カーネル法とガウス過程の関係について＆近況

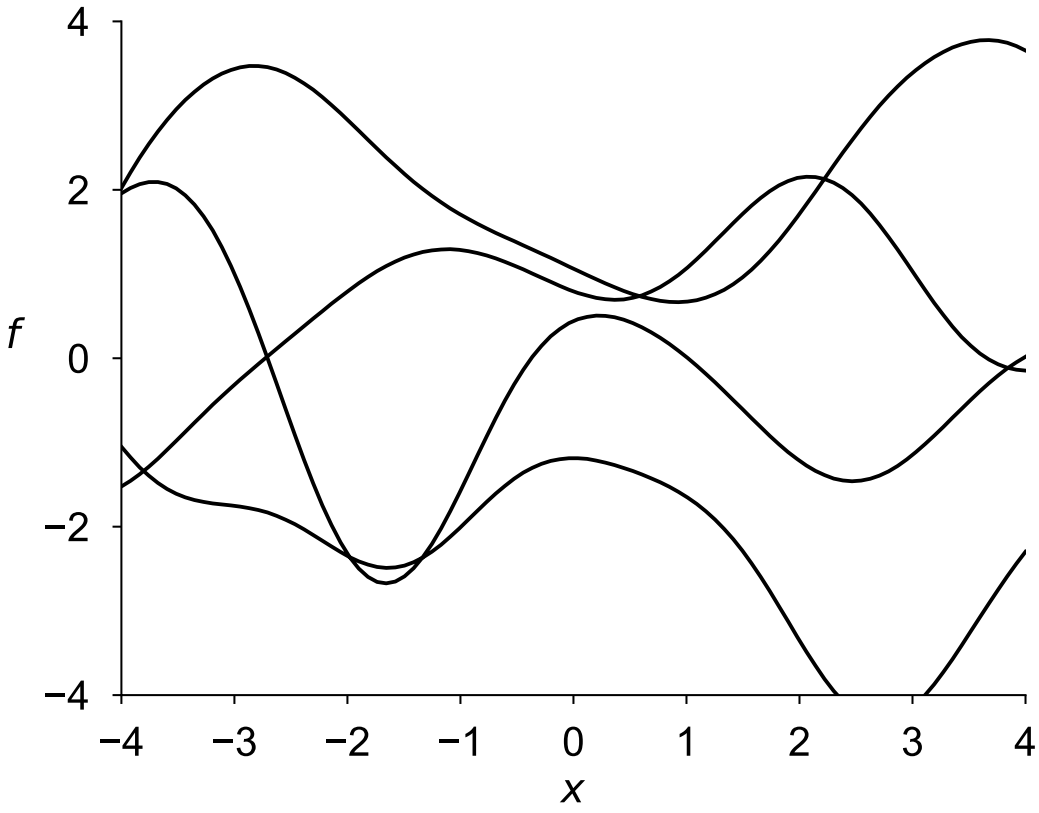
持橋大地
統計数理研究所

SVM 2018

2018-9-14 (金)

ガウス過程とは

- ランダムな関数を生成する確率過程
  (関数の確率分布)



