

Kernel Bayesian Inference with Posterior Regularization

Song et al. NIPS2016

第3回NIPS+ 読み会

Osaka univ.
Kano Lab.
Yuchi Matsuoka

自己紹介

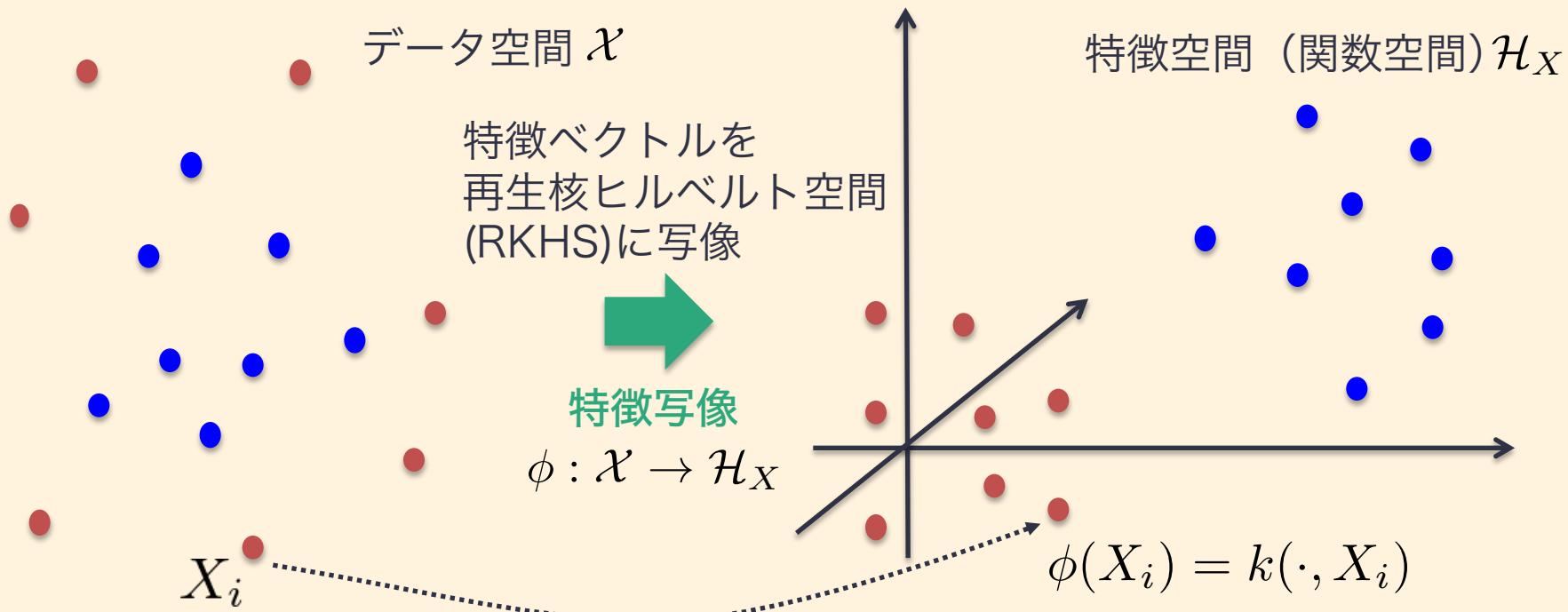
- 所属
 - 大阪大学基礎工学研究科システム創成専攻
数理科学領域データ科学研究室
(狩野研究室)
- 研究内容
 - 数理統計, 機械学習
 - 主にカーネル法, 因果推論, 統計的学習理論など.
- 来年度勉強したい. . .
 - 経験過程 (と統計的学習理論)
 - 位相的データ解析

興味ある方, 勉強会
しましょう!

発表のアジェンダ

1. カーネル平均の導入
 1. カーネル平均とその応用
 2. 同時, 条件付き分布のカーネル平均
 3. Kernel Bayes' Rule
2. 論文紹介
 1. モチベーションと流れ
 2. 論文詳細
 3. 実験（状態空間モデル or HMMのフィルタリングへの応用）
- 方針
 - 厳密にやると関数解析の前提知識と, かなりの導入が必要になるので, ところどころ雰囲気で説明します.
気になる方は途中でも質問していただくな, Appendixを参考にしてください. m(_ _)m

カーネル法とは



例えば、ガウシアンカーネル $k(x, y) = \exp(-||x - y||^2/\sigma^2)$

特殊な内積を持つ. $\langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}} = k(X_i, X_j)$

再生性 $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_X}, \forall f \in \mathcal{H}_X.$

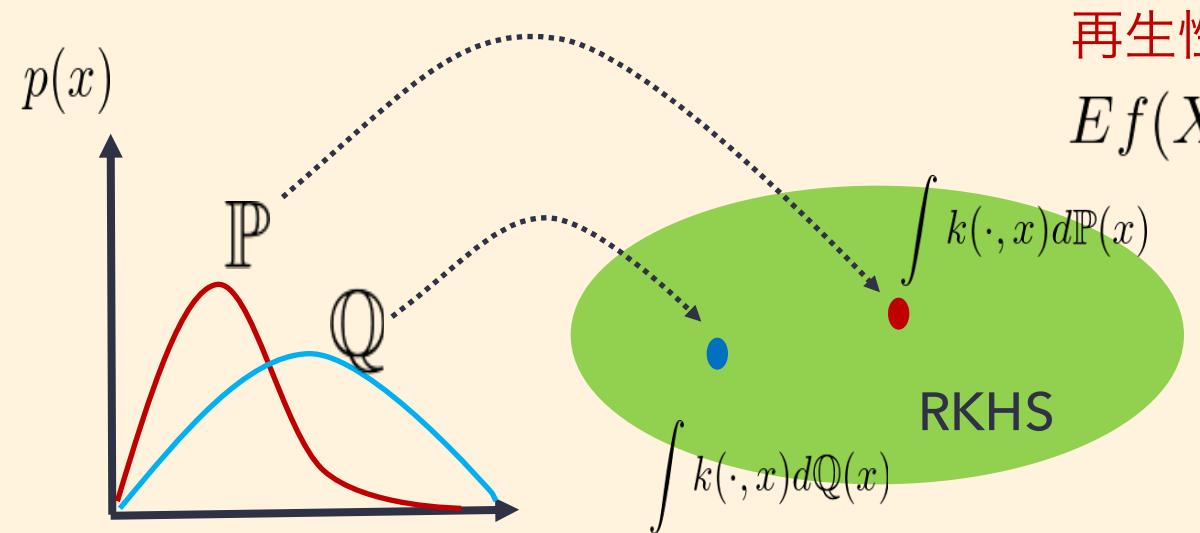
データ点から確率測度へ

- $X \sim p_X$ とし, カーネル $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}_X}$ とする. このとき, \mathbb{P} のカーネル平均を以下で定義:

$$\mu_X := \int k(\cdot, x) dp_X(x) = \int \phi(x) dp_X(x)$$

- p_X をディラック測度とすれば, いつもの特徴写像!

$$\int k(\cdot, y) \delta_x(y) = k(\cdot, x)$$



再生性 :

$$Ef(X) = \langle f, \int k(\cdot, x) d\mathbb{P}(x) \rangle_{\mathcal{H}_k}$$

カーネル平均を考える利点

- 特性的と呼ばれるクラスのカーネルを用いれば、元の確率測度の情報をすべて保存している。 $(P \rightarrow \mu_P)$ は单射)
 \therefore たとえば、 $k(x, x') = e^{xx'}$ としたときのXのカーネル平均は、

$$\begin{aligned}\mu_X(t) &= \mathbb{E}_{X \sim P}[k(t, X)] = \int e^{tx} dP(x) \\ &= \int 1 + tx + \frac{t^2 x^2}{2} + \frac{t^3 x^3}{3!} + \dots dP(x) \\ &= 1 + t\mathbb{E}_{X \sim P}[X] + \frac{t^2}{2}\mathbb{E}_{X \sim P}[X^2] + \frac{t^3}{3!}\mathbb{E}_{X \sim P}[X^3] + \dots\end{aligned}$$

