Kernel conditional density operators

read: how to solve your GP problems

Ingmar Schuster, Mattes Mollenhauer, Stefan Klus, Krikamol Muandet

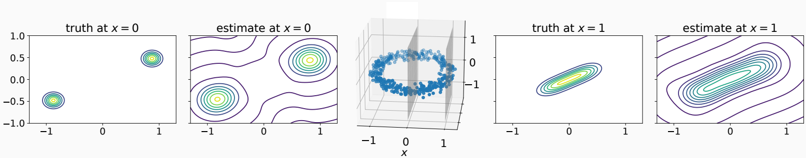
July 11, 2019

Zalando Research

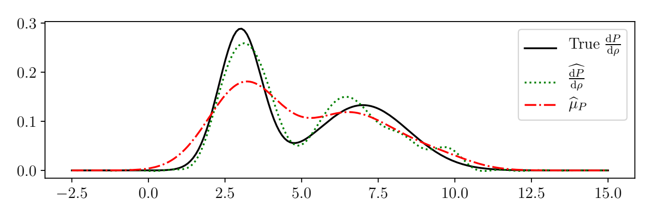
Overview

Overview (1)

- the conditional density operator (CDO) is a kernel-based model for estimation of conditional densities
- multi-modal, multivariate output densities improve over vanilla Gaussian Processes
- experimental performance is competitive with neural conditional density models



Overview (2)



- to derive CDO, focus on reconstructing densities from their kernel mean embedding
- clarify when density $p \in L_2(\rho)$ has RKHS representer $\tilde{p}$ so

$$p = \tilde{p} \text{ holds } \rho\text{-almost everywhere}$$

- finite sample bounds on stochastic error
- guidelines on regularization

Reproducing Kernel Hilbert spaces and embeddings of distributions

RKHS overview

- continuous, symmetric psd kernel $k\colon \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ inducing an RKHS H

$$\langle k(x,\cdot), f\rangle = f(x) \text{ for } f \in H$$

- RKHS element form $\sum_{i=1}^{\infty} \alpha_i k(x_i, \cdot) \in H$ for $\alpha_i \in \mathbb{R}, x_i \in \mathbb{X}$
- RKHSs are vector spaces of functions

$$g, f \in H \text{ and } \alpha, \beta \in \mathbb{R} \Rightarrow \alpha f + \beta g \in H$$

- operators on RKHSs provide powerful tool (think matrices on real vector spaces)

Algebra of distributions

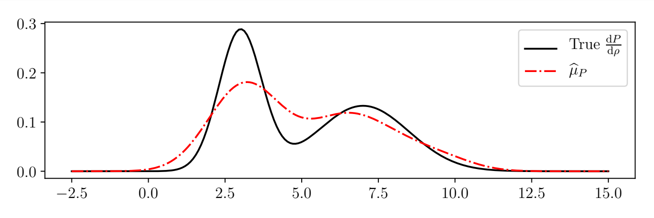
Our view

- RKHS operators provide an algebra of distributions
- given dataset, we *just solve* for distribution/density of interest

Example: distribution embedding, covariance operator and density

$$\mu_{\mathbb{P}} = C_\rho u$$

Embedding of distributions



- kernel mean embedding $\mu_{\mathbb{P}} := \int_{\mathbb{X}} \phi(x)\,\mathrm{d}\mathbb{P}(x) \in H$ for finite measure $\mathbb{P}$ with $\phi(x) := k(x, \cdot)$
- stores mean for linear, higher moments for polynomial kernel
- injective for characteristic kernels (stores all distribution information)

Covariance and integral operator

- covariance operator on H

$$C_\rho := \int_{\mathbb{X}} \phi(x) \otimes \phi(x) \, \mathrm{d}\rho(x)$$

  is well-defined operator s.t. for $f \in H$ by reproducing property

$$C_\rho f = \int_{\mathbb{X}} \phi(x) \langle \phi(x), f \rangle \, \mathrm{d}\rho(x) = \int_{\mathbb{X}} \phi(x) f(x) \, \mathrm{d}\rho(x)$$

