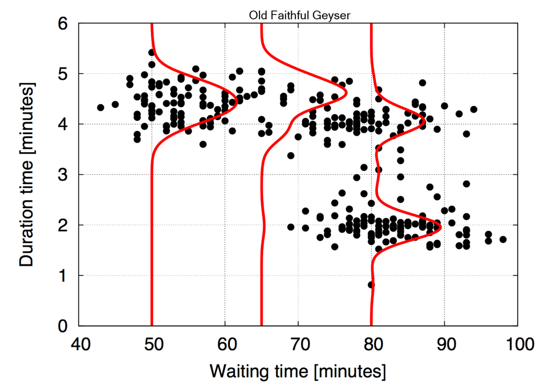
Kernel Conditional Exponential Family

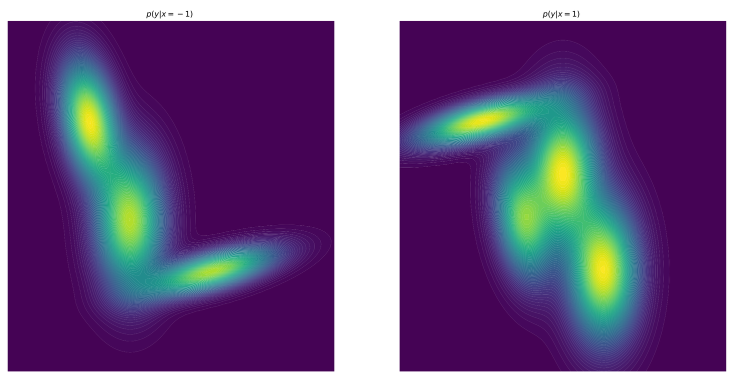
Michael Arbel and Arthur Gretton
Gatsby Computational Neuroscience Unit

Learning Conditional Distributions

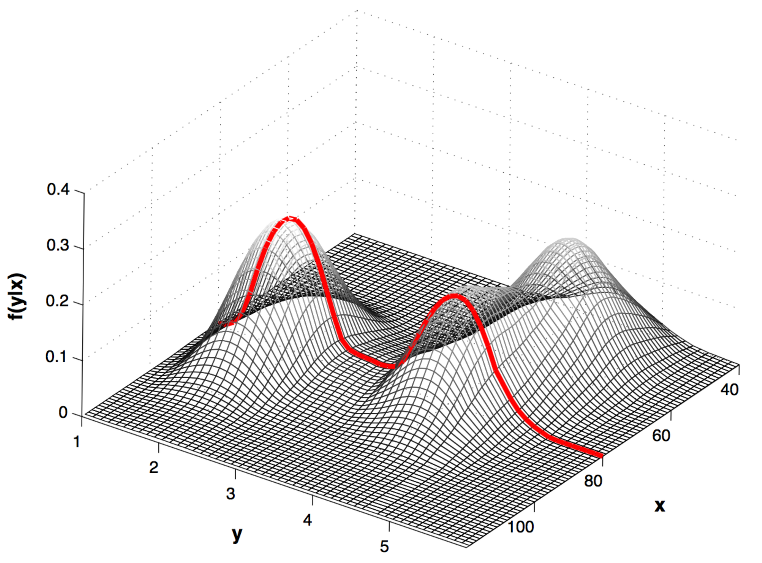
Goal: Learning conditional densities in a non-parametric fashion.



Number of modes can vary



Densities can be heteroscedastic



Density ratios are often simpler to learn than the full joint

Contribution:
- ✓ A particular form of Conditional Exponential Family based on vector valued RKHS.
- ✓ A method for approximating conditional densities using the KCEF with statistical guarantees.

Kernel Exponential Family (Sriperumbudur et al. (2017))

Idea: Parametrize densities with functions in an RKHS $\mathcal{G}$ with kernel k

$$p_\theta(y) = q_0(y)e^{\langle \theta, k(y,.)\rangle_{\mathcal{G}} - A(\theta)} \qquad A(\theta) = \log \int q_0(y)e^{\langle \theta, k(y,.)\rangle_{\mathcal{G}}}dy$$

θ is the natural parameter and $k(y,.)$ the sufficient statistic. Both are 'infinite' dimensional vectors.