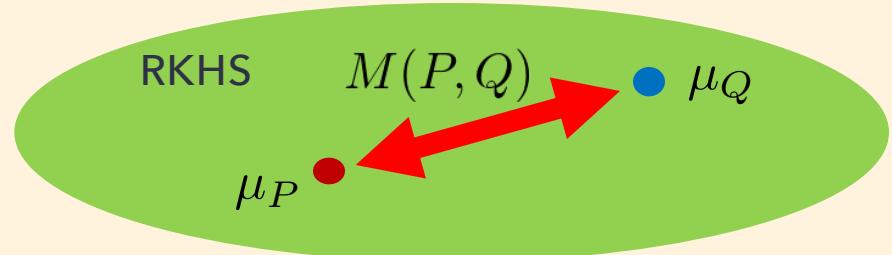
→モーメント母関数の役割を果たしている。

- ノンパラメトリックに確率推論を行うことが可能。
 - 推定量:
$$\hat{\mu}_X := \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i)$$
 - カーネル密度推定などのノンパラメトリック手法と異なり次元の呪いを受けない。

応用例(2標本問題)

参考:
http://qiita.com/yuchi_m/items/7132b426d848dc81ad9f
に解説記事を書きました.

- 2標本問題 (Gretton et al. NIPS 2009)
 - データ $X_1, \dots, X_\ell \sim P$ $Y_1, \dots, Y_n \sim Q$ (多次元でもOK)
 - $H_0 : P = Q$ vs $H_1 : P \neq Q$ の仮説検定.
- 検定統計量:
 - MMD(maximum mean discrepancy) : $M^2(P, Q) := \|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2$ の推定量 (を不偏化したもの).
 - $P=Q$ ならMMDは0となる.
- 異種データ間マッチング (Yoshikawa et al. NIPS2015)
 - 異なるドメインの相関のあるデータ (例えば, Wikipedia上の日本語と英語の同じ内容の記事) を潜在分布間のMMDを推定することでマッチング.

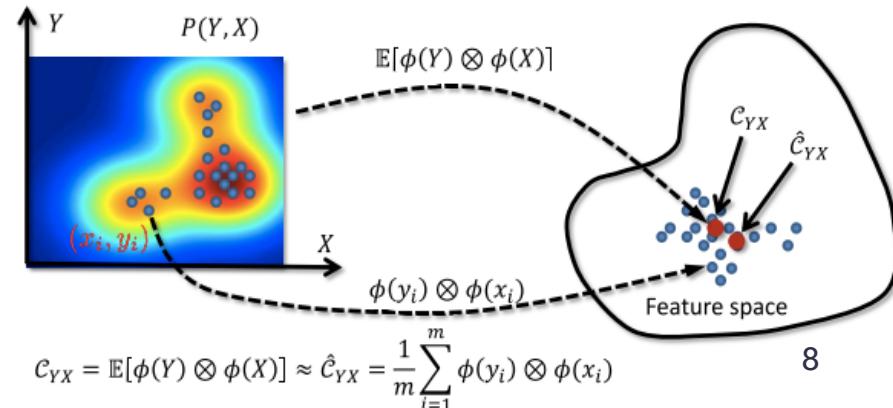


同時確率への拡張

- 記号の準備
 - $X, Y \sim p$, それぞれのRKHSと特徴写像: \mathcal{H}_X and $\phi(x)$, \mathcal{H}_Y and $\psi(y)$.
- 共分散作用素 $C_{YX} : \mathcal{H}_X \rightarrow \mathcal{H}_Y$ を
以下で定義:
$$\langle g, C_{YX} f \rangle_{\mathcal{H}_Y} = E[f(X)g(Y)] \quad \forall f \in \mathcal{H}_X, g \in \mathcal{H}_Y$$
- これは積力一ネル $k((x_1, y_1), (x_2, y_2)) = k_X(x_1, x_2)k_Y(y_1, y_2)$ により定義されるテンソル積 $\mathcal{H}_X \otimes \mathcal{H}_Y$ 上の測度 p のカーネル平均と同等である, i.e.

$$C_{YX} = E[\psi(Y) \otimes \phi(X)] \in \mathcal{H}_Y \otimes \mathcal{H}_X$$

共分散作用素は同時分布の
テンソル積への
カーネル平均埋め込み.



From Song et al. 2013

条件付き確率への拡張

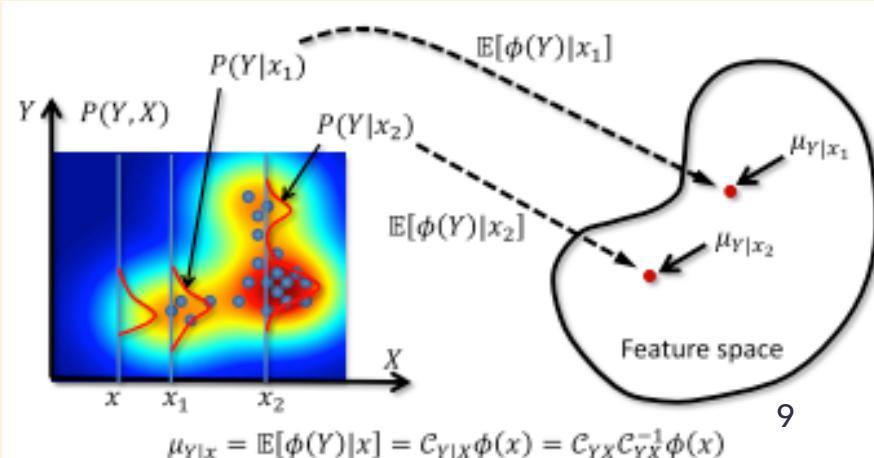
- x が与えられたもとでの Y の確率測度 $P(Y|X=x)$ のカーネル平均：

$$\mu_{Y|X=x} = E[\psi(Y)|X = x]$$

- いくつかの条件下で共分散作用素を用いて
 $\mu_{Y|X=x} = C_{YX}C_{XX}^{-1}\phi(x)$ と書ける.(Thm.1 Proof in Appdendix)
- 条件付きカーネル平均の推定量：
 $\hat{\mu}_{Y|X=x} = \hat{C}_{YX}(\hat{C}_{XX} + \lambda I)^{-1}\phi(x)$
具体的なグラム行列表現は、Song et al. 2013などを参照

- 再生性：

$$E[g(Y)|X = x] \approx \langle g, \hat{\mu}_{Y|X=x} \rangle_{\mathcal{H}_Y}, \quad \forall g \in \mathcal{H}_Y.$$



From Song et al. 2013

カーネルベイズのモチベ

- Bayes' rule

$$p^\pi(Y|X = x) \propto \underbrace{\pi(Y)}_{\text{事前分布}} \underbrace{p(X = x|Y)}_{\text{尤度}}$$

- 言わずもがな、統計学、機械学習のいたるところで重要な定理です。
- 難点
 - 数値計算が大変：MCMC, SMC, ABC, etc…
 - 尤度に対する適切なパラメトリックモデルの仮定が必要。

- カーネルベイズのモチベーション

- 事後分布の代わりにそのカーネル平均埋め込みを推定することで、ノンパラメトリックかつ次元の呪いを受けない推定を行いたい！



Kernel Bayes Rule

(Fukumizu et al. NIPS2011)

- Y に対する事前分布 $\pi(Y)$, 尤度 $p(X = x|Y)$ とする.
- Goal:
 - 観測 x と, 事前分布 $\pi(Y)$ が与えられたときの事後分布 $p^\pi(Y|X = x)$ に対応するカーネル平均埋め込み $\mu_Y^\pi(X = x)$ が求めたい.
- 難しさ :
 - 条件付きカーネル平均埋め込み $\mu_Y^\pi(X = x) = C_{YX}^\pi C_{XX}^{\pi^{-1}} \phi(x)$ を求めればよいので一見簡単そう.



But

$$C_{YX}^\pi = E_{p^\pi(Y,X)}[\phi_Y(Y) \otimes \phi_X(X)] \text{ であり,}$$

一般に

$$p^\pi(Y, X) = \pi(Y)p(X|Y) \neq p(Y, X).$$

単純に, $p(Y, X)$ からの標本平均で推定することができない.

