

人工知能の数理

- なぜ開催するのか

東京には人工知能(機械学習) の勉強会
が沢山あるが,
札幌には少ないし, さらにあってもあ
まりレベルの高いものではない.



自己紹介



- 11月から機械学習エンジニア
- 趣味はカメラとチェロ
- 数学が好き. (勉強中)

Before

情報幾何の応用と 最近の機械学習の動向

大下範晃

発表の概要

- ・機械学習に使われることのある数学
- ・情報幾何とその応用事例の説明
- ・カーネル法とその応用事例の説明
- ・ガウス過程の基礎とその応用事例の説明

※私は数学の初心者な為、間違いや指摘内容があれば、発表途中でもご指摘ください。

(使われることがあるか)

機械学習にどれくらい数学が必要か

機械学習の理論を突き詰めたい人
どういう原理で動いているのか知りたい

機械学習を仕事or趣味で使えればよい

微分積分、線型代数、
関数解析、測度論、
確率論、統計学、
集合、位相、代数学、
微分幾何(情報幾何)、
数論, etc



最低限
微分積分と線型代数

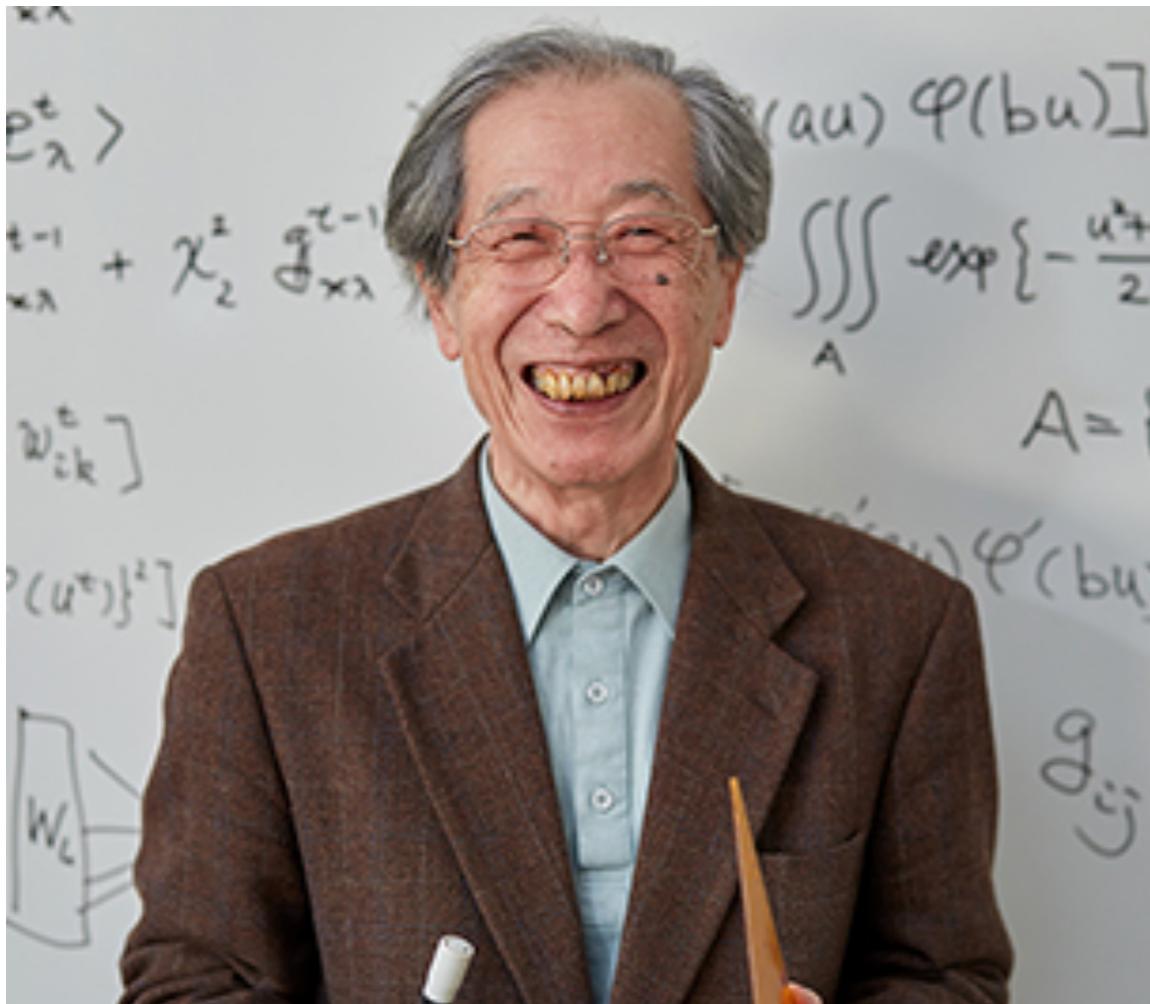
全部理解する必要はない。

(使われることがあるか)