- strong connection to integral operator on $L_2(\rho)$

$$(\mathcal{E}_\rho g) := \int \phi(x) g(x) \, \mathrm{d}\rho(x)$$

  for $g \in L_2(\rho)$
- both share eigenvalues & -functions up to rescaling

Fitting from data

To fit these objects from data, we use the standard approximations

$$C_\rho \approx \frac{1}{M} \sum_{i=1}^{M} \phi(x_i) \otimes \phi(x_i)$$

for $x_i \sim \rho$ and

$$\mu_{\mathbb{P}} \approx \frac{1}{N} \sum_{j=1}^{N} \phi(x_j)$$

for $x_i \sim \mathbb{P}$.

Example for algebra of distributions

- if $u \in H$ is density of $\mathbb{P}$, then

$$C_\rho u = \int_{\mathbb{X}} \phi(x) u(x) \, \mathrm{d}\rho(x)$$

- Whats the density of $\mathbb{P}$? If ρ is Lebesgue measure then $u = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$
- thus for the kernel mean embedding

$$\mu_{\mathbb{P}} = \int_{\mathbb{X}} \phi(x) \, \mathrm{d}\mathbb{P}(x) = \int_{\mathbb{X}} \phi(x) \frac{\mathrm{d}\mathbb{P}(x)}{\mathrm{d}\rho(x)} \mathrm{d}\rho(x) = C_\rho u$$

- estimating $\mu_{\mathbb{P}}$, C_ρ from samples, we can estimate u!
  (if it's in H)

Cross covariance and conditional mean operators

- cross covariance operator for joint distribution $\mathbb{P}_{XY}$ is

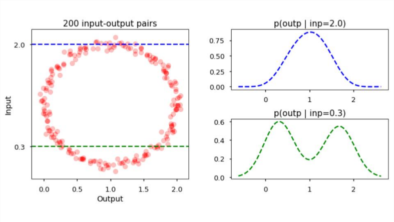
$$C_{YX} := \int_{\mathbb{X}} \psi(y) \otimes \phi(x) \, \mathrm{d}\mathbb{P}_{XY}(x, y)$$

- use another kernel ℓ for space $\mathbb{Y}$
  inducing RKHS F and $\psi(y) := \ell(y, \cdot)$
- conditional mean operator $\mathcal{U}_{Y|X} = C_{YX} C_X^\dagger$ maps from H to F

$$\mu_{\mathbb{P}_{Y|X=x'}} = \mathcal{U}_{Y|X} k(x', \cdot)$$

  ($C_X^\dagger$ being a pseudoinverse)
- i.e. we can get the mean embedding of the conditional distribution on output Y given input $X = x'$

Conditional mean operator example



if input is x, output y:

$$\widehat{\mathcal{U}}_{Y|X} k(x^*, \cdot) = \begin{bmatrix} \ell(y_1, \cdot) \\ \vdots \\ \ell(y_N, \cdot) \end{bmatrix}^\top K_X^{-1} \begin{bmatrix} k(x_1, x^*) \\ \vdots \\ k(x_N, x^*) \end{bmatrix}$$

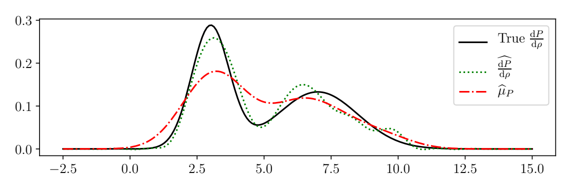
- estimate output embedding using only samples, where K_X is gram matrix for kernel k and all $x_1, \ldots, x_N$

Kernel conditional density operators

Paper gist

- we want the conditional output density rather than embedding
- even if it is not in output RKHS F
- previous attempts of reconstructing densities
  - did not use that $u = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$
  - only have theory when everything is in F
  - use inverses $C_{\rho_y}^{-1}$ that don't exist
  - dont have error bounds for reconstruction