- 解決法：
 - 共分散作用素はテンソル積空間上のカーネル平均埋め込みと同一視できる. →条件付き埋め込みとして推定可能.

- Theorem 1を用いて,

$$C_{YX}^\pi = C_{(YX)Y} C_{YY}^{-1} \pi_Y, \quad C_{XX}^\pi = C_{(XX)Y} C_{YY}^{-1} \pi_Y$$

となることが示せる. (データから推定可能)

- データからこれらを推定すれば,

$$\hat{\mu}_Y^\pi(X = x) = \hat{C}_{YX}^\pi ([\hat{C}_{XX}^\pi]^2 + \lambda I)^{-1} \hat{C}_{XX}^\pi \phi(x)$$

で解が得られる.

注 \hat{C}_{XX}^π が正定値とは限らないため, $(B + \lambda I)^{-1} z$ ではなく,
 $(B^2 + \lambda I)^{-1} B z$ として正則化.

- 具体的なグラム行列表現は元論文参照.

本論文のモチベーション

- Observations
 - 条件付き分布のカーネル埋め込みは最適化問題として定式化できる. (Grünewälder et al. ICML2012)
 - ベイズ事後分布は最適化問題として定式化できる (RegBayes, Zhu et al. JMLR2014).

カーネルベイズは最適化問題として定式化できるのか？



YES!

その定式化により新しい知見は得られるのか？



Thresholding Regularization, kRegBayesなどのより優れた推定アルゴリズムを提案.

ベイズルールの最適化問題

- ベイズルールは以下の最適化問題の解である. (Proof in Appendix)

$$\min_{p(Y|X=x)} \text{KL}(p(Y|X=x) || \pi(Y)) - \int \log p(X=x|Y) dp(Y|X=x)$$

s.t. $p(Y|X=x) \in \mathcal{P}_{\text{prob}}$

- RegBayes(Zhu et al. JMLR2014)
 - 事後分布に対する正則化を加えて, 定式化.

$$\min_{p(Y|X=x), \xi} \text{KL}(p(Y|X=x) || \pi(Y)) - \int \log p(X=x|Y) dp(Y|X=x) + U(\xi)$$

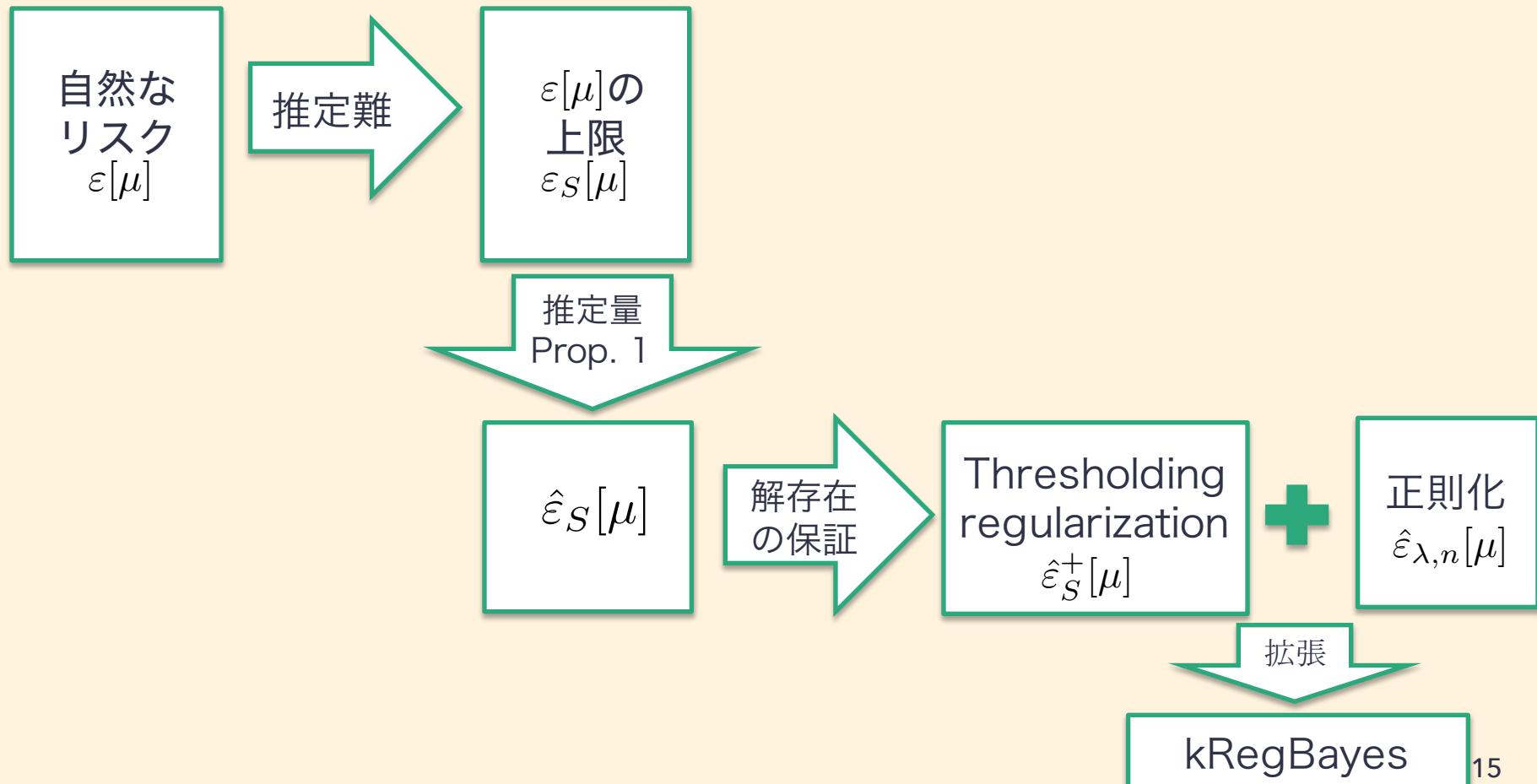
s.t. $p(Y|X=x) \in \mathcal{P}_{\text{prob}}(\xi)$