機械学習にどれくらい数学が必要か



情報幾何とは

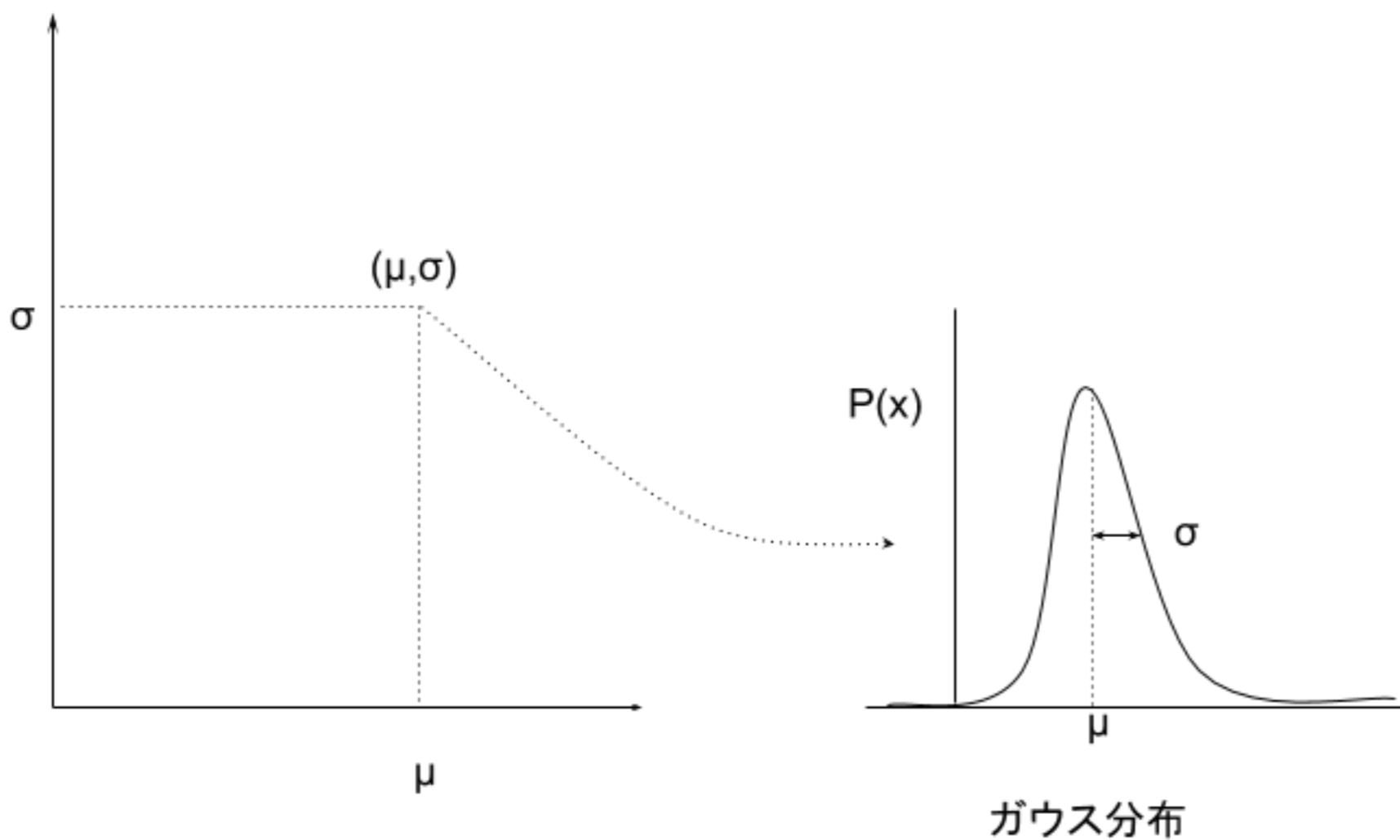


- 情報幾何の創設者
- 骨組みは修士と博士で考えたそうです。

甘利俊一先生

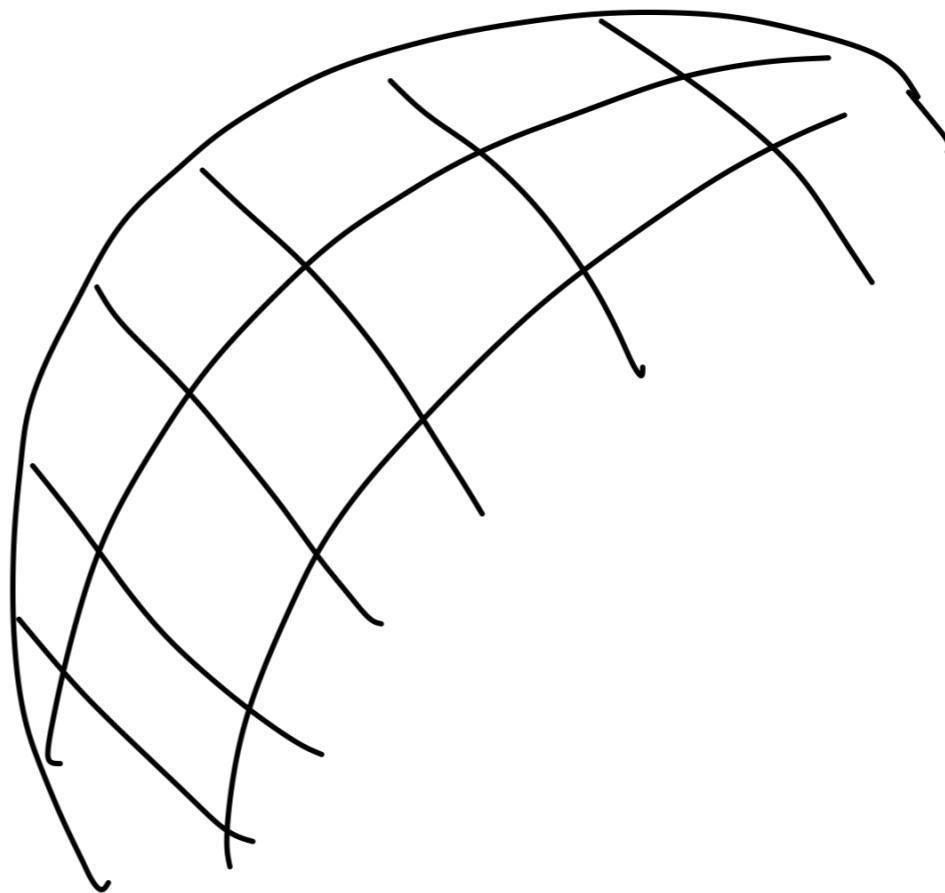
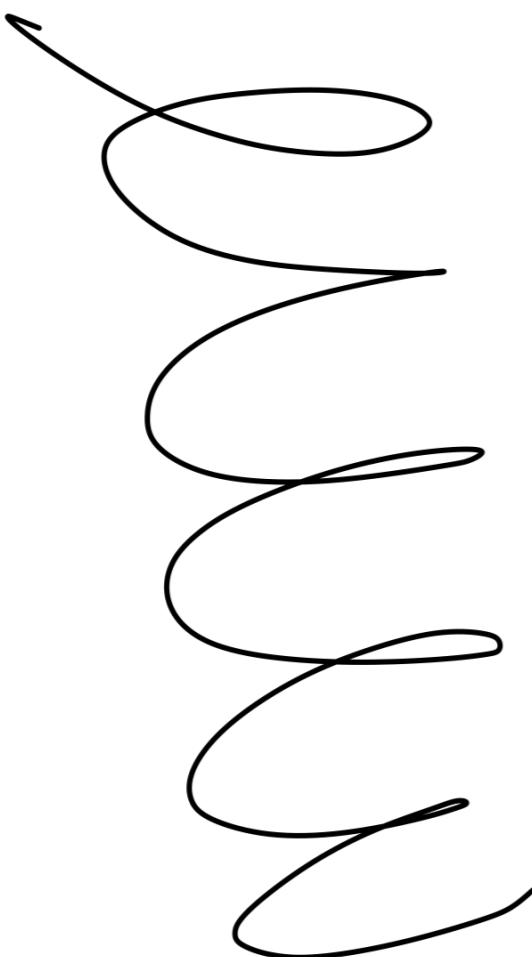
情報幾何とは

- ・ 双対アファイン接続の微分幾何
- ・ パラメータ空間の幾何学 (\neq データ空間の幾何学)



微分幾何とは

- ひとことで言えば微分を用いた幾何.