- ✓ Richer than finite dimensional exponential family
- × Intractable log-partition function $A(\theta)$: MLE is hard to compute.
- ✓ Learning via Score-Matching (Hyvarinen (2005))
- ✓ Good statistical properties (Sriperumbudur et al. (2017))

Kernel Conditional Exponential Family

Idea: Extend the KEF to conditional densities:

$$p_\theta(y|x) = q_0(y)e^{\langle \theta_x, k(y,.)\rangle_{\mathcal{G}} - A(\theta_x)} \qquad A(\theta_x) = \log \int q_0(y)e^{\langle \theta_x, k(y,.)\rangle_{\mathcal{G}}}dy$$

$x \mapsto \theta_x$ constrained to be in a vector valued RKHS $\mathcal{H}$ with vector valued kernel $\Gamma_{x,x'}$.
$\mathcal{H}$ contains functions $\theta : \mathcal{X} \mapsto \mathcal{G}$ that satisfy the vector valued reproducing property (Micchelli and Pontil (2005)):

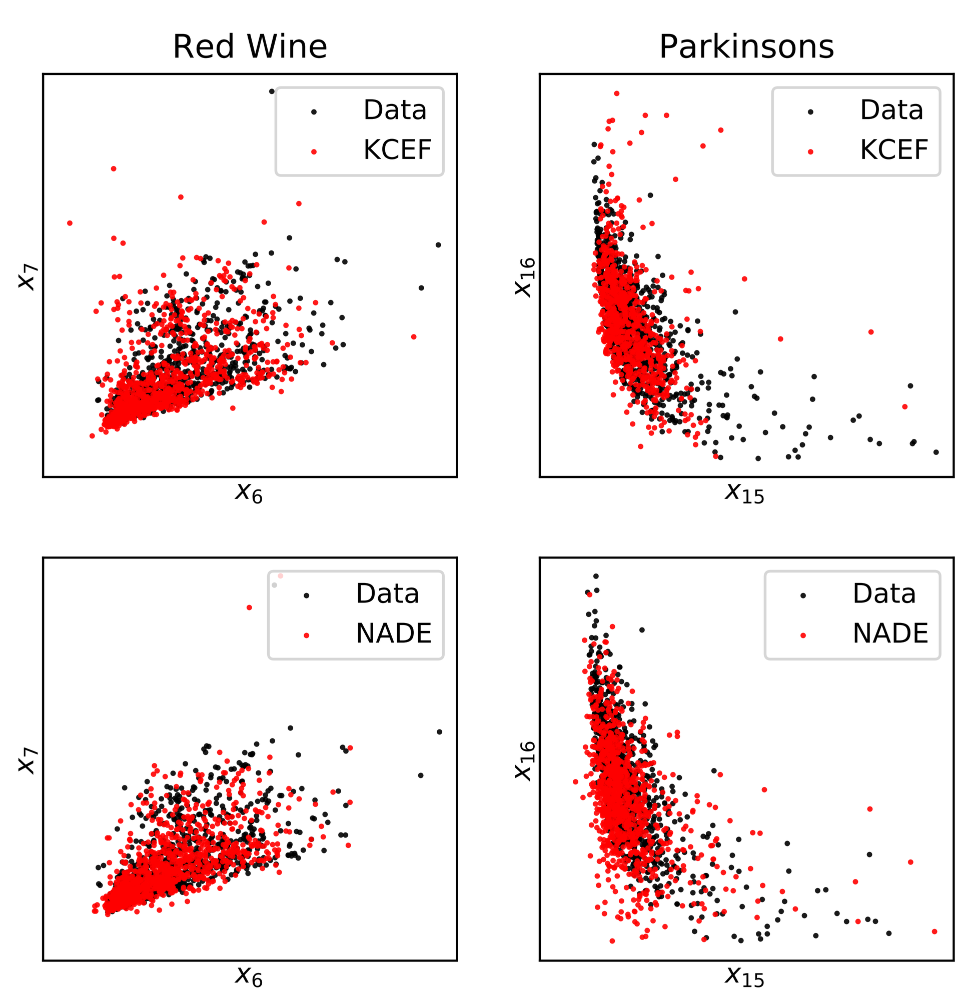
$$\langle \theta_x, f\rangle_{\mathcal{G}} = \langle \theta, \Gamma_{x,.}f\rangle_{\mathcal{H}}; \qquad \forall f \in \mathcal{G}$$

By this property, p_θ can also be written as:

$$p_\theta(y|x) = q_0(y)e^{\langle \theta, \Gamma_{x,.}k(y,.)\rangle_{\mathcal{H}} - A(\theta_x)}$$

Experiments: Sampling from KCEF

Motivation: Sampling from a high dimensional distribution $p(x_1, ..., x_d)$ can suffer from a slow mixing time.



Idea:
- ✓ Approximate p by a product of conditional densities $p \simeq \hat{p}(x_1)\hat{p}(x_2|x_1)...\hat{p}(x_d|x_{\pi(d)})$
- ✓ Use Ancestral Hamiltonian Monte Carlo to sample from $\hat{p}(X_i|X_{\pi(i)})$ given a sample $X_{\pi(i)}$.

Model comparison: Samples from different models can be compared using the test for relative similarity in (Bounliphone et al. (2015)).