これをカーネル化したい

ここからの流れ

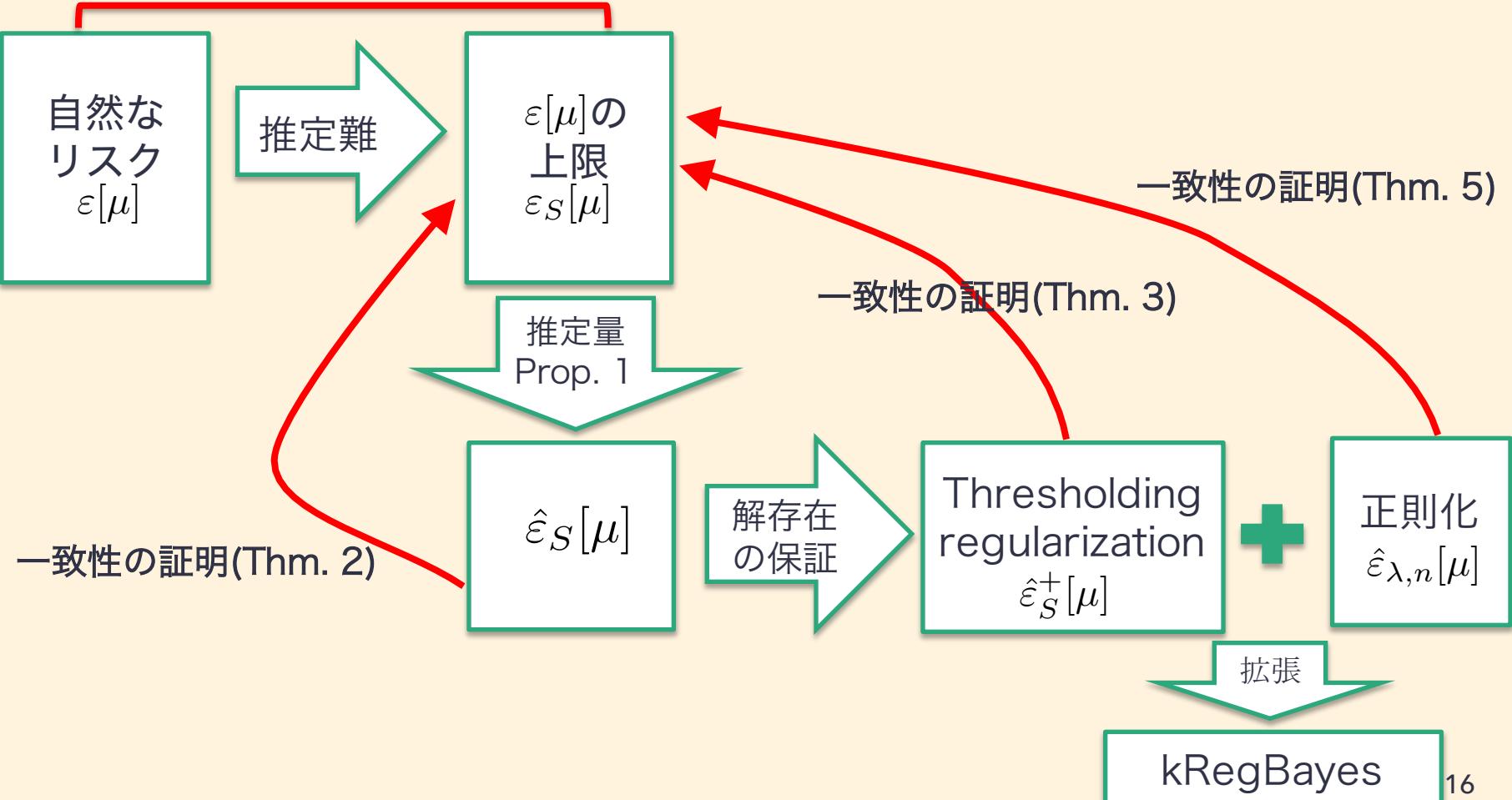
目的：カーネルベイズを解として持つ最適化問題を得て、
推定，拡張していく。



ここからの流れ

妥当性の証明：新しい問題がカーネルベイズ解に一致性をもつことを確認。

解が等しいことの証明(Thm. 4)



カーネルベイズに対する自然な目的関数

- 事後分布 $p^\pi(Y|X=x)$ の埋め込み $\mu_Y^\pi(X=x) \in \mathcal{H}_Y$ は、任意の $h \in \mathcal{H}_Y$ に対して、

$$\langle h, \mu_Y^\pi(X=x) \rangle_{\mathcal{H}_Y} = E_{p^\pi(Y|X=x)}[h(Y)]$$

- これにより、最小化したい自然なリスクは、

$$\varepsilon[\mu] := \sup_{\|h\|_{\mathcal{H}_Y} \leq 1} E_X[(E_Y[h(Y)|X]) - \langle h, \mu(X) \rangle_{\mathcal{H}_Y}]^2$$

- ただし、 $E_X[\cdot]$ は $p^\pi(X)$, $E_Y[\cdot|X]$ は $p^\pi(Y|X)$ での期待値。

目標：データ $(x_i, y_i)_{i=1}^n$ から、
 $\varepsilon[\mu]$ を最小化する $\mu : \mathcal{X} \rightarrow \mathcal{H}_Y$ を見つける。

→ 次の上限を得る。(Proof in Appendix)

$$\varepsilon_S[\mu] = E_{(X,Y) \sim p^\pi(X,Y)}[\|\psi(Y) - \mu(X)\|_{\mathcal{H}_Y}^2]$$

Theorem 4 (More detail in Appendix)

ある仮定のもとで、 $\mu^* := \operatorname{argmin}_{\mu \in \mathcal{H}_K} \varepsilon[\mu] = \operatorname{argmin}_{\mu \in \mathcal{H}_K} \varepsilon_S[\mu]$ $p_X - a.s.$

→ 最小化したいリスク $\varepsilon[\mu]$ と最小化しやすいリスク $\varepsilon_S[\mu]$ が等価。

$\varepsilon_S[\mu]$ の推定量は？

- 難しさ：
私たちは $p^\pi(X, Y)$ からのサンプルを持っていない。

- 構成方法

- 仮定：

1. $f(x, y) := \|\psi(y) - \mu(x)\|_{\mathcal{H}_Y}^2 \in \mathcal{H}_X \otimes \mathcal{H}_Y$

$$\rightarrow \varepsilon_s[\mu] = E_{(X, Y)}[\|\psi(Y) - \mu(X)\|_{\mathcal{H}_Y}^2] = \langle f, \mu_{(X, Y)} \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}$$

2. 事前分布のカーネル平均の一致推定量：

$$\hat{\pi}_Y = \sum_{i=1}^{\ell} \tilde{\alpha}_i \psi(\tilde{y}_i) \text{ given.}$$



$\hat{\varepsilon}_s[\mu] = \langle \hat{\mu}_{(X, Y)}, f \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}$ ($\varepsilon_S[\mu]$ の一致推定量, Thm. 2)

where, $\hat{\mu}_{(X, Y)} = \hat{C}_{(X, Y)Y} (\hat{C}_{YY} + \lambda I)^{-1} \hat{\pi}_Y$

Prop. 1 (推定量の具体的表現, Proof in Appendix)

$$\hat{\varepsilon}_s[\mu] = \sum_{i=1}^n \beta_i \|\psi(y_i) - \mu(x_i)\|_{\mathcal{H}_Y}^2,$$

where. $\boldsymbol{\beta} = (G_Y + n\lambda I)^{-1} \tilde{G}_Y \tilde{\boldsymbol{\alpha}}$, $(G_Y)_{ij} = k_Y(y_i, y_j)$, $(\tilde{G}_Y)_{ij} = k_Y(y_i, \tilde{y}_j)$

Thresholding regularization

- 実は、Prop1は β_i が負の値をとることもあるので、常に最小値を持つとは限らない。

$$\begin{array}{ccc} \beta_i & \xrightarrow{\text{置き換え}} & \beta_i^+ := \max(0, \beta_i) \\ & & \text{Thresholding regularization} \\ \text{リスク : } \epsilon_s[\mu] & \xrightarrow{} & \epsilon_s^+[\mu] \end{array}$$

- Theorem 3. いくつかの条件下で、

$$|\hat{\epsilon}_s^+[\mu] - \epsilon_s[\mu]| \xrightarrow{p} 0.$$

注 thresholding regularizationは、Nishiyama et al. arXiv2012における部分観測マルコフ決定過程(POMDPs)のカーネル埋め込みでも使われていた。しかし、そこではTheorem 3のconsistencyは未証明であった。

Minimizing $\hat{\epsilon}_s^+[\mu]$

- 正則化項を加えて,

$$\hat{\epsilon}_{\lambda,n}[\mu] = \sum_{i=1}^n \beta_i^+ \|\psi(y_i) - \mu(x_i)\|_{\mathcal{H}_Y}^2 + \lambda \|\mu\|_{\mathcal{H}_K}^2$$

を目的関数とする.