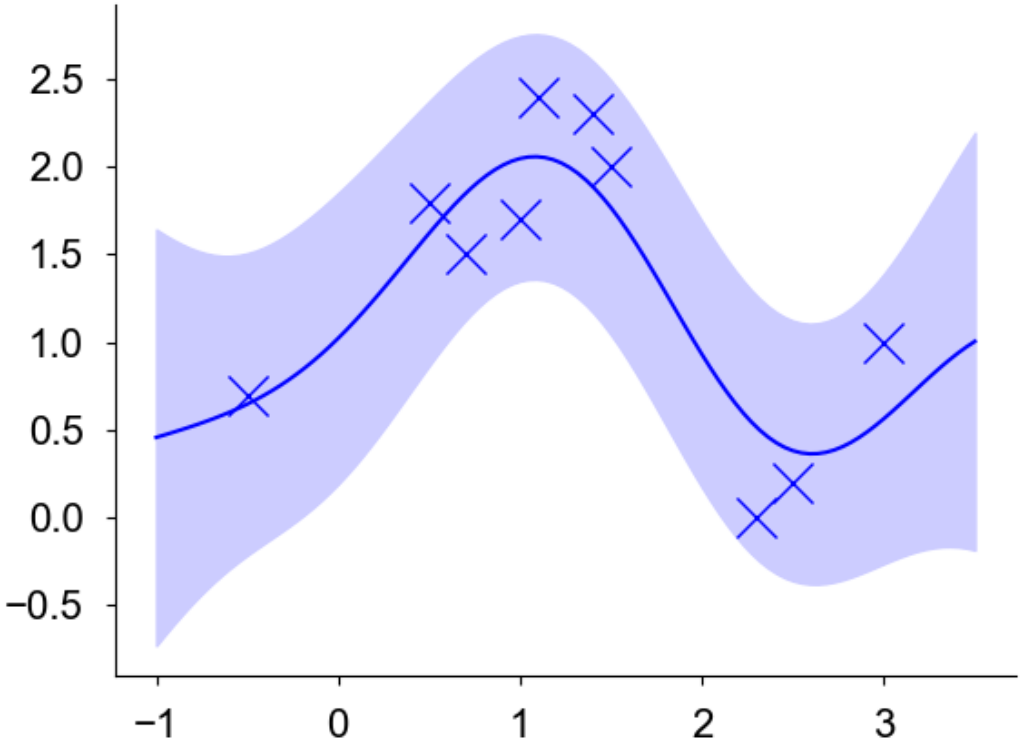
$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

$K_{ij}=k(x_i,x_j)$
ベイズ的な
カーネルマシン

詳しくは、
ガウス過程本で

カーネルリッジ回帰とGP回帰

- カーネルを用いた回帰モデル
- 結果はほとんど同じ→ガウス過程回帰には、分散が存在



カーネルリッジ回帰とGP回帰 (2)

- カーネルリッジ回帰の目的関数

$$\hat{f} = \arg\min \frac{1}{N} \sum_{n=1}^{N} (f(x_n) - y_n)^2 + \lambda ||f||^2$$

- 解は、

$$\hat{f}(x) = \mathbf{k}(\mathbf{K} + n\lambda \mathbf{I}_N)^{-1}\mathbf{y} = \sum_{n=1}^{N} \alpha_n k(x, x_n)$$

- ガウス過程回帰の事後分布

$$\mathbf{f} \,|\, \mathbf{X}, \mathbf{y} \sim \mathcal{N}\left(\mathbf{k}(\mathbf{K}+\sigma^2\mathbf{I}_N)^{-1}\mathbf{y}, k - \mathbf{k}^T(\mathbf{K}+\sigma^2\mathbf{I}_N)^{-1}\mathbf{k}\right)$$

  - カーネルリッジ回帰の平均＋事後分布の分散

カーネル埋め込み

- 確率分布*P*とカーネル*k*(*x*,*x*')について、
  *P*の*k*によるカーネル埋め込みは

$$\mu_P = \int k(\cdot, x) P(dx)$$

- 独立性の尺度であるHSICで使用：
  確率変数X,Yについて、

$$\mathrm{HSIC} = ||\mu_{P_{XY}} - \mu_{P_X P_Y}||^2$$

  - カーネルで経験分布から計算できる

HSIC

- HSICは、カーネルで経験分布から計算できる

$$\mathrm{HSIC} = ||\mu_{P_{XY}} - \mu_{P_X P_Y}||^2 = \frac{1}{N^2}\mathrm{tr}(\mathbf{KHLH})$$