Density reconstruction



- density p of distribution $\mathbb{P}$ is $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$ if ρ is Lebesgue measure (Radon–Nikodym derivative wrt Lebesgue)
- we suggest to reconstruct density by using the fact that $C_\rho^\dagger \mu_{\mathbb{P}} = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho}$ ρ-almost everywhere
- applied to conditional mean operator results in conditional density operator (CDO)

Plan

- we concentrate on density reconstruction for the theory
- CDO results directly follow from this
- plan
  - show that reconstruction is unique
  - derive conditions for RKHS representative in correct $L_2(\rho)$ equivalence class
  - derive conditional density operator and clarify how it reconstructs densities

Uniqueness of reconstruction

Uniqueness (no solutions other then the density of interest)
Let $\mathbb{P} \ll \rho$ and $p := \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho} \in L_2(\rho)$, also let there be an RKHS function in the equivalence class of p.
Then solving $C_\rho u = \mu_{\mathbb{P}}$ for $u \in H$ uniquely yields the solution $u^\dagger$ and $u^\dagger = p$ holds ρ-almost everywhere.

Conditions for existance of representer

Representer (When does the density have an RKHS representer?)
Let $(\lambda_i, e_i)_{i \in I}$ be the eigenvalue/eigenfunction pairs of $\mathcal{E}_\rho$. If

$$\left(\langle p,\, e_i \rangle_{L_2(\rho)} \lambda_i^{-1/2} \right)_{i \in I} \in \ell_2(I),$$

then $p \overset{\rho\text{-a.e.}}{=} \tilde{p}$ for some $\tilde{p} \in H$.

Conditional density operator (CDO)

Conditional density operator

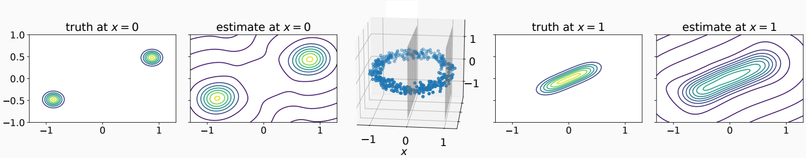
Assume $\mathbb{P}_y(\cdot) = \int_{\mathbb{X}} \mathbb{P}_{Y|X=x}(\cdot)\,\mathrm{d}\mathbb{P}(x)$ admits a Radon–Nikodym derivative with respect to the reference measure ρ_y called $p_y \in L_2(\rho_y)$. Assume the conditonal mean operator $\mathcal{U}_{Y|X}$ for $\mathbb{P}_{Y|X}$ exists.

Then

$$p_y \overset{\rho_y-\text{a.e.}}{=} \mathcal{A}_{Y|X}\mu_{\mathbb{P}}$$

for $\mathcal{A}_{Y|X} = C_{\rho_y}^{\dagger}\mathcal{U}_{Y|X} = C_{\rho_y}^{\dagger} C_{YX} C_X^{\dagger}$.

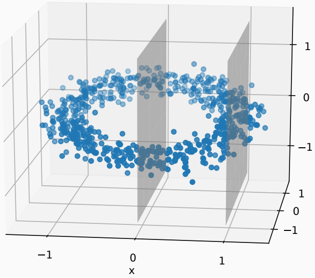
Advantages of CDOs



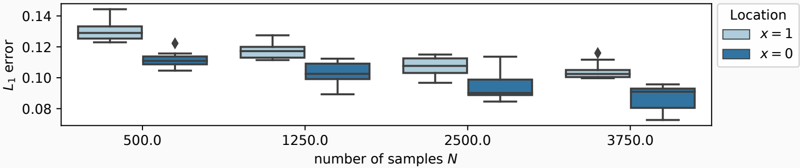
- can represent multimodal, multivariate output densities unlike vanilla GPs
- output can be mixture of Student-t, mixture of Laplace, etc
- remedies on the GP side
  - multimodality could be achieved with mixtures of GPs
  - multi-output GPs, for example using vector values RKHS
- experimentally competitive with neural conditional density models, while providing theoretical guarantees