Proposition 2 (Proof in Appendix)

$\hat{\mu}_{\lambda,n} = \operatorname{argmin}_{\mu} \hat{\epsilon}_{\lambda,n}[\mu]$ は以下で求まる.

$$\hat{\mu}_{\lambda,n}(x) = \Psi(K_X + \lambda_n \Lambda^+)^{-1} K_{:x}$$

where. $\Psi = (\psi(y_1), \dots, \psi(y_n))$, $(K_X)_{ij} = k_X(x_i, x_j)$,

$$\Lambda^+ = \operatorname{diag}(1/\beta_1^+, \dots, 1/\beta_n^+), K_{:x} = (k_X(x, x_1), \dots, k_X(x, x_n))^T.$$

- Theorem 5

- 正則化パラメータに関する適切な仮定のもとで,

$$\varepsilon_S[\hat{\mu}_{\lambda_n,n}] \xrightarrow{P} \min_{\mu} \varepsilon_S[\mu].$$

ちゃんと
カーネルベイズ解
にいく.

kRegBayes

- Thresholding regularizationをさらに発展させ, 次のように正則化項をいれる.
- kRegbayes:
 - 目的関数 :

$$\mathcal{L} := \sum_{i=1}^m \beta_i^+ \|\psi(y_i) - \mu(x_i)\|_{\mathcal{H}_Y}^2 + \lambda \|\mu\|_{\mathcal{H}_K}^2 + \delta \sum_{i=m+1}^n \|\mu(x_i) - \psi(t_i)\|_{\mathcal{H}_Y}^2.$$

- この解はProp. 2 と全く同様にして, 以下で得られる:

Proposition 3

$$\hat{\mu}_{reg}(x) = \Psi(K_X + \lambda \Lambda^+)^{-1} K_{:x}$$

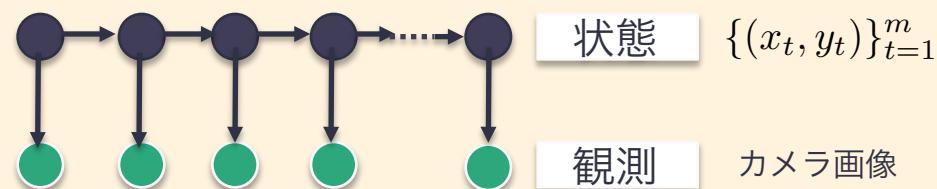
where. $\Psi = (\psi(y_1), \dots, \psi(y_n))$, $(K_X)_{ij} = k_X(x_i, x_j)$,

$$\Lambda^+ = \text{diag}(1/\beta_1^+, \dots, 1/\beta_m^+, 1/\delta, \dots, 1/\delta), K_{:x} = (k_X(x, x_1), \dots, k_X(x, x_n))^T.$$

- 利点は?
 - ただ, 一部の β を定数 δ に変えただけに見える??
➡ テストデータに関する事前知識を利用できる!

Experiment

- Camera position recovery(状態空間モデルへの応用)
 - 状態：各時刻tでのカメラの位置, $\{(x_t, y_t)\}_{t=1}^m$.
 $\theta_{t+1} = \theta_t + 0.2 + \mathcal{N}(0, 4e - 1)$, $r_{t+1} = \max(R_2, \min(R_1, r_t + \mathcal{N}(0, 1)))$
として, $x_{t+1} = r_{t+1} \cos \theta_{t+1}$, $y_{t+1} = r_{t+1} \sin \theta_{t+1}$
 - 観測：各時刻tでの100×100カメラ画像.(10000次元ベクトル)
 - 目的：
状態空間モデル（右図）



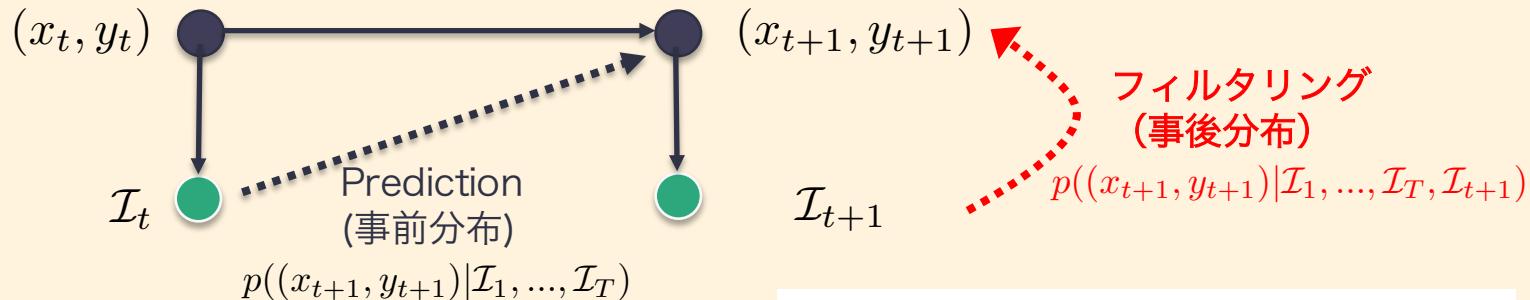
をトレーニングデータから学習し, テストデータのカメラ画像から, その時点のカメラの位置を予測する(フィルタリング).



Figure 2: First several frames of training data (upper row) and test data (lower row).

- トレーニングデータでは $R_1 = 0, R_2 = 10$
テストデータでは $R_1 = 5, R_2 = 7$ とする.
- この事前知識をkRegBayesでは組み込むことができる.
→ 半径6の円上の一様分布から, Supervision dataを生成し,
kRegBayesの正則化項を構成する.

大体半径r=6の円上
が答えっぽいという
事前知識

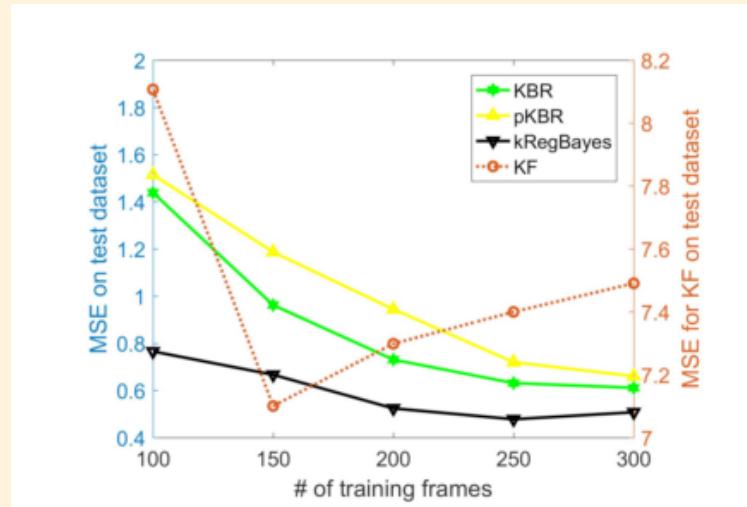


より半径6の円上に近づくように
学習が進む.

- カーネルベイズ(KBR),
カルマンフィルタ(KF),
Thresholding regularization(pKBR),
kRegBayesでMSEを比較.

結果

- いずれもkRegBayesは他手法を大きく上回っている.
スモールサンプルかつ事前知識が活用できるとき,
特に性能が高い.



References

- Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *In Advances in neural Information Processing Systems*, pages 513–520, 2007.
- Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. *In Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968. ACM, 2009.
- Le Song, Kenji Fukumizu. and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine* 30.4 : 98-11. 2013
- Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel bayes' rule. *In Advances In Advances in neural information processing systems*, pages 1737–1745, 2011.
- Yang Song, Jun Zhu, and Yong Ren. Kernel Bayesian Inference with Posterior Regularization. *In Advances in neural information processing systems*, pages 4763-4771, 2016.
- Yuya Yoshikawa, Tomoharu Iwata, Hiroshi Sawada, and Takeshi Yamada. Cross-domain matching for bag-of-words data via kernel embeddings of latent distributions. *In Advances in neural information processing systems*, pages 1405-1413. 2015.
- Yu Nishiyama, Abdeslam Boularias, Arthur Gretton, and Kenji Fukumizu. Hilbert space embeddings of pomdps. *arXiv preprint arXiv:1210.4887*, 2012.
- Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. *In Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1823–1830, 2012.
- Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.

Kernel Bayesian Inference with Posterior Regularization (Appendix)

Yuchi Matsuoka

2017 年 3 月 18 日

1 Preliminaries

$(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ を確率変数の可測空間とし, p_X をそれ上の確率測度とする. $\mathcal{H}_{\mathcal{X}}$ をカーネル $k(\cdot, \cdot)$ に対応する RKHS とする. p_X のカーネル平均は $\mu_X = E_{p_X}[\phi(X)] \in \mathcal{H}_{\mathcal{X}}$ で定義される. ただし, $\phi(X) = k(X, \cdot)$. ¹通常の再生成と同様に, $f \in \mathcal{H}$ に対して, $E_{p_X}[f(X)] = E_{p_X}[\langle f, \phi(X) \rangle] = \langle f, \mu_X \rangle$ が成り立つ. universal kernel と呼ばれるカーネルは対応する RKHS \mathcal{H} が sup norm の意味で $\mathcal{C}_{\mathcal{X}}$ で稠密となる. 次に線形作用素のカーネル平均を定義する.