Expected Conditional Score Matching

Motivation: Define a loss between two unnormalized conditional densities : $\mathcal{J}(p,q)$.
Idea: Adapt the score objective from (Hyvarinen (2005)) to conditional densities:

$$\mathcal{J}(p,q) = \frac{1}{2}\mathbb{E}_{X,Y}\left[\left\|\nabla_y \log \frac{p(Y|X)}{q(Y|X)}\right\|^2\right]$$

The expectation is under the true joint distribution. Using integration by part and some regularity conditions:

$$\mathcal{J}(p,q) = \mathbb{E}_{X,Y}\left[\Delta_y \log q(Y|X) + \frac{1}{2}\|\nabla_y \log q(Y|X)\|^2\right] + const$$

For q_θ in the KCEF the score is convex and quadratic in θ:

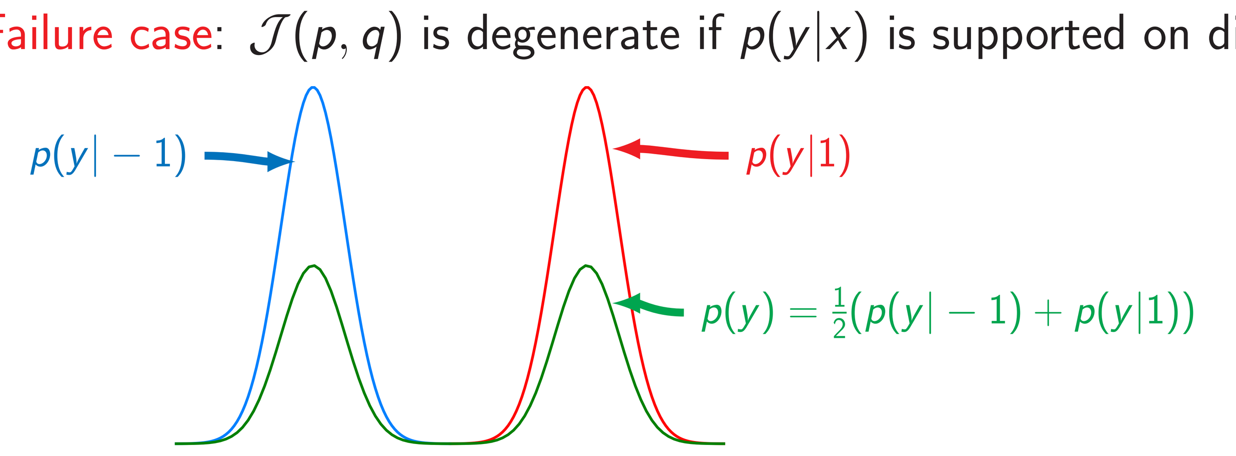
$$\mathcal{J}(p,q_\theta) = \frac{1}{2}\langle \theta, C\theta\rangle_{\mathcal{H}} + \langle \xi, \theta\rangle_{\mathcal{H}} + const$$

$$\mathbb{E}_{X,Y}\left[\sum_{i=1}^d \Gamma_{X,.}\partial_i k(Y,.) \otimes \Gamma_{X,.}\partial_i k(Y,.)\right] \qquad \mathbb{E}_{X,Y}\left[\sum_{i=1}^d \Gamma_{X,.}\partial_i^2 k(Y,.) + \partial_i \log g(Y)\Gamma_{X,.}\partial_i k(Y,.)\right]$$

C is a symmetric positive trace-class operator and ξ is a vector in $\mathcal{H}$:
- ✓ No need to compute the intractable normalizer.
- ✓ Convex quadratic loss: Guarantees existence and uniqueness of an optimal solution.
- ✓ A provably convergent algorithm can be used to estimate the optimal θ.
- × The score can become degenerate if $p(y|x)$ is not supported on the whole space.

Truth in Advertising

Failure case: $\mathcal{J}(p,q)$ is degenerate if $p(y|x)$ is supported on disjoint subsets.



$$\mathcal{J}(\underbrace{p(y|x)}_{\text{target}}, \underbrace{p(y)}_{\text{model}}) = 0$$

Easy Fix: Add a small gaussian noise to the data!