- 非線形な相互情報量: “Pointwise”にしたもの
  → Pointwise HSIC (PHSIC) (Yokoi+ EMNLP2018)

$$\mathrm{PHSIC}(x, y) = \mathbb{E}_{(x', y')}[\widetilde{k}(x, x')\widetilde{\ell}(y, y')]$$

  – Yokoi&Mochihashi (IJCAI 2017)から派生

カーネル法とガウス過程

- Kanagawa+ (2018.7)
  - 超労作！！

Gaussian Processes and Kernel Methods:
A Review on Connections and Equivalences

Motonobu Kanagawa[1], Philipp Hennig[1],
Dino Sejdinovic[2], and Bharath K Sriperumbudur[3]

[1]University of Tübingen and Max Planck Institute for Intelligent Systems,
Max-Planck-Ring 4, 72076 Tübingen, Germany
{motonobu.kanagawa,ph}@tue.mpg.de

[2]Department of Statistics, University of Oxford,
24-29 St Giles', Oxford OX1 3LB, UK
dino.sejdinovic@stats.ox.ac.uk

[3]Department of Statistics, Pennsylvania State University,
University Park, PA 16802, USA
bks18@psu.edu

HSICとガウス過程

- $\mathbf{f} \sim \mathrm{GP}(0, k),\ \mathbf{g} \sim \mathrm{GP}(0, \ell)$ のとき、

$$\mathrm{HSIC}(X, Y) = \mathbb{E}_{\mathbf{f},\mathbf{g}}[\mathrm{cov}^2(f(X), g(Y))]$$

$$\mathrm{cov}(f(X), g(Y)) =$$

$$\mathbb{E}_{X,Y}[(f(X) - \langle f(X)\rangle)(g(Y) - \langle g(Y)\rangle)|\mathbf{f}, \mathbf{g}]$$

(Kanagawa+ 2018)

カーネル埋め込みとは?