曲率・捩率（れいりつ）

- 曲率

$p(s) = (x(s), y(s), z(s)) (a \leq s \leq b)$ を3次元ユークリッド空間中の曲線とする。 (また曲線P(s)によって表される運動の速さは一定で1になるようにパラメータをとってある。)

すなわち速度ベクトル $e_1(s) = p'(s) = (x'(s), y'(s), z'(s))$ が長さ1であるとする。具体的には

$$e_1(s) \cdot e_1(s) = x'(s)^2 + y'(s)^2 + z'(s)^2 = 1$$

となっているとする。

曲率・捩率（れいりつ）

そのとき加速度ベクトル $e'_1(s)$ を考えてみると

$$0 = \frac{d(e_1(s) \cdot e_1(s))}{ds} = 2e'_1(s) \cdot e_1(s)$$

であるから、 $e'_1(s)$ が $e_1(s)$ に直交している。 $e'_1(s)$ の長さを $k(s)$ と書き、曲線 $P(s)$ の曲率と呼ぶ。すなわち

$$\begin{aligned}\kappa(s) &= \sqrt{e'_1(s) \cdot e'_1(s)} \\ &= \sqrt{x''(s)^2 + y''(s)^2 + z''(s)^2}.\end{aligned}$$

曲率・捩率（れいりつ）

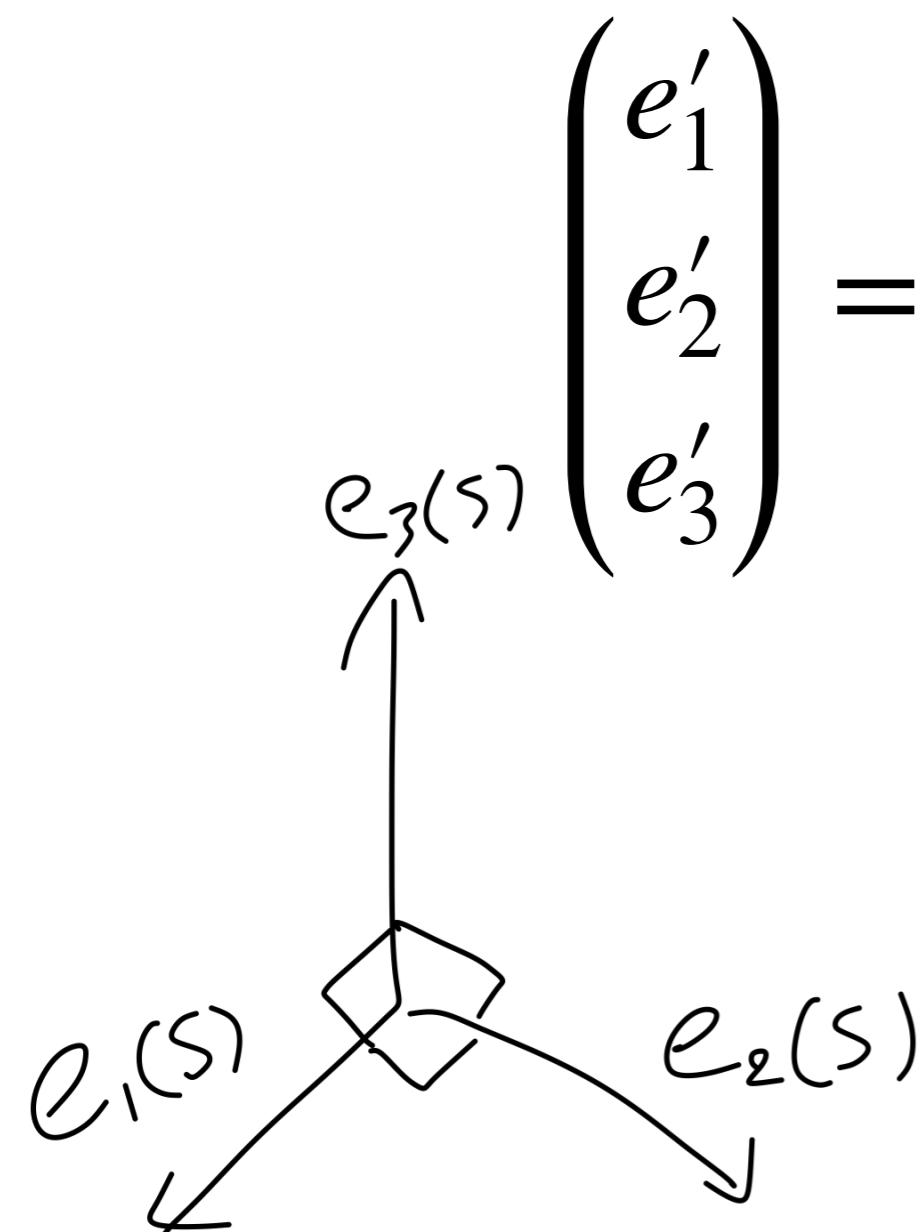
- 捩率

$$e_3'(s) \begin{pmatrix} e'_1 \\ e'_2 \\ e'_3 \end{pmatrix} = \begin{pmatrix} 0 & k & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

e₁(s) *e₂(s)* *e₃(s)*

捩率といふ

導出は省略します。



双対アファイン接続

アファイン接続 ∇ を持つ Riemann 多様体 (M, g) において

$$X_g(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X^* Z) \quad (X, Y, Z) \in \mathcal{X}(M)$$

で定義されるアファイン接続 ∇^* を計量 g に関する ∇ の双対アファイン接続という。



情報幾何の応用

- ・自然勾配法
- ・サポートベクトルマシン
- ・Boosting
- ・主成分分析
- ・
- ・
- ・
- ・など、ありとあらゆるものに応用されている
- ・詳しくは情報幾何学の新展開



情報幾何の応用

- Bayesian shrinkage prediction for the regression problem (正規分布のベイズ予測における事前分布の構成)

[https://www.sciencedirect.com/science/article/pii/
S0047259X08000365](https://www.sciencedirect.com/science/article/pii/S0047259X08000365)

- Statistical Inference with Unnormalized Discrete Models and Localized Homogeneous Divergences(非正規化モデルの推定理論)

