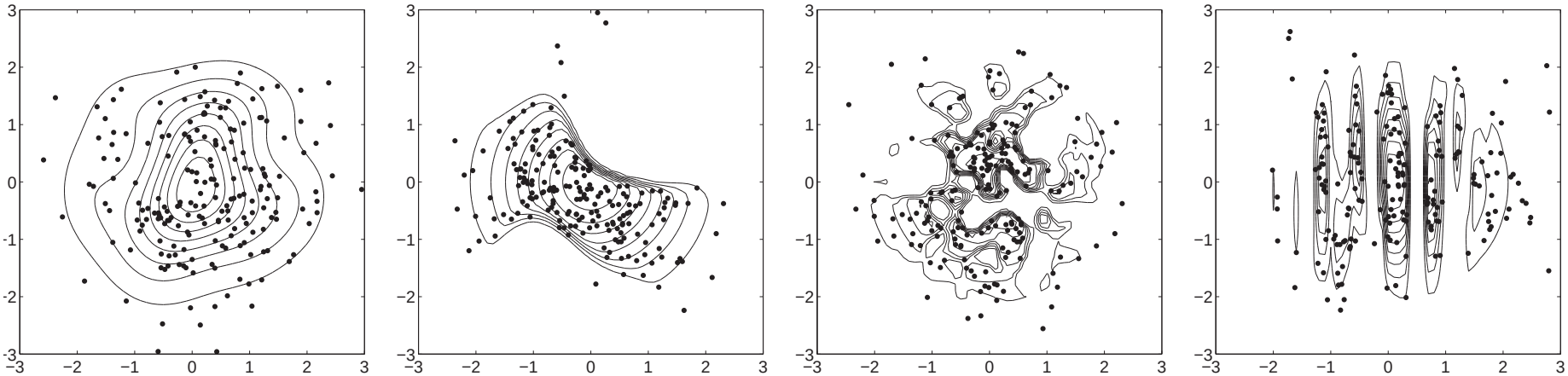
$$p(\alpha) \propto \prod_i \alpha_i^{-1},\ p(\beta) \propto \prod_i \beta_i^{-1}. \tag{74}$$

- 対数尤度は

$$\begin{aligned} -\log p(\mathbf{Y}, \mathbf{X}, \alpha, \beta) =& \frac{1}{2}\mathrm{tr}(\mathbf{K}_X^{-1}\mathbf{X}_-\mathbf{X}_-^T) + \frac{1}{2}\mathrm{tr}(\mathbf{K}_Y^{-1}\mathbf{Y}\mathbf{W}^2\mathbf{Y}^T) \\ & + \frac{d}{2}\log|\mathbf{K}_X| + \frac{D}{2}\log|\mathbf{K}_Y| \quad \text{(正則化項)} \\ & \underbrace{-\log|\mathbf{W}| + \sum_j \log\alpha_j + \sum_j \log\beta_j}_{\text{(定数)}} \end{aligned} \tag{75}$$

$$\longrightarrow \text{最小化.} \tag{76}$$

Gaussian Process Density Sampler (1)



(a) $\ell_x = 1, \ell_y = 1, \alpha = 1$ (b) $\ell_x = 1, \ell_y = 1, \alpha = 10$ (c) $\ell_x = 0.2, \ell_y = 0.2, \alpha = 5$ (d) $\ell_x = 0.1, \ell_y = 2, \alpha = 5$

- GP は任意の関数の prior →確率密度関数のモデルに使えないか?

$$p(\mathbf{x}) = \frac{1}{Z(f)} \Phi(f(\mathbf{x}))\pi(\mathbf{x}) \tag{77}$$

  - $f(\mathbf{x}) \sim \mathrm{GP}(\mathbf{x})$ ； $\pi(\mathbf{x})$：事前分布
  - $\Phi(\mathbf{x}) \in [0, 1]$：シグモイド関数
    - ex. $\Phi(x) = 1/(1 + \exp(-x))$

Gaussian Process Density Sampler (2)

$$p(\mathbf{x}) = \frac{1}{Z(f)} \Phi(f(\mathbf{x}))\pi(\mathbf{x}) \tag{78}$$

- 生成プロセス: Rejection sampling
  1. Draw $\mathbf{x} \sim \pi(\mathbf{x})$.
  2. Draw $r \sim \mathrm{Uniform}[0, 1]$.
  3. If $r < \Phi(g(\mathbf{x}))$ then accept $\mathbf{x}$; else reject $\mathbf{x}$
- Accept された N 個の観測データの背後に, reject された M 個のデータとその場所が存在 (隠れ変数)
  - 分配関数 $Z(f)$ は求まらないが, $\Phi(g(\mathbf{x}))$ は求まる
    $\rightarrow$ MCMC!
  - Infinite Mixture とは別の確率密度のモデル化

まとめ

- Gaussian process…連続的な関数のベイズモデル
  - カーネル関数で定義される, 無限次元のガウス分布
  - 基底関数の空間での線形モデルで, 重みを積分消去したもの
- カーネル設計が重要
  - カーネルの違いにより, さまざまな振舞い
  - カーネルのパラメータは, 確率モデルなのでデータから最適化できる
- 回帰問題だけでなく, 教師なし学習にも使える (GPLVM, GPDM)
- 計算量が課題→変分近似による効率的計算

Literature

- “Gaussian Process Dynamical Models”. J. Wang, D. Fleet, and A. Hertzmann. NIPS 2005. http://www.dgp.toronto.edu/ jmwang/gpdm/
- “Gaussian Process Latent Variable Models for Visualization of High Dimensional Data”. Neil D. Lawrence, NIPS 2003.
- “The Gaussian Process Density Sampler”. Ryan Prescott Adams, Iain Murray and David MacKay. NIPS 2008.
- “Archipelago: Nonparametric Bayesian Semi-Supervised Learning”. Ryan Prescott Adams and Zoubin Ghahramani. ICML 2009.

参考文献・教科書

- 「パターン認識と機械学習」(*Pattern Recognition and Machine Learning*), Chapter 6. Christopher Bishop, Springer, 2006. http://ibisforest.org/index.php?PRML
- "*Gaussian Processes for Machine Learning*". Rasmussen and Williams, MIT Press, 2006. http://www.gaussianprocess.org/gpml/
- "Gaussian Processes — A Replacement for Supervised Neural Networks?". David MacKay, Lecture notes at NIPS 1997. http://www.inference.phy.cam.ac.uk/mackay/GP/
  - Videolectures.net: "*Gaussian Process Basics*". http://videolectures.net/gpip06_mackay_gpb/
- ガウス過程に関するメモ (1). 正田備也, 2007. http://www.iris.dti.ne.jp/~tmasada/2007071101.pdf

Codes

- GPML Toolbox (in MATLAB): http://www.gaussianprocess.org/gpml/code/
- GPy (in Python): http://sheffieldml.github.io/GPy/