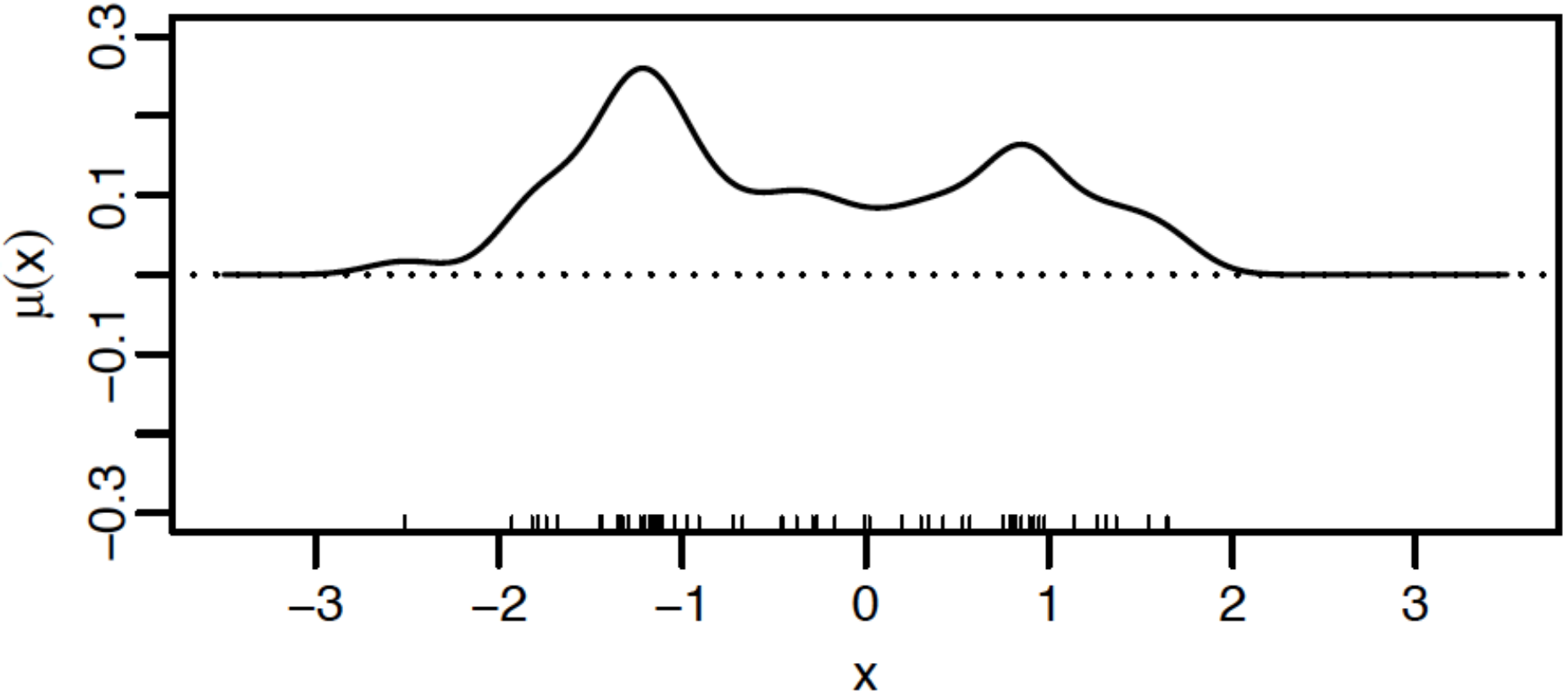
$$\mu_P = \int k(\cdot, x) P(dx)$$

- カーネル埋め込み $\mu_P : \mathcal{X} \to \mathbb{R}$ の姿は、そもそも何?
- μ_P は $\mathcal{X}$ 上の関数！(測度のようなもの)
- Empirical estimator:

$$\widehat{\mu}_P = \sum_{n=1}^{N} k(\cdot, x_n)$$

  - ほとんどカーネル密度推定! (の現代版)

μ_P の姿



- Flaxman+ (2016)の図より

“Bayesian Learning of Kernel Embeddings”

- Flaxman+ (UAI 2016)
- カーネル埋め込みをガウス過程を使って確率モデルとして捉えることで、ハイパーパラメータが学習できる！
- 従来手法：Median heuristic (Takeuchi 06他)
  - データの分散の中央値をカーネルの分散とする
  - 実は、かなり性能が悪い場合がある
  - ヒューリスティックに決めているだけ

カーネル埋め込み μ_P の分布

- リプリゼンター定理によれば、$\mathbf{f} \sim \mathrm{GP}(0, k)$ について

$$\mathbf{f} = \sum_{n=1}^{N} \alpha_n k(\cdot, x_n)$$

$$= \sum_{n=1}^{N} z_n \lambda_n^{1/2} \phi_n \qquad z_n \sim \mathcal{N}(0, 1)$$

- これは直感的にも自然：ガウス過程fは

$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K}) = \mathcal{N}(0, (\lambda\mathbf{\Phi})(\lambda\mathbf{\Phi})^T)$$

から導かれる

カーネル埋め込み μ_P の分布 (2)

- よって、μ_P はガウス過程に従っている
  → 経験分布から得られる $\widehat{\mu}_P$ は、μ_P の近似
- 尤度関数をどうするか?
  → 最も単純に、ガウスノイズとする
  (中心極限定理)