<http://jmlr.org/papers/v18/15-596.htm>

正規分布のベイズ予測における事前分布の構成

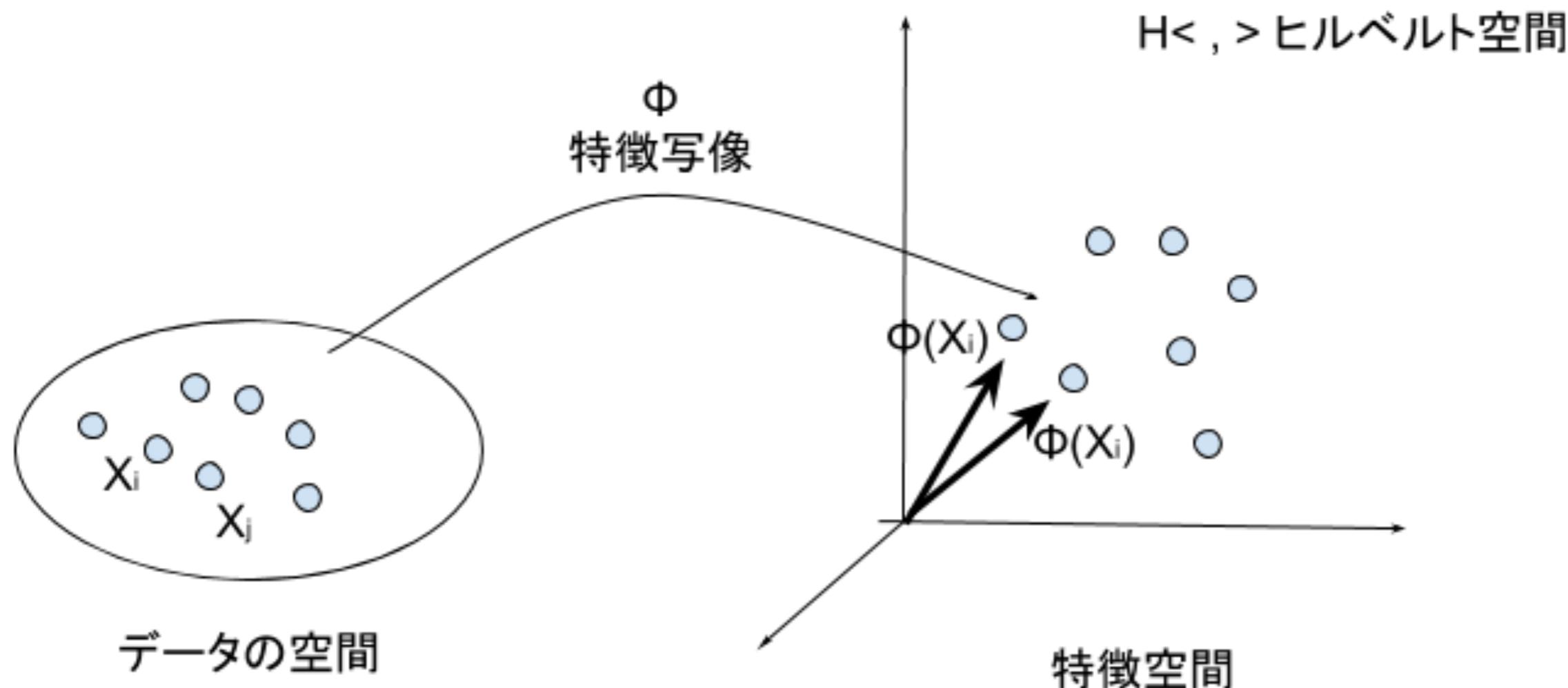
- ・この論文の基礎

今以下の多変量正規分布が観測されるとする.

$$y \sim N_d(y; \mu, \Sigma)$$

N_d は平均 μ , 共分散 Σ からなる,
d 次元の多変量正規分布の密度関数である.

カーネル法の簡単な説明



※実際は $\langle \Phi(X_i), \Phi(X_j) \rangle$ がカーネル関数 $k(X_i, X_j)$ によって計算される。
(グラム行列を用いる。)

カーネル法の重要な定理

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad (\forall x \in \mathcal{X}, \forall f \in \mathcal{H})$$

再生性：ヒルベルト空間上の任意の f と
カーネル関数 $k(\cdot, X)$ の内積によって,
 $f(x)$ が再生される

(Moore Aronszjnの定理)

正定値カーネル

2. 1. 1 正定値カーネルの定義と基本的性質

まず、実数値の正定値カーネルから定義する。

\mathcal{X} を集合とするとき、次の2条件を満たすカーネル $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ を (\mathcal{X} 上の) 正定値カーネル(positive define kernel)という。

- 対称性：任意の $x, y \in \mathcal{X}$ に対し $k(x, y) = k(y, x)$
- 正値性：任意の $n \in \mathbb{N}, x_1, \dots, x_n \in \mathcal{X}, c_1, \dots, c_n \in \mathbb{R}$

に対し

$$\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0 .$$

グラム行列

対称性のもと、正値性の条件は対称行列

$$(k(x_i, x_j))_{i,j=1}^n = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

が半正定値であることを意味する。

この対称行列をグラム行列という。

正定値・半正定値

$$z^T M z = [z_1, z_2, \dots, z_n] \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ \vdots & \ddots & & \\ m_{n1} & m_{21} & \cdots & m_{nn} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

M が正定値とは $z^T M z > 0$ が必ずなりたつとき正定値という。

M が正定値とは $z^T M z \geq 0$ が成り立つとき半正定値という。

$z \equiv$ 非ゼロ列ベクトル

$M \equiv n \times n$ の実数対称行列

A Linear-Time Kernel Goodness-of-Fit Test

- NIPS2017のベストペーパー！！
- 分布の分からぬところを推定する。
- 例えればこれは標準正規分布といって良いのかどうかを判定する
- A Kernelized Stein Discrepancy for Goodness-of-fit Tests (Qiang Liu, Jason D. Lee, Michael I. Jordan)という手法を用いて（後日Qiitaに概要を書く予定）2つの確率分布が似ているかを線形時間で測定する。

<https://arxiv.org/abs/1602.03253>

A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum¹ Wenkai Xu¹ Zoltán Szabó² Kenji Fukumizu³ Arthur Gretton¹

¹Gatsby Unit, University College London

²CMAP, École Polytechnique

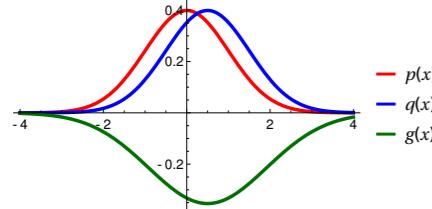
³The Institute of Statistical Mathematics

Summary

- Given: $\{\mathbf{x}_i\}_{i=1}^n \sim q$ (unknown), and a density p .
- Goal: Test $H_0 : p = q$ vs $H_1 : p \neq q$ quickly.
- New multivariate goodness-of-fit test (FSSD):
 - Nonparametric: arbitrary, unnormalized p . $\mathbf{x} \in \mathbb{R}^d$.
 - Linear-time: $\mathcal{O}(n)$ runtime complexity. Fast.
 - Interpretable: tell where p does not fit the data.

Previous: Kernel Stein Discrepancy (KSD)

- Let $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})} \nabla_{\mathbf{x}} [k(\mathbf{x}, \mathbf{v}) p(\mathbf{x})] \in \mathbb{R}^d$.
- Stein witness function: $g(\mathbf{v}) = \mathbb{E}_{\mathbf{x} \sim q} [\xi(\mathbf{x}, \mathbf{v})]$ where $g = (g_1, \dots, g_d)$ and each $g_i \in \mathcal{F}$, an RKHS associated with kernel k .