Finite Sample estimate

Given n samples $(X_i, Y_i)_{1\le i\le n}$, the regularized empirical version of the score is:

$$\hat{\mathcal{J}}(p,q) = \frac{1}{2}\langle \theta, \hat{C}\theta\rangle_{\mathcal{H}} + \langle \hat{\xi}, \theta\rangle_{\mathcal{H}} + \frac{\lambda}{2}\|\theta\|^2_{\mathcal{H}}$$

kernel trick: The generalized representer theorem ensures $\hat{\theta}$ is of the form:

$$\hat{\theta} = -\frac{1}{\lambda}\hat{\xi} + \sum_{b\in[n];i\in[d]} \beta_{(b,i)}\Gamma_{X_b}\partial_i k(Y_b,.)$$

where β is obtained by solving a linear system of size $n \times d$:

$$(G + n\lambda I)\beta = \frac{1}{\lambda}h$$

$$(\langle \partial_i k(Y_a,.), \Gamma(X_a,X_b)\partial_j k(Y_b,.)\rangle_{\mathcal{H}})_{(a,i),(b,j)\in[n]\times[d]} \qquad (\partial_i\hat{\xi}(X_a,Y_a))_{(a,i)\in[n]\times[d]}$$

Theory

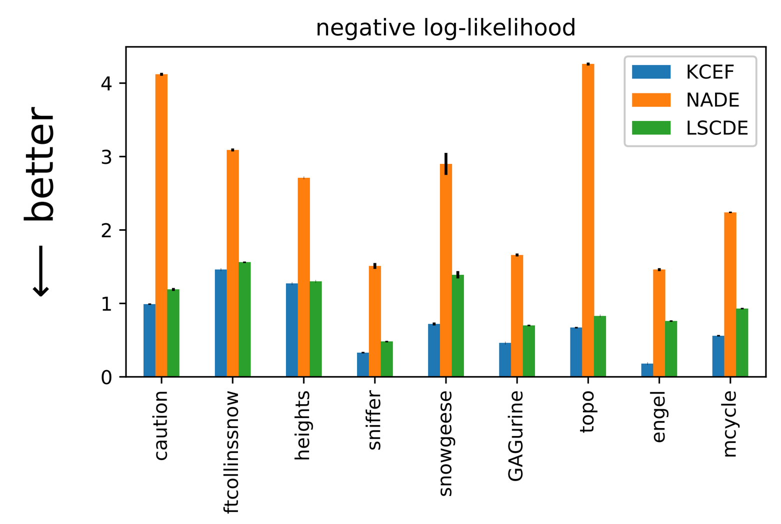
The paper provides asymptotic rates of convergence of $\hat{\theta}$ in the well-specified case. If θ_0 is the true natural parameter, then:

$$\|\hat{\theta} - \theta_0\| = \mathcal{O}_{p_0}(n^{-\frac{1}{2}+\alpha})$$

with $\lambda = n^{-\alpha}$ and $\frac{1}{4} < \alpha < \frac{1}{2}$ depends on the kernels and p_0.

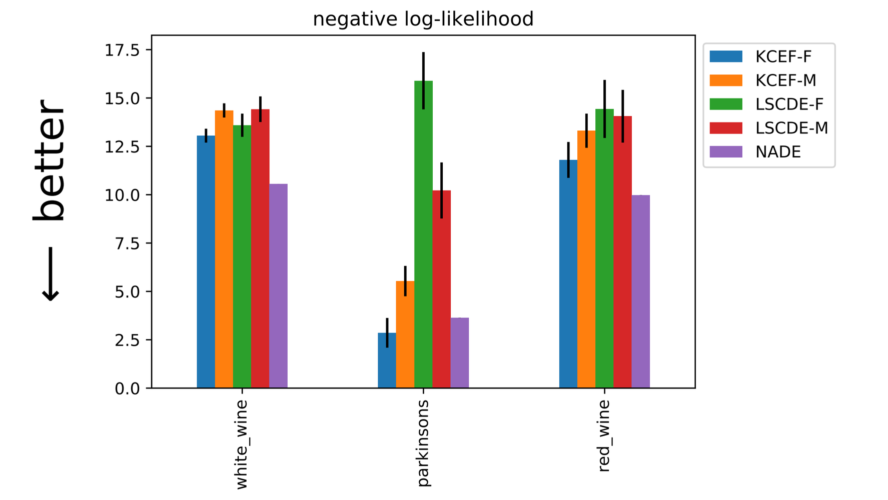
Experiments: Comparison with other methods: Real NADE and LSCDE

Estimating $p(y|x)$.



R datasets: $N_{train} = N_{test}$ from 10 to 300. $d_y = 1$ and d_x from 1 to 13.

Estimating $p(x_1, ..., x_d)$ as a product of conditional densities.



UCI datasets: $N_{train} = 10 * N_{test}$ from 1500 to 5000. d from 11 to 15.

Bibliography

Bounliphone, W., Belilovsky, E., Blaschko, M. B., Antonoglou, I., and Gretton, A. (2015). A test of relative similarity for model selection in generative models. *CoRR*.
Hyvarinen, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*
Micchelli, C. A. and Pontil, M. A. (2005). On learning vector-valued functions. *Neural Comput.*
Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., and Hyvarinen, A. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*.

This work was funded by the Gatsby Charitable Foundation.

Contact: michael.n.arbel@gmail.com