2つの可測空間 $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}), (\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ を考え, $\phi(x), \psi(y)$ で対応する有界なカーネルによる RKHS $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ の可測な特徴写像とする. p は $\mathcal{X} \times \mathcal{Y}$ 上の (X, Y) の同時分布とする. 共分散作用素 \mathcal{C}_{XY} は $\mathcal{C}_{XY} = E_p[\phi(X) \otimes \psi(Y)]$ で定義される. カーネル $k((x_1, y_1), (x_2, y_2)) = k_{\mathcal{X}}(x_1, x_2)k_{\mathcal{Y}}(y_1, y_2)$ に対応する RKHS $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ 上の $\mu_{(XY)}$ と同等である. 以下の定理は重要である.

Theorem 1 \mathcal{C}_{XX} が单射で, $\mu_X \in R(\mathcal{C}_{XX})$ かつ任意の $g \in \mathcal{H}_{\mathcal{Y}}$ に対して, $E[g(Y)|X = \cdot] \in \mathcal{H}_{\mathcal{X}}$ であれば,

$$\mu_Y = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\mu_X, \quad \mu_{Y|X=x} = E[\psi(Y)|X = x] = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\phi(x).$$

2 推定量とその一致性について

μ_X は p_X からのサンプル $\{x_i\}_{i=1}^N$ により, $\hat{\mu}_X = \frac{1}{N} \sum_{i=1}^N \phi(x_i)$, 共分散作用素 \mathcal{C}_{XY} は $\hat{\mathcal{C}}_{XY} = \frac{1}{N} \sum_{i=1}^N \phi(x_i) \otimes \psi(y_i)$ により推定する. どちらも RKHS ノルムの意味で $O_p(N^{-1/2})$ で収束する.

¹この期待値はカーネルが有界,i.e. $\sup_x k_{\mathcal{X}}(x, x) < \infty$ なら常に存在する.

3 Theorem 1 の証明

Theorem 1 ((Song et al., 2009, Equation 6)) m_{Π} と m_{Q_y} は H_X 上の Π 及び, H_Y 上の Q_Y のカーネル平均とする. ここで, C_{XX} が単射で, $m_{\Pi} \in \mathcal{R}(C_{XX})$, さらに任意の $g \in H_Y$ に対して $E[g(Y)|X = \cdot] \in H_X$ が成り立つとする. このとき

$$m_{Q_y} = C_{YX} C_{XX}^{-1} m_{\Pi}.$$

が成り立つ. ただし $C_{XX}^{-1} m_{\Pi}$ は C_{XX} によって m_{Π} に写像される関数を表す.

Proof $C_{XX}f = m_{\Pi}$ となるような $f \in H_X$ を取ってくる. このとき任意の $g \in H_Y$ に対して,

$$\begin{aligned} \langle C_{YX}f, g \rangle &= \langle f, C_{XY}g \rangle = \langle f, C_{XX}E[g(Y)|X = \cdot] \rangle \\ &= \langle C_{XX}f, E[g(Y)|X = \cdot] \rangle = \langle m_{\Pi}, E[g(Y)|X = \cdot] \rangle = \langle m_{Q_y}, g \rangle. \end{aligned}$$

ここで最後の等号, $\langle m_{\Pi}, E[g(Y)|X = \cdot] \rangle = \langle m_{Q_y}, g \rangle$ を説明する. まず, カーネル平均の再生性

$$\langle f, m_X \rangle = E[f(X)]$$

より, $(X, Y) \sim p(x, y)$. $U \sim \pi(u)$. $(Z, W) \sim q(x, y) = \pi(x)p(y|x)$, $q_Y(y) = \int q(x, y)dx$ とおくと,

$$\begin{aligned} \langle m_{Q_y}, g \rangle_{\mathcal{H}_Y} &= E[g(W)] = \int g(w)q_Y(w)dw \\ \langle m_{\Pi}, E[g(Y)|X = \cdot] \rangle &= E_U[E_Y[g(Y)|U]] = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} g(y)p(y|u)dx \right) \pi(u)du \\ &= \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} g(y)q(u, y)du \right) dx = \int_{\mathcal{Y}} g(y)q_Y(y)dy. \end{aligned}$$

最後に $m_{Q_y} = C_{YX} C_{XX}^{-1} m_{\Pi}$ の m_{Π} を $k_{\mathcal{X}}(\cdot, x)$ で置き換えると $E[k_{\mathcal{Y}}(\cdot, Y)|X = x] = C_{YX} C_{XX}^{-1} k_{\mathcal{X}}(\cdot, x)$ が得られる. \square

4 カーネルベイズについて

$\pi(Y)$ を Y の事前分布, $p(X = x|Y)$ を尤度, $p^{\pi}(Y|X = x)$ は $\pi(Y)$ と観測 x が与えられた下での事後分布とし, $p^{\pi}(X, Y) = \pi(Y)p(X|Y)$ とする. 目的は事前分布のカーネル平均 π_Y と共に分散作用素 C_{XY} が与えられたときの事後分布のカーネル平均 $\mu_Y^{\pi}(X = x)$ を得ることである. ベイズルールより, $p^{\pi}(Y|X = x) \propto \pi(Y)p(X = x|Y)$.

条件付き分布が $p(X|Y)$ と一致するような $\mathcal{X} \times \mathcal{Y}$ 上の同時分布 p が存在する. そして C_{XY} はその共分散作用素であると仮定する. Thm. 1 より C_{YX}^π が同時分布 p^π に対応する共分散作用素, C_{XX}^π が p^π の X に関する周辺分布の共分散作用素であるとすれば,

$$\mu_Y^\pi(X = x) = C_{YX}^\pi C_{XX}^{\pi^{-1}} \phi(x).$$

C_{YX}^π が $\mathcal{H}_Y \otimes \mathcal{H}_X$ 上の $\mu_{(YX)}$ と同等であることを思い出せば, Thm 1. より,

$$\mu_{(YX)} = C_{(YX)Y} C_{YY}^{-1} \pi_Y, \text{ where. } C_{(YX)Y} := E[\psi(Y) \otimes \phi(X) \otimes \psi(Y)].$$

同様に, C_{XX}^π も

$$\mu_{(XX)} = C_{(XX)Y} C_{YY}^{-1} \pi_Y$$

で表現される.

5 ベイズルールの最適化問題

Regularized Bayesian inference (RegBayes) はベイズルールの別の定式化である. 事後分布は $\mathcal{P}_{\text{prob}}$ を有効な確率測度の集合としたとき,

$$\min_{p(Y|X=x)} \text{KL}(p(Y|X=x)||\pi(Y)) - \int \log p(X=x|Y) dp(Y|X=x)$$

$$s.t. \quad p(Y|X=x) \in \mathcal{P}_{\text{prob}}$$

の解としてみることができる.