Experimental results

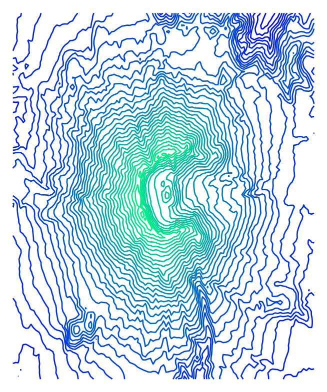
Toy example: Gaussian Donut



- embedd 2d circle in 3d ambient space, tilt it around y-axis
- pick 50 equidistant points on circle for gaussian location mixture
- draw 50 samples from each mixture component
- experimentally check L_1 convergence at two locations



Rough terrain reconstruction (1)



- rough terrain reconstruction for robotics and navigation [1, 2]
- estimate altitude for unseen logitude, latitude on a map
- reproduce GP experiment from [3] using Mount St Helens data
- a random 90% split of the data as training, the rest as test
- compute scaled mean absolute error (SMAE)

Rough terrain reconstruction (2)

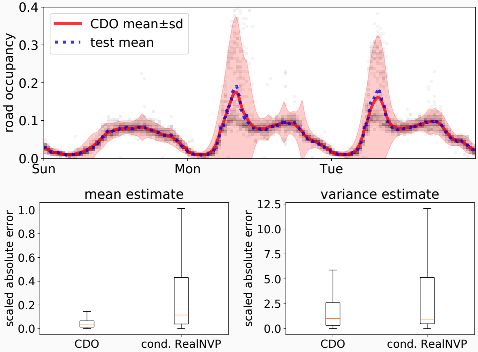
- use only Gaussian kernels for all methods
- GP lengthscale is optimized wrt training marginal likelihood
- CDO input lengthscale chosen based on median heuristic
- CDO reference samples given by equidistant grid points
- output lengthscale from distance between adjacent grid points

SMAE *GP*: 0.0358 ± 0.00062 *CDO*: 0.0269 ± 0.00055

Traffic prediction (1)

- predict occupancy of freeways in bay area
- encoded as number between 0 and 1 for 963 locations
- *day of week* and *time of day* in 10 minute intervals as predicting variables
- each dow occured 32 times in training, 20 times in test data
- thus $32 \times 144 \times 7 = 32\,256$ training examples (i.e. a lot for exact kernel methods)
- fitted a CDO using Gaussian output kernels, lengthscale based on distance of adjacent grid points
- Laplacian input kernels provided much smoother estimates compared to Gaussians

Traffic prediction (2)



- compare to conditional RealNVP using the same predicting variables (no temporal structure)
- measure scaled absolute error of estimating mean and variance of different time points in test data
- CDO clearly outperforms RealNVP based model

Nonasymptotic error bounds

Tikhonov solution

$$\mathcal{A}_{Y|X} = C_{\rho_y}^{\dagger} C_{YX} C_X^{\dagger}$$

- one way of approximating pseudoinverses is Tikhonov regularization

$$C_\rho^{\dagger}\mu_{\mathbb{P}} \approx (C_\rho + \alpha I_H)^{-1}\mu_{\mathbb{P}} = u_\alpha$$

- assume N data samples, M reference measure samples
- decompose total error into deterministic and stochastic parts

$$\left\| u^{\dagger} - \hat{u} \right\|_H \leq \left\| u^{\dagger} - u_\alpha \right\|_H + \| u_\alpha - \hat{u} \|_H$$