$$\widehat{\mu}_P \,|\, \mu_P \sim \mathcal{N}(\mu_P, \tau^2/N)$$

Marginalized likelihood

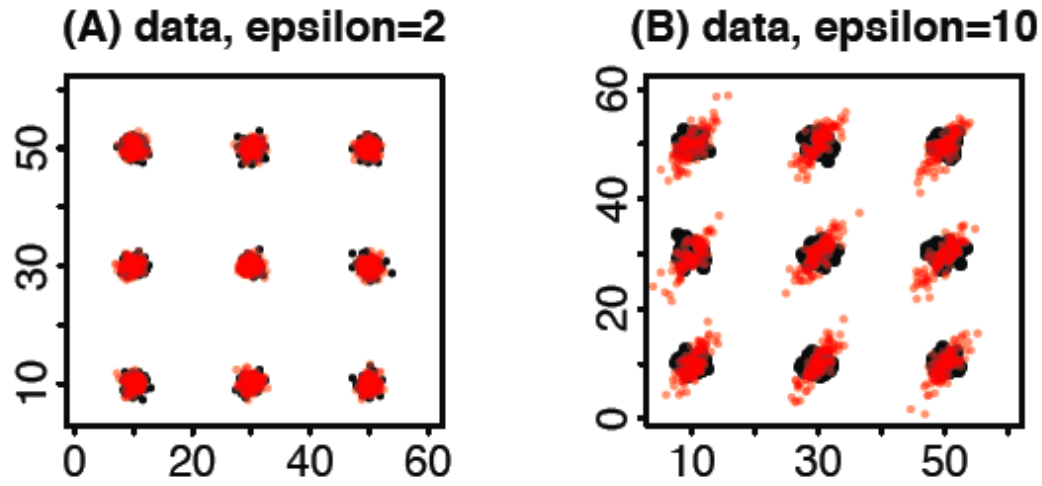
- μ_Pを積分消去して、データ x1..xN の尤度が以下のように得られる
  → カーネルのハイパーパラメータθを最適化

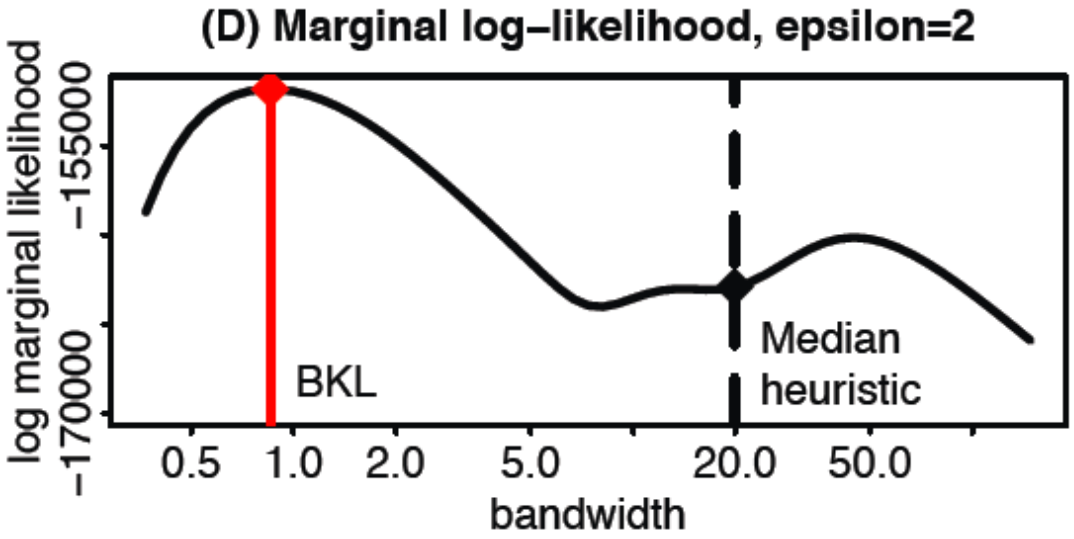
$$p(x_1,\ldots,x_n|\theta) = \int p(x_1,\ldots,x_n|\mu_\theta,\theta)p(\mu_\theta|\theta)d\mu_\theta$$

$$= \int \mathcal{N}\left(\phi_{\mathbf{z}}(\mathbf{x}); \mathbf{m}_\theta(\mathbf{z}), \tau^2 I_{mn}\right) \left[\prod_{i=1}^{n} \gamma_\theta(x_i)\right] p(\mu_\theta|\theta)d\mu_\theta$$

$$= \mathcal{N}\left(\phi_{\mathbf{z}}(\mathbf{x}); \mathbf{0}, \mathbf{1}_n\mathbf{1}_n^\top \otimes R_{\theta,\mathbf{zz}} + \tau^2 I_{mn}\right) \prod_{i=1}^{n} \gamma_\theta(x_i).$$