Proof

$$\begin{aligned} \text{KL}(p(Y|X=x)||\pi(Y)) - \int \log p(X=x|Y) dp(Y|X=x) &= \int \log \frac{p(Y|X=x)}{\pi(Y)} dp(Y|X=x) - \int \log p(X=x|Y) dp(Y|X=x) \\ &= \int \log \frac{p(Y|X=x)}{\pi(Y)p(X=x|Y)} dp(Y|X=x) \\ &= \int \log \frac{p(Y|X=x)}{\frac{\pi(Y)p(X=x|Y)}{p^\pi(X=x)}} dp(Y|X=x) + \int \log p^\pi(X=x) dp(Y|X=x) \\ &= \text{KL} \left(p(Y|X=x) \middle\| \frac{\pi(Y)p(X=x|Y)}{p^\pi(X=x)} \right) + \log p^\pi(X=x). \end{aligned}$$

よって, $\arg \min_{p(Y|X=x)} \text{KL}(p(Y|X=x)||\pi(Y)) - \int \log p(X=x|Y) dp(Y|X=x) = \frac{\pi(Y)p(X=x|Y)}{p^\pi(X=x)}$. □

6 Vector-valued regression の一般論について補足

ベクトル値 (RKHS 値) 回帰では以下の目的関数を最小化する。

$$E(f) := \sum_{i=1}^n \|y_j - f(x_j)\|_{\mathcal{H}_Y}^2 + \lambda \|f\|_{\mathcal{H}_K}^2,$$

where. $y_j \in \mathcal{H}_Y, f : \mathcal{X} \rightarrow \mathcal{H}_Y$.

f は RKHS \mathcal{H}_Y に値をとる関数で, f はベクトル値 RKHS \mathcal{H}_K に属すると仮定する。

これについて, 元論文にもとづきもう少し詳しく解説する。以下の記号はこの小節のみにおいて有効である。

6.0.1 Vector-values regression and RKHSs

サンプル $\{(x_i, v_i)\}_{i \leq m}$ が $\mathcal{X} \times \mathcal{V}$ 上の確率分布から i.i.d に得られたとする。ただし, \mathcal{X} は空でない集合で, $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ はヒルベルト空間とする。目的は

$$E_{(X, V)}[\|f(X) - V\|_{\mathcal{V}}^2]$$

を最小化する $f : \mathcal{X} \rightarrow \mathcal{V}$ を見つけることである。これを *vector-valued regression problem* と呼ぶ。

[Definition] 関数 $h : \mathcal{X} \rightarrow \mathcal{V}$ からなるヒルベルト空間 $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\Gamma})$ は、任意の $x \in \mathcal{X}, v \in \mathcal{V}$ について、線形関数 $h \mapsto \langle v, h(x) \rangle_{\mathcal{V}}$ が連続であれば RKHS である。 \mathcal{H}_{Γ} と書く。

引用文献では上を定義としているが、これは再生各ヒルベルト空間のいくつかの同値な定義の一つである。

Riesz の表現定理²により、各 $x \in \mathcal{X}, v \in \mathcal{V}$ に対して、 \mathcal{V} から \mathcal{H}_{Γ} への線形作用素 Γ_x が存在して ($\Gamma_x v \in \mathcal{H}_{\Gamma}$ で書かれる), 任意の $h \in \mathcal{H}_{\Gamma}$ に対して,

$$\langle v, h(x) \rangle_{\mathcal{V}} = \langle h, \Gamma_x v \rangle_{\Gamma}$$

が成り立ち、これは再生性である。よって \mathcal{H}_{Γ} は RKHS である。

次に、ベクトル値再生カーネルを導入し、それと Γ_x の関連を述べる。 $\mathcal{L}(\mathcal{V})$ で \mathcal{V} から \mathcal{V} への有界線形作用素の空間を表すとし、再生カーネル $\Gamma(x, x') \in \mathcal{L}(\mathcal{V})$ を

$$\Gamma(x, x')v \in (\Gamma_{x'}v)(x) \in \mathcal{V}$$

²(Riesz の表現定理) ヒルベルト空間 \mathcal{H} について \mathcal{H} から \mathbb{R} への線形作用素を線形汎関数という。有界な（連續性と同値）線形汎関数全体のなすベクトル空間 \mathcal{H}^* を \mathcal{H} の双対空間という。任意の $\phi \in \mathcal{H}^*$ に対し、 $y_{\phi} \in \mathcal{H}$ があって、任意の $x \in \mathcal{H}$ に対して、 $\phi(x) = \langle x, y_{\phi} \rangle$ が成り立つ。

で定義する。定義と再生性から以下が成り立つ。

[Proposition 2.1] 関数 $\Gamma : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{L}(\mathcal{V})$ は以下が成り立つときカーネルである。

$$(1) \quad \Gamma(x, x') = \Gamma(x', x)^*.$$

$$(2) \quad \text{任意の } n \in \mathbb{N}, \{(x_i, v_i)\}_{i \leq n} \subset \mathcal{X} \times \mathcal{V} \text{ に対して, } \sum_{i,j \leq n} \langle v_i, \Gamma(x_i, x_j) v_j \rangle_{\mathcal{V}} \geq 0.$$

ここから回帰問題を定式化していく。まず目的関数 $E_{(X,V)}[||f(X) - V||_{\mathcal{V}}^2]$ の推定量 $\sum_{i=1}^n ||v_i - f(x_i)||_{\mathcal{V}}^2$ を考え, f を RKHS \mathcal{H}_{Γ} で制限する (ベクトル値関数空間)。そして \mathcal{H}_{Γ} ノルムによる正則化を考える。これにより正則化経験リスク

$$\hat{\epsilon}_{\lambda}(f) := \sum_{i=1}^n ||v_i - f(x_i)||_{\mathcal{V}}^2 + \lambda ||f||_{\Gamma}^2.$$

が得られた。次に興味があるのはこの最適化問題はユニークな解を持つのか、その解はどのような形で書けるのか (リプリゼンタ一一定理が成り立って Γ_{x_i} の線形和で書けてほしい) ということである。それに答えるのが次の定理である。

Theorem 2.2.(Adapted from G. Lever and S. Grünwald+ 2012) f^* が $\hat{\epsilon}_{\lambda}$ を \mathcal{H}_{Γ} で最小化するとする。このとき

$$f^* = \sum_{i=1}^n \Gamma_{x_i} c_i$$

とユニークに書ける。ここで係数 $\{c_i\}$, $c_i \in \mathcal{V}$ は以下の連立方程式のユニークな解である。

$$\sum_{i \leq n} (\Gamma(x_j, x_i) + \lambda \delta_{ji}) c_i = v_j, \quad 1 \leq j \leq n.$$

$\hat{\epsilon}_{\lambda}(f)$ のような形をした、カーネルベイズに対応する経験リスクを導出していくのが本論文の大筋である。

7 $\epsilon_s[\mu]$ が $\varepsilon[\mu]$ の上限であることの証明

Proof

$$\begin{aligned}
\varepsilon[\mu] &:= \sup_{\|h\|_{\mathcal{H}_Y} \leq 1} E_X[(E_Y[h(Y)|X]) - \langle h, \mu(X) \rangle_{\mathcal{H}_Y})^2] \\
&= \sup_{\|h\|_{\mathcal{H}_Y} \leq 1} E_X[(E_Y[\langle h, \psi(Y) \rangle_{\mathcal{H}_Y}|X] - \langle h, \mu(X) \rangle_{\mathcal{H}_Y})^2] \\
&\leq \sup_{\|h\|_{\mathcal{H}_Y} \leq 1} E_{X,Y}[\langle h, \psi(Y) - \mu(X) \rangle_{\mathcal{H}_Y}^2] \\
&\leq \sup_{\|h\|_{\mathcal{H}_Y} \leq 1} \|h\|_{\mathcal{H}_Y}^2 E_{X,Y}[|\psi(Y) - \mu(X)|_{\mathcal{H}_Y}^2] \\
&= E_{X,Y}[|\psi(Y) - \mu(X)|_{\mathcal{H}_Y}^2] = \epsilon_s[\mu].
\end{aligned}$$

□

8 Proposition 1 の証明

Proposition 1 (X, Y) は $\mathcal{X} \times \mathcal{Y}$ 上の確率変数とし, Y に対する prior を $\pi(Y)$, 尤度を $p(X|Y)$ とする. $\mathcal{H}_{\mathcal{X}}$ はカーネル $k_{\mathcal{X}}$, 特徴写像 $\phi(x)$ に対応する RKHS, $\mathcal{H}_{\mathcal{Y}}$ はカーネル $k_{\mathcal{Y}}$, 特徴写像 $\psi(y)$ に対応する RKHS とする. そして, $\phi(x, y)$ は $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ の特徴写像とする.

$\hat{\pi}_Y = \sum_{i=1}^{\ell} \tilde{\alpha}_i \psi(\tilde{y}_i)$ は π_Y の一致推定量とし, $\{(x_i, y_i)\}_{i=1}^n$ は尤度 $p(X|Y)$ を表現するサンプルとする.