Known: Under some conditions, $\|g\|_{\mathcal{F}^d} = 0 \iff p = q$.
[Chwialkowski et al., 2016, Liu et al., 2016]

Statistic: $KSD^2 = \|g\|_{\mathcal{F}^d}^2 = \mathbb{E}_{\mathbf{x} \sim q} \mathbb{E}_{\mathbf{y} \sim q} h_p(\mathbf{x}, \mathbf{y}) \approx \frac{2}{n(n-1)} \sum_{i < j} h_p(\mathbf{x}_i, \mathbf{x}_j)$, where

$$h_p(\mathbf{x}, \mathbf{y}) := [\nabla_{\mathbf{x}} \log p(\mathbf{x})] k(\mathbf{x}, \mathbf{y}) [\nabla_{\mathbf{y}} \log p(\mathbf{y})] + \nabla_{\mathbf{x}} \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) + [\nabla_{\mathbf{y}} \log p(\mathbf{y})] \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) + [\nabla_{\mathbf{x}} \log p(\mathbf{x})] \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}).$$

Characteristics of KSD:

- Nonparametric. Applicable to a wide range of p .
- Do not need the normalizer of p .
- Runtime: $\mathcal{O}(n^2)$. Computationally expensive. ☹

Linear-Time KSD (LKS) Test: [Liu et al., 2016]

$$\|g\|_{\mathcal{F}^d}^2 \approx \frac{2}{n} \sum_{i=1}^{n/2} h_p(\mathbf{x}_{2i-1}, \mathbf{x}_{2i}).$$

Runtime: $\mathcal{O}(n)$. High variance. Low test power. ☹

The Finite Set Stein Discrepancy (FSSD)

Idea: Evaluate witness g at J locations $\{\mathbf{v}_1, \dots, \mathbf{v}_J\}$. Fast.

$$FSSD^2 = \frac{1}{dJ} \sum_{j=1}^J \|g(\mathbf{v}_j)\|_2^2.$$

Proposition (FSSD is a discrepancy measure). Main conditions:

- (Nice kernel) Kernel k is C_0 -universal, and real analytic (Taylor series at any point converges) e.g., Gaussian kernel.
- (Vanishing boundary) $\lim_{\|\mathbf{x}\| \rightarrow \infty} p(\mathbf{x})g(\mathbf{x}) = 0$.
- (Avoid "blind spots") Locations $\{\mathbf{v}_1, \dots, \mathbf{v}_J\}$ are drawn from a distribution η which has a density.

Then, for any $J \geq 1$, η -a.s. $FSSD^2 = 0 \iff p = q$.

Characteristics of FSSD:

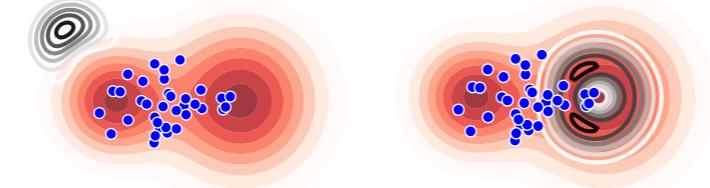
- Nonparametric. Do not need the normalizer of p .
- Runtime: $\mathcal{O}(n)$. Higher test power than LKS. ☺☺

Model Criticism with FSSD

Proposal: Optimize locations $\{\mathbf{v}_1, \dots, \mathbf{v}_J\}$ and kernel bandwidth by $\arg \max$ score = $FSSD^2 / \sigma_{H_1}$ (runtime: $\mathcal{O}(n)$).

Proposition: This procedure maximizes the true positive rate = $\mathbb{P}(\text{detect difference} \mid p \neq q)$.

score: 0.034 score: 0.44

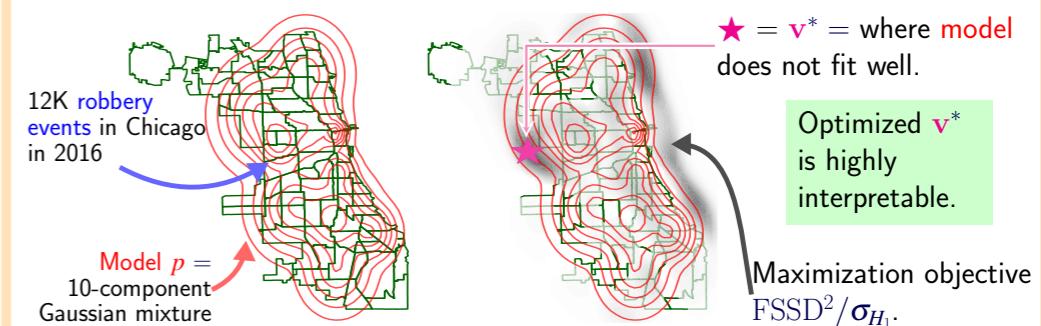


WJ, WX, and AG thank the Gatsby Charitable Foundation for the financial support. ZSz was financially supported by the Data Science Initiative. KF has been supported by KAKENHI Innovative Areas 25120012.

Contact: wittawat@gatsby.ucl.ac.uk

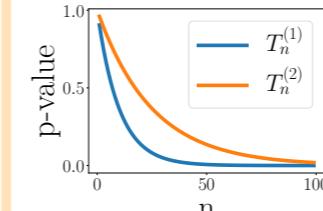
Code: github.com/wittawatj/kernel-gof

Interpretable Features for Model Criticism



Bahadur Slope and Bahadur Efficiency

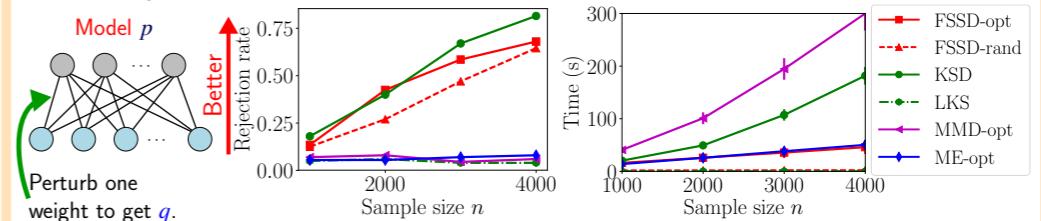
- Bahadur slope \approx rate of p-value $\rightarrow 0$ of statistic T_n under H_1 . High = good.
- Bahadur efficiency = ratio $\frac{\text{slope}^{(1)}}{\text{slope}^{(2)}}$ of slopes of two tests. > 1 means test⁽¹⁾ better.
- Results: Slopes of FSSD and LKS tests when $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(\mu_q, 1)$.