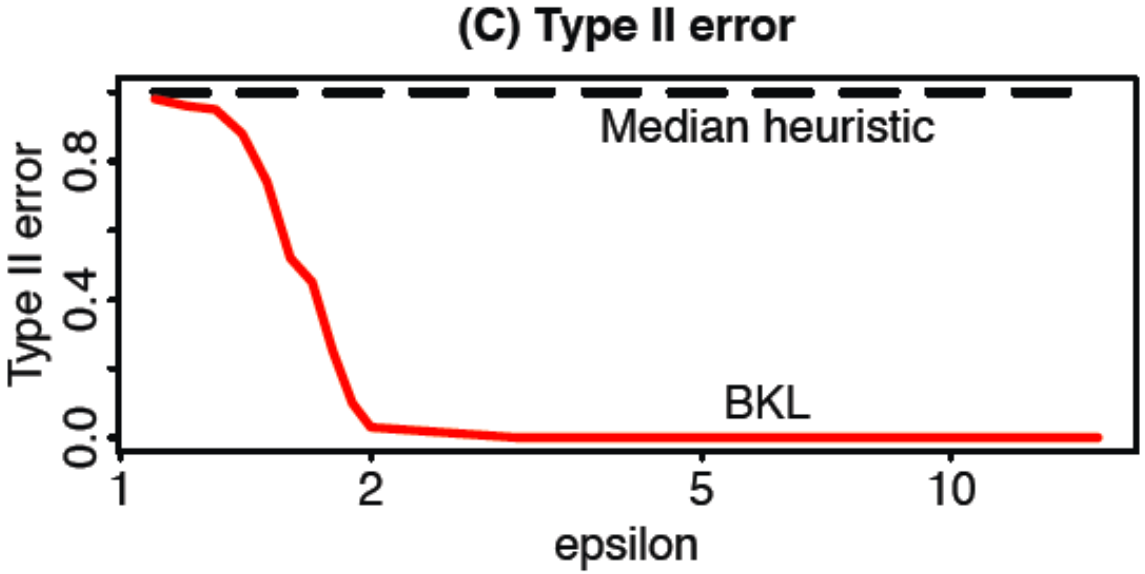
HSICによる検定

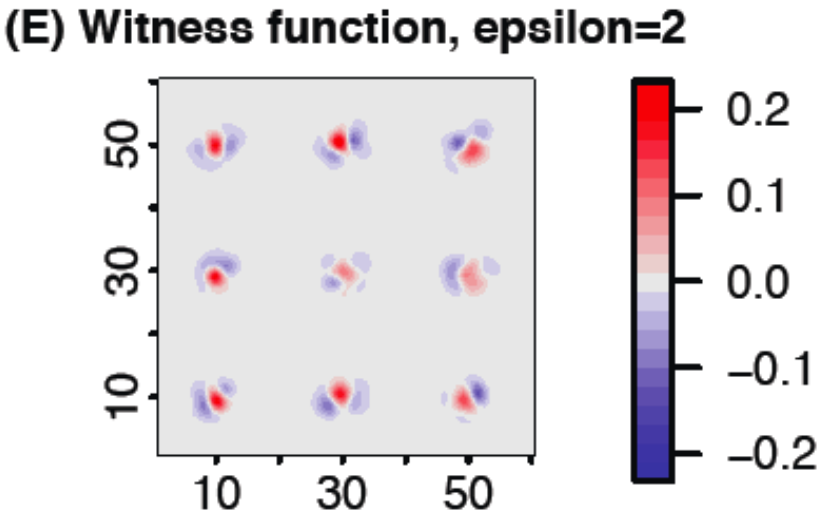




- 赤の点の分布が、黒の点の分布と違うことを検定
- εによって難しさが異なる
- Median heuristicは大きすぎるカーネルのバンド幅
→検定失敗

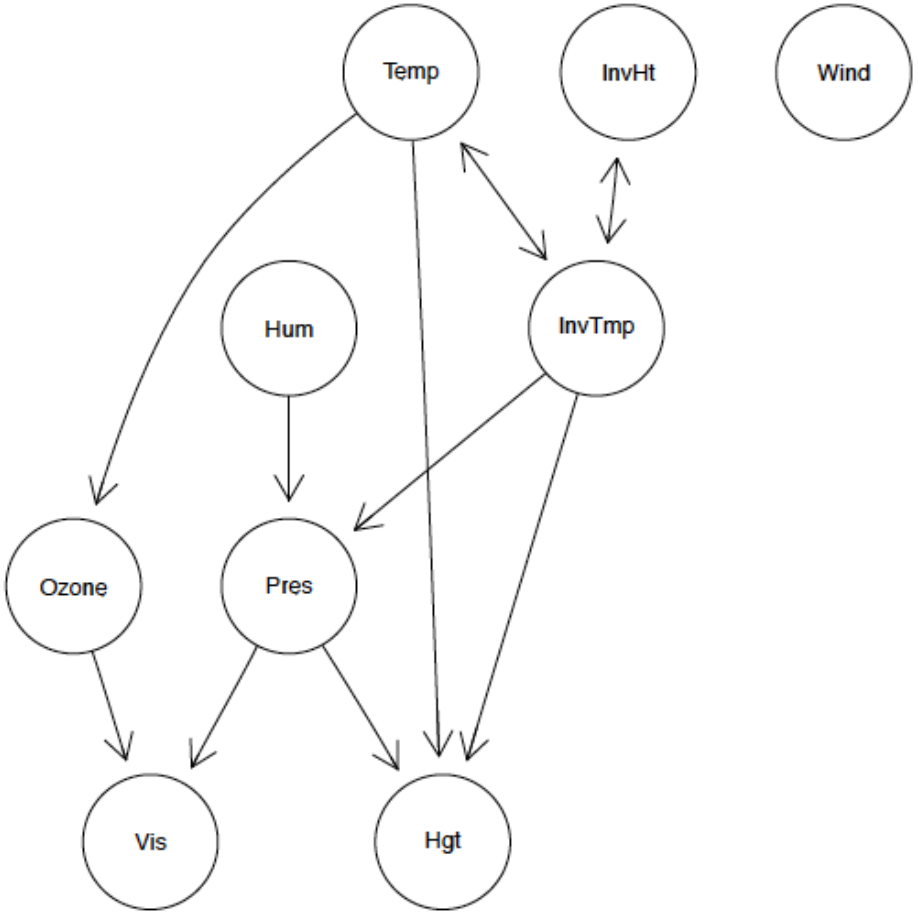
HSICによる検定





- ε=2でも、検定誤差はほぼ0
- Median heuristicは常に誤り
- “Witness function” $\mu_P - \widehat{\mu}_P$ で様子がわかる (分散もわかる)

HSICによる因果推論



- HSICによって、事象の独立性を検定
  → 因果グラフが得られる
- カーネルが定義できれば、複雑な対象でも検定可能
- 結果はカーネルのハイパーパラメータに依存→学習できる

まとめ

- カーネル法とガウス過程は、深い関係
  - Kanagawa+ (2018)で色々な考察
- カーネル法では、
  - ハイパーパラメータが推定できない
  - 確率モデルと結びつけることができない

ガウス過程として解釈すれば可能に！

- HSIC、PHSICをガウス過程で解釈することで、どう拡張できるか？