このとき, $f(x, y) = |\psi(y) - \mu(x)|_{\mathcal{H}_Y}^2$ とおき, $f \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ を仮定すると,

$$\hat{\epsilon}_s[\mu] = \sum_{i=1}^n \beta_i |\psi(y_i) - \mu(x_i)|_{\mathcal{H}_Y}^2,$$

ただし, $\beta = (\beta_1, \dots, \beta_n)^T$ は $\beta = (G_Y + n\lambda I)^{-1} \tilde{G}_Y \tilde{\alpha}$, $(G_Y)_{ij} = k_{\mathcal{Y}}(y_i, y_j)$, $(\tilde{G}_Y)_{ij} = k_{\mathcal{Y}}(y_i, \tilde{y}_j)$, そして $\tilde{\alpha} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_{\ell})^T$.

Proof この証明は K. Fukumizu 2016. Kernel Bayes Rule. の Proposition 4 とほとんど同じにできる.

$\Phi_{X,Y} = (\phi(x_1, y_1), \dots, \phi(x_n, y_n))$ と置いたとき, $\hat{\mu}_{(X,Y)} = \Phi_{X,Y}\beta = \Phi_{X,Y}(G_Y + n\lambda I)^{-1}\tilde{G}_Y\tilde{\alpha}$ とかけることを示せばよい. これが示せれば, $\mathcal{H}_X \otimes \mathcal{H}_Y$ の再生性により,

$$\begin{aligned}\epsilon_s[\mu] &= \langle \hat{\mu}_{(X,Y)}, f \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \\ &= \langle \Phi_{X,Y}(G_Y + n\lambda I)^{-1}\tilde{G}_Y\tilde{\alpha}, f \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \\ &= \langle \Phi_{X,Y}\beta, f \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} \\ &= \sum_{i=1}^n \beta_i \|\psi(y_i) - \mu(x_i)\|_{\mathcal{H}_Y}^2.\end{aligned}$$

として示せる.

まず, $\hat{\mu}_{(X,Y)} = \hat{C}_{(X,Y)Y}(\hat{C}_{YY} + \lambda I)^{-1}\hat{\pi}_Y$ であることを思い出す. $h = (\hat{C}_{YY} + \lambda I)^{-1}\hat{\pi}_Y$ とおいて, このを直交成分で分解した

$$h = \sum_{i=1}^n a_i \psi(y_i) + h_\perp$$

を考える. h_\perp は h の $\text{span}(\psi(y_1), \dots, \psi(y_n))$ に直交する成分である. $(\hat{C}_{YY} + \lambda I)h = \hat{\pi}_Y$ を展開することにより,

$$\frac{1}{n} \sum_{i,j \leq n} a_i k_Y(y_i, y_j) \psi(y_j) + \lambda \left(\sum_{i \leq n} a_i \psi(y_i) + h_\perp \right) = \sum_{i \leq \ell} \tilde{\alpha}_i \psi(\tilde{y}_i)$$

が得られる. この両辺に, $\psi(y_k)|_{k=1}^n$ をそれぞれ掛けると

$$\frac{1}{n} G_Y^2 \mathbf{a} + \lambda G_Y \mathbf{a} = \tilde{G}_Y \tilde{\alpha} \Leftrightarrow \frac{1}{n} (G_Y + n\lambda I) G_Y \mathbf{a} = \tilde{G}_Y \tilde{\alpha} \Leftrightarrow \frac{1}{n} G_Y \mathbf{a} = (G_Y + n\lambda I)^{-1} \tilde{G}_Y \tilde{\alpha}$$

が得られる. よって, $\hat{\mu}_{(X,Y)}$ は

$$\hat{\mu}_{(X,Y)} = \frac{1}{n} \left[\sum_{i \leq n} \phi(x_i, y_i) \otimes \psi(y_i) \right] h = \frac{1}{n} \Phi_{X,Y} G_Y \mathbf{a} = \Phi_{X,Y} (G_Y + n\lambda I)^{-1} \tilde{G}_Y \tilde{\alpha}$$

と書け, 示せた. □

9 Proposition 2 の証明

Proposition 2 一般性を失わず、任意の i で $\beta_i^+ \neq 0$ と仮定する。 $\mu \in \mathcal{H}_K$ として、 \mathcal{H}_K のカーネルを $K(x_i, x_j) = k_{\mathcal{X}}(x_i, x_j)\mathcal{I}$ と選ぶ。ただし、 $\mathcal{I} : \mathcal{H}_K \rightarrow \mathcal{H}_K$ は恒等写像とする。すると、

$$\hat{\mu}_{\lambda,n}(x) = \Psi(K_X + \lambda_n \Lambda^+)^{-1} K_{:x}$$

ただし、 $\Psi = (\psi(y_1), \dots, \psi(y_n))$, $(K_X)_{ij} = k_{\mathcal{X}}(x_i, x_j)$, $\Lambda^+ = \text{diag}(1/\beta_1^+, \dots, 1/\beta_n^+)$, $K_{:x} = (k_{\mathcal{X}}(x, x_1), \dots, k_{\mathcal{X}}(x, x_n))^T$ で λ_n は正の正則化定数である。

Proof $\beta_i^+ = 0$ となるときは、対応するデータ (x_i, y_i) を除いて議論しても結果は変わらない。よって一般性を失わず、任意の i で $\beta_i^+ \neq 0$ と仮定できる。

$\mu = \mu_0 + g$, ただし $\mu_0 = \sum_{i=1}^n K_{x_i} c_i$ とおく。これを $\hat{\epsilon}_{\lambda,n}[\mu]$ に代入すると、

$$\begin{aligned} \hat{\epsilon}_{\lambda,n}[\mu] &= \sum_{i=1}^n \beta_i^+ \|\psi(y_i) - \mu(x_i)\|_{\mathcal{H}_Y}^2 + \lambda_n \|\mu\|_{\mathcal{H}_K}^2 = \sum_{i=1}^n \beta_i^+ \|\psi(y_i) - (\mu_0(x_i) + g(x_i))\|_{\mathcal{H}_Y}^2 + \lambda_n \|\mu_0 + g\|_{\mathcal{H}_K}^2 \\ &= \sum_{i=1}^n \beta_i^+ \|\psi(y_i) - \mu_0(x_i)\|^2 + \lambda_n \|\mu_0\|^2 + \sum_{i=1}^n \beta_i^+ \|g(x_i)\|^2 + \lambda_n \|g\|^2 + 2\lambda_n \langle \mu_0, g \rangle - 2 \sum_{i=1}^n \beta_i^+ \langle g(x_i), \psi(y_i) - \mu_0(x_i) \rangle. \end{aligned}$$

と展開できる。

ここで、各 i に対して、 $\psi(y_i) - \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j)c_j = \frac{\lambda_n}{\beta_i^+}c_i$ と推測してみる。実際これが成り立つとすれば、 $\hat{\epsilon}_{\lambda,n}[\mu]$ の後半 2 項は

$$\lambda_n \langle \mu_0, g \rangle - \sum_{i=1}^n \beta_i^+ \langle g(x_i), \psi(y_i) - \mu_0(x_i) \rangle = 0$$

となり、

$$\hat{\epsilon}_{\lambda,n}[\mu] = \hat{\epsilon}_{\lambda,n}[\mu_0] + \sum_{i=1}^n \beta_i^+ \|g(x_i)\|^2 + \lambda_n \|g\|^2 \geq \hat{\epsilon}_{\lambda,n}[\mu_0]$$

が得られる。これより、 $\psi(y_i) - \sum_{j=1}^n k_{\mathcal{X}}(x_i, x_j)c_j = \frac{\lambda_n}{\beta_i^+}c_i$ の関係を満たす c_i による $\mu_0 = \sum_{i=1}^n K_{x_i} c_i$ が解であることが分かる。この関係式を行列で表現すると、 $(K_X + \lambda_n \Lambda^+)c = \Psi$ であり、

$$\mu_0(x) = \sum_{i=1}^n k_{\mathcal{X}}(x, x_i)c_i = \Psi(K_X + \lambda_n \Lambda^+)^{-1} K_{:x}$$

が得られた。 □