Proposition. Let σ_k^2, κ^2 be kernel bandwidths of FSSD and LKS. Fix $\sigma_k^2 = 1$. Then, $\forall \mu_q \neq 0$, $\exists \mathbf{v} \in \mathbb{R}$, $\forall \kappa^2 > 0$, the Bahadur efficiency $\frac{\text{slope}^{(\text{FSSD})}(\mu_q, \mathbf{v}, \sigma_k^2)}{\text{slope}^{(\text{LKS})}(\mu_q, \kappa^2)} > 2$. FSSD is statistically more efficient than LKS.

Experiment: Restricted Boltzmann Machine

- 40 binary hidden units. $d = 50$ visible units. Significance level $\alpha = 0.05$.
- Perturb one weight to get q .
- FSSD-opt, (FSSD-rand) = Proposed tests. $J = 5$ optimized, (random) locations.
- MMD-opt [Gretton et al., 2012] = State-of-the-art two-sample test (quadratic-time).
- ME-opt [Jitkrittum et al., 2016] = Linear-time two-sample test with optimized locations.
- Key: FSSD ($\mathcal{O}(n)$), KSD ($\mathcal{O}(n^2)$) have comparable power. FSSD is much faster.



ガウス過程の基礎

- まず、ガウス過程ではない例で考えてみましょう。
たとえば、1次元の入力 x について $\phi(x) = (1, x, x^2, x^3)^T$ という特徴ベクトルを考えれば、 x の3次関数

$$y = w_0 + w_1x + w_2x^2 + w_3x^3$$

は、対応する重み $w = (w_0, w_1, w_2, w_3)^T$ を使った線形回帰モデルを

$$y = W^T \phi(x)$$

と表すことが出来る。

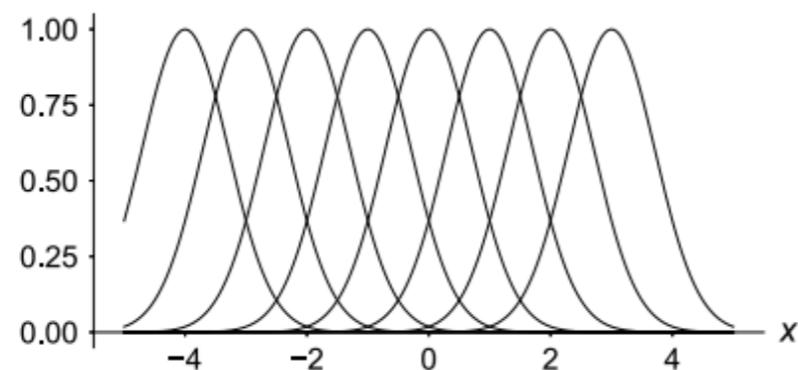
と定義する。

ガウス過程の基礎

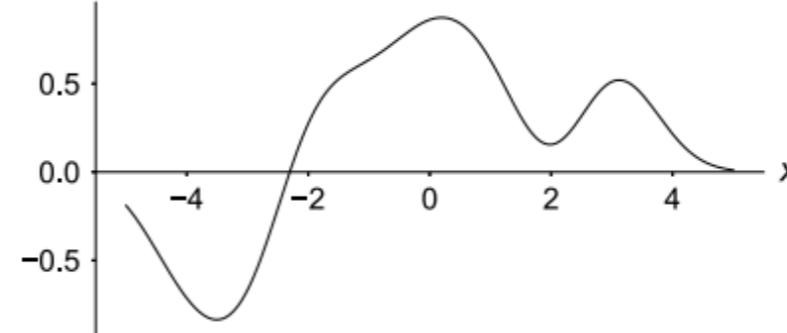
先ほどの回帰関数 $\phi(x)$ を $\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2, \phi_3(x) = x^3$ と 4 つの基底関数で表しているとみることができる。

また、基底関数を次のように定義し、任意の関数を表すことを考える。

$$\phi(x) = \exp\left(-\frac{x - \mu}{\sigma^2}\right)$$



(a) グリッド上のガウス基底関数
 $\phi_m(x) = \exp\left(-\frac{(x - \mu_m)^2}{\sigma^2}\right)$.



(b) 左の基底関数を重み $\mathbf{w} = (-0.48, -0.64, 0.41, 0.28, 0.57, 0.50, -0.26, 0.60)$ で混合して得られた関数。