Stochastic error bounds

- stochastic error of the pseudoinverse solution u_α satisfies

$$\begin{aligned}\Pr&\left[\|u_\alpha - \hat{u}\|_H \leq \frac{M^{-2b}}{\alpha^2}(\|\mu_{\mathbb{P}}\|_H + N^{-2a}) + \frac{N^{-2a}}{\alpha}\right] \\ &\geq \left[1 - 2\exp\left(-\frac{N^{1-2a}}{8c^2}\right)\right]\left[1 - 2\exp\left(-\frac{M^{1-2b}}{8c^4}\right)\right]\end{aligned} \tag{1}$$

  independent of problem dimension
- we are free to choose $a, b \in (0, 0.5)$
- $c < \infty$ is a constant depending on kernel and domain

Regularization choice

- we can derive a principled regularization scheme
- guarantees convergence and yields tight bound
  $\Pr\left[\|u_\alpha - \hat{u}\|_H \le \frac{M^{-2b}}{\alpha^2}(\|\mu_{\mathbb{P}}\|_H + N^{-2a}) + \frac{N^{-2a}}{\alpha}\right]$
- pick $a, b \in (0, 0.5)$ and $c' \in (0, 1)$ and set

$$\widetilde{\alpha}(M, N) = \max(M^{-b}, N^{-2a})^{c'}$$

- smaller c' implies
  - larger approximation error (i.e. bias)
  - tighter bounds on the stochastic error

Summary and outlook

Summary (1)

- general reconstruction of densities by using $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\rho} = C_\rho^\dagger \mu_\mathbb{P}$
- clarify when a density $p \in L_2(\rho)$ has an RKHS representer
- our method uniquely reconstructs RKHS representer $\tilde{p}$ and

$$p = \tilde{p} \text{ holds } \rho\text{-almost everywhere}$$

- finite sample bounds on stochastic error
- guidelines on regularization

Summary (2)

- construct the conditional density operator from this
- kernel-based method comparable mostly to GPs but multivariate, multimodal
- good experimental performance compared to conditional neural density models and GPs

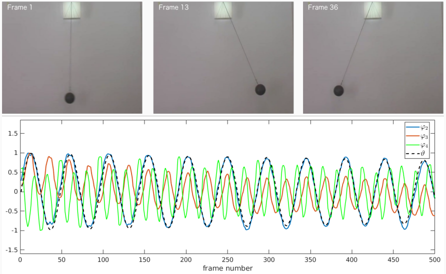
Outlook: closed form posterior densities

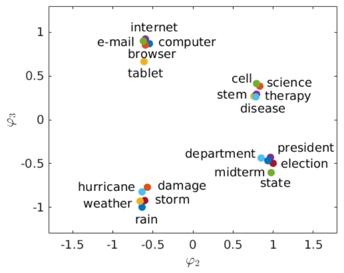
- in Bayesian setting generate θ_i from prior, $x_i \mid \theta_i$ from likelihood
- conditional mean operator $\widehat{\mathcal{U}}_{X|\Theta}$ fitted with artificial data
- given actual observed data, posterior embedding is

$$\widehat{\mathcal{U}}_{X|\Theta}\, \widehat{\mu}_{\text{observed}}$$

  (i.e. closed form fit linear in data set size)
- now, we can also get the posterior density directly!

Outlook: data exploration





- SVD of conditional density operator as possible data exploration tool
- as demonstrated before for dynamical systems and eigendecomposition [4]

Thank you!

References

[1] Raia Hadsell, J Andrew Bagnell, Daniel Huber, and Martial Hebert. Space-carving kernels for accurate rough terrain estimation. *The International Journal of Robotics Research*, 29(8):981–996, 2010.

[2] David Gingras, Tom Lamarche, Jean-Luc Bedwani, and Érick Dupuis. Rough terrain reconstruction for rover motion planning. In *2010 Canadian Conference on Computer and Robot Vision*, pages 191–198. IEEE, 2010.

[3] David Eriksson, Kun Dong, Eric Lee, David Bindel, and Andrew G Wilson. Scaling Gaussian process regression with derivatives. In *Advances in Neural Information Processing Systems*, pages 6867–6877, 2018.

[4] S. Klus, I. Schuster, and K. Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. *arXiv preprint arXiv:1712.01572*, 2017.