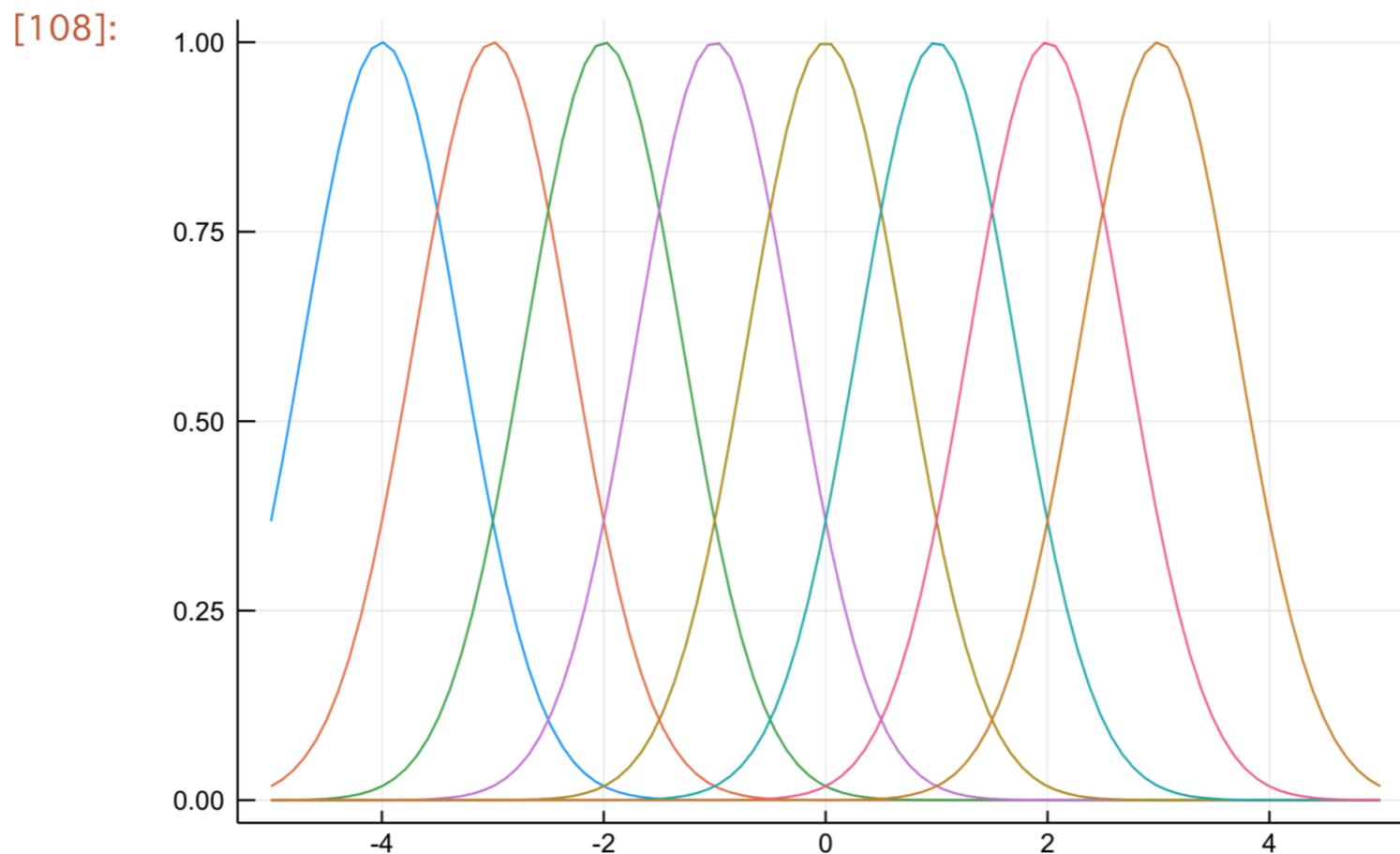
図 4.1 基底関数とその重みづけによる関数の表現。

ガウス過程の基礎

using Plots

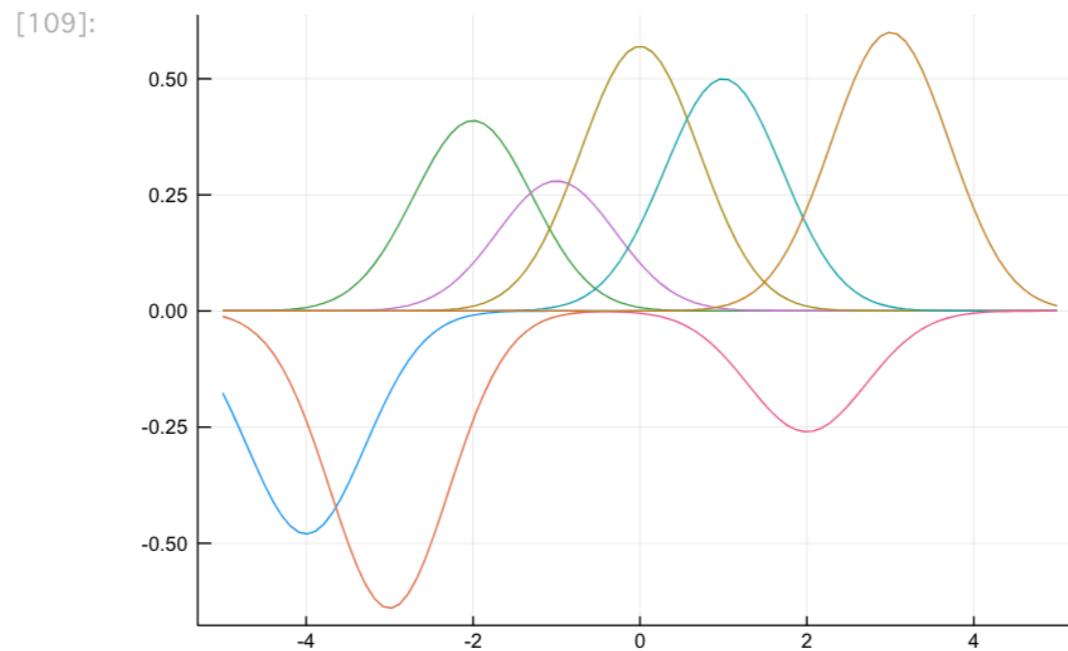
```
ϕ(x,μ,σ=1) = exp(-(x - μ)^2 / σ^2)
x=range(-5,stop=5,length=100)
μ = [-4,-3,-2,-1,0,1,2,3]
w = [-0.48,-0.64,0.41,0.28,0.57,0.50,-0.26,0.60]
```

```
[108]: plot(x,[ϕ.(x,μ[i]) for i in 1:length(μ)],legend=false)
```

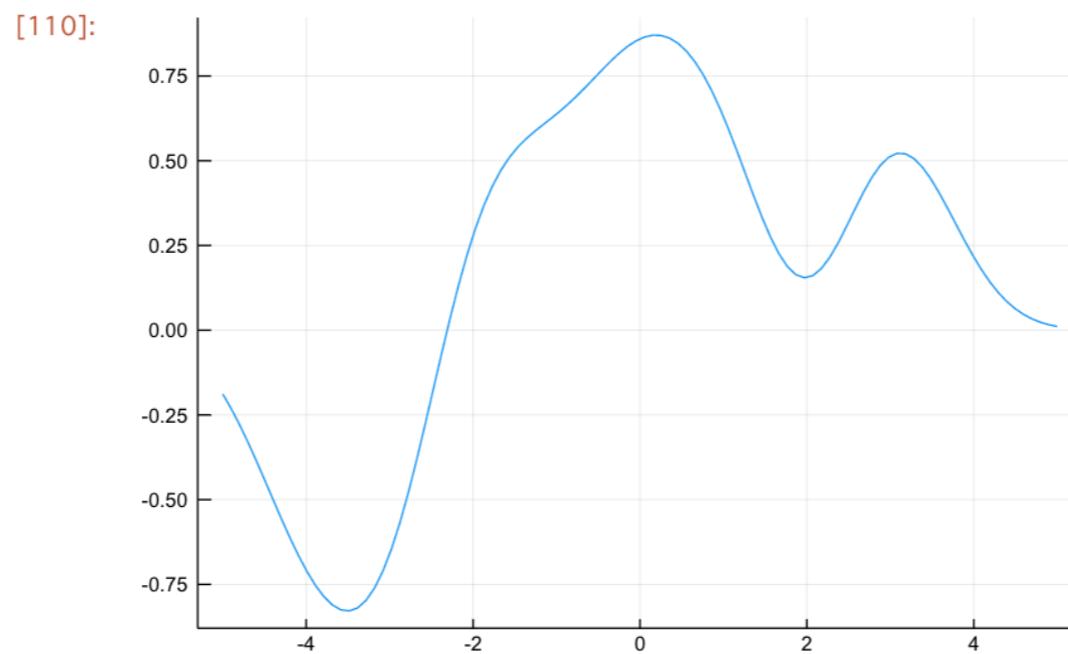


ガウス過程の基礎

```
[109]: plot(x,[w[i]*ϕ.(x,μ[i]) for i in 1:length(μ)],legend=false)
```



```
[110]: plot(x,sum([w[i]*ϕ.(x,μ[i]) for i in 1:length(μ)]),legend=false)
```



ガウス過程の基礎

しかし、この方法では x の次元が小さいときしか利用できない。 x の次元が1の場合、-10から10まで、間隔1.0で基底関数の中心 μ_m ならべたとしましょう。

$$y = W^T \phi(x)$$

これに対応するベクトル w は21になります。

すなわち $w =$

(-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10)

ガウス過程の基礎

では x が2次元の場合はどうなるでしょう

$$y = W^T \phi(x_1, x_2)$$

答えは

$$21^2 = 441 \text{ } (\dim=2)$$

$$21^3 = 9261 \text{ } (\dim=3)$$

⋮

$$21^{10} = 16,679,880,978,201 \text{ } (\dim=10)$$

ガウス過程の基礎

x が高次元の場合でもさきほどの図のように柔軟な回帰モデルを実現するにはどうすればいいでしょう。

解決法は線形回帰モデルのパラメータ W について期待値をとって、モデルから積分消去してしまうことです。

線形回帰モデルは次のように書くことも出来ます。

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \vdots & \ddots & & \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_M(x_N) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix}$$

ガウス過程の基礎

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{pmatrix} = \underbrace{\begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \vdots & \ddots & & \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_M(x_N) \end{pmatrix}}_{\phi} \underbrace{\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix}}_{w}$$

上のように置き換えると $\hat{y} = \phi w$ と表せます。

次に今回は簡単なために y が x から正確に誤差なく回帰されるとすると

$y = \phi w$ が成り立つと仮定します。

ガウス過程の基礎

ここで重み W が

$$w \sim N(0, \lambda^2 I)$$

で生成されるとします. $I =$ 単位行列

期待値を取ることは積分することとイコール

このとき, 共分散行列は次のようになります.

$$\begin{aligned}\mathbb{E} [\mathbf{y} \mathbf{y}^T] - \mathbb{E}[\mathbf{y}] \mathbb{E}[\mathbf{y}]^T &= \mathbb{E} [(\Phi \mathbf{w})(\Phi \mathbf{w})^T] = \Phi \mathbb{E} [\mathbf{w} \mathbf{w}^T] \Phi^T \\ &= \lambda^2 \Phi \Phi^T\end{aligned}$$

結果として \mathbf{y} の分布は次の多変量ガウス分布にしたがいます.

$$\mathbf{y} \sim \mathcal{N} (\mathbf{0}, \lambda^2 \Phi \Phi^T) \quad (4.1.1)$$

Wが消えました！！

ガウス過程

- 定義4.1

どんなN個の入力の集合 (x_1, x_2, \dots, x_N) についても、対応する出力 $y = (y_1, y_2, \dots, y_N)$ の同時分布が多変量ガウス分布に従うとき, $p(y)$ はガウス過程(Gaussian process)に従うといいます。

ガウス過程のまとめ

- 確率過程とはもともとは時系列に対する理論として生まれましたが、必ずしも時系列であることを要請しているわけではありません。
- 時系列ではなくても理論は同様に成り立つ。
- 入力の個数N,すなわち出力の次元Nはいくら大きくても成り立ちます。
- ガウス過程とは実は無限次元のガウス分布のことです。

ガウス過程の なにがうれしいのか

式4.1.1の共分散行列を

$$\mathbf{K} = \lambda^2 \Phi \Phi^T$$

とおくと、(n,m)要素は、図のように

$$\mathbf{K} = \lambda^2 \Phi \Phi^T = \lambda^2 \underbrace{\begin{pmatrix} \vdots \\ \boxed{\phi(\mathbf{x}_n)^T} \\ \vdots \end{pmatrix}}_{\Phi} \underbrace{\left(\cdots \boxed{\phi(\mathbf{x}_m)} \cdots \right)}_{\Phi^T}$$

\mathbf{x} の特徴ベクトルを $\phi(x) = (\phi_0, \dots, \phi_M(x))^T$ として、

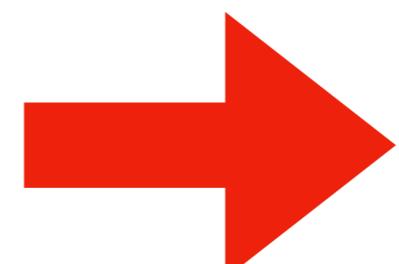
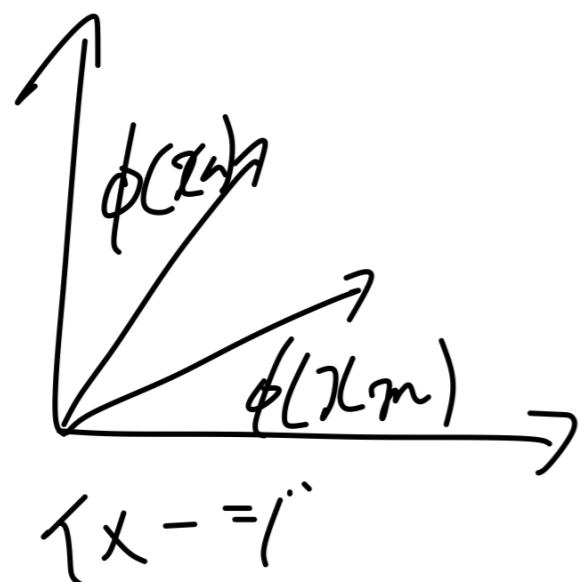
$$K_{nm} = \lambda^2 \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

で表されます。

ガウス過程の なにがうれしいのか

共分散行列 $\mathbf{K} = \lambda^2 \Phi \Phi^T$ が互いに似ていれば
特徴ベクトル $\phi(x_n)$ と $\phi(x_m)$ の内積の定数倍が、共分散行列
 K の(n,m)要素 K_{nm} になっています。

すなわち特徴ベクトル空間において x_n と x_m が似ているなら
対応する y_n と y_m も似た値を持つことになります。



大きいならば

y_n と y_m も似た値になる。

入力xが似ているならばyも似た値にある

ガウス過程の応用

Variational Learning on Aggregate Outputs with Gaussian Processes

<https://arxiv.org/abs/1805.08463>

- ・教師あり学習は入力と出力が同様の精度で観測されることを想定しているが、一般的な教師あり学習では入力よりも出力の方が粗く観測される。この問題に対する、アプローチをガウス過程を用いた変分学習を提案している。

参考文献

- ・情報幾何学の新展開：甘利俊一
- ・情報幾何学の基礎：藤原彰夫
- ・曲線と曲面の微分幾何：小林昭七
- ・カーネル法入門：福水健次
- ・Gaussian Process for Machine Learning (<http://www.gaussianprocess.org/gpml/>)
- ・ガウス過程の基礎と教師なし学習 (<http://www.ism.ac.jp/~daichi/lectures/H26-GaussianProcess/gp-lecture2-daichi.pdf>)
- ・『ガウス過程と機械学習』サポートページ (<http://chasen.org/~daiti-m/gpbook/>)

