Representing and comparing probabilities with kernels: Part 3

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

MLSS Madrid, 2018

Training GANs with MMD

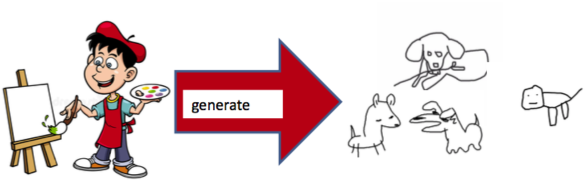
What is a Generative Adversarial Network (GAN)?

- Generator (student)

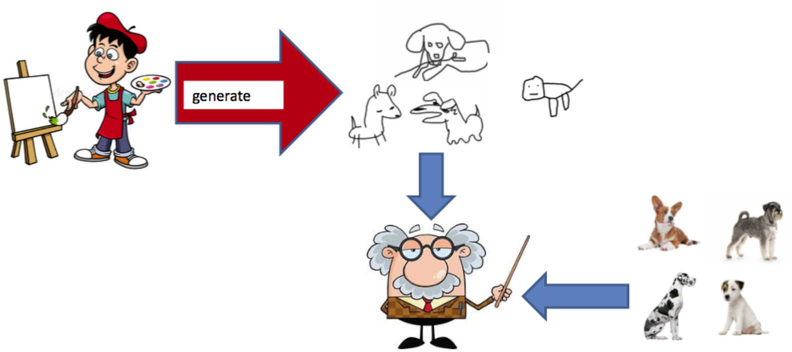
- Task: critic must teach generator to draw images (here dogs)

- Critic (teacher)

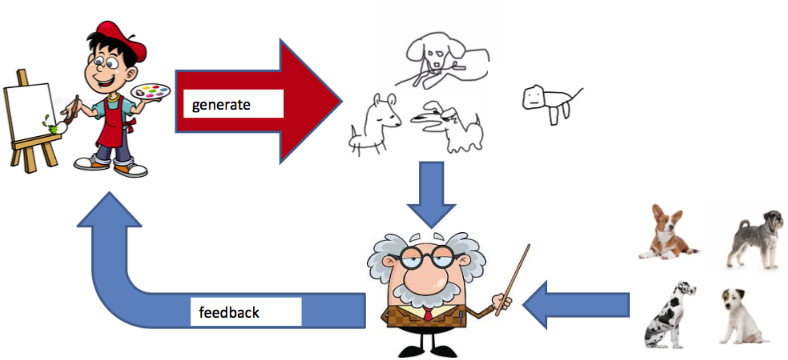
What is a Generative Adversarial Network (GAN)?



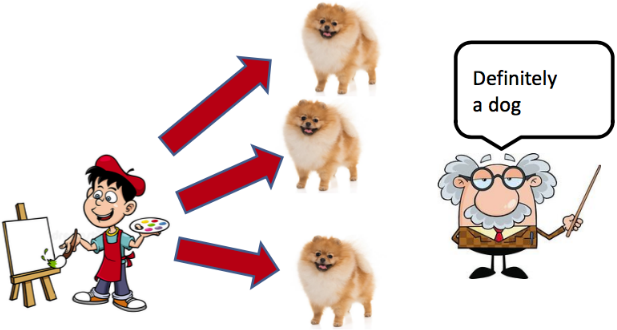
What is a Generative Adversarial Network (GAN)?



What is a Generative Adversarial Network (GAN)?



Why is classification not enough?



Classification **not** enough!

Need to compare **sets**

(otherwise student can just produce the same dog over and over)

MMD for GAN critic

Can you use MMD as a critic to train GANs?

From ICML 2015:

Generative Moment Matching Networks

Yujia Li[1] YUJIALI@CS.TORONTO.EDU
Kevin Swersky[1] KSWERSKY@CS.TORONTO.EDU
Richard Zemel[1,2] ZEMEL@CS.TORONTO.EDU
[1]Department of Computer Science, University of Toronto, Toronto, ON, CANADA
[2]Canadian Institute for Advanced Research, Toronto, ON, CANADA

From UAI 2015:

Training generative neural networks via Maximum Mean Discrepancy optimization

Gintare Karolina Dziugaite
University of Cambridge

Daniel M. Roy
University of Toronto

Zoubin Ghahramani
University of Cambridge

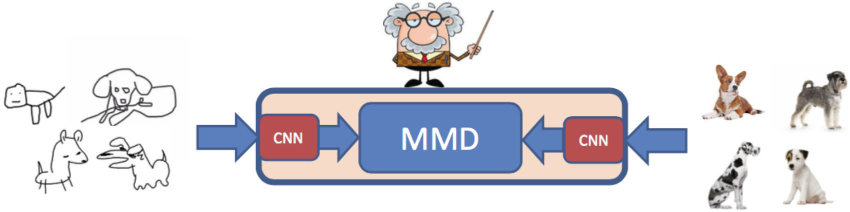
MMD for GAN critic

Can you use MMD as a critic to train GANs?

Need better image features.

How to improve the critic witness

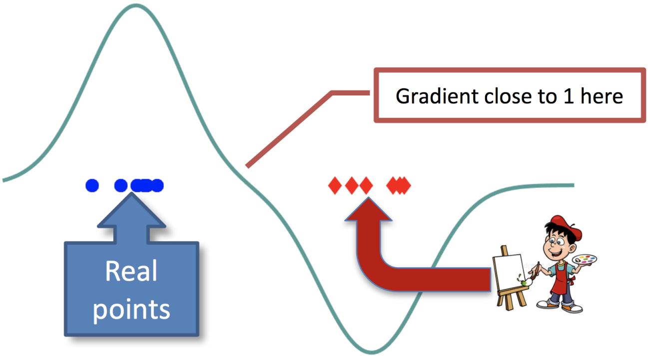
- Add convolutional features!
- The critic (teacher) also needs to be trained.
- How to regularise?



MMD GAN Li et al., [NIPS 2017]
Coulomb GAN Unterthiner et al., [ICLR 2018]

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]



WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]

-  Given a generator G_θ with parameters θ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$
-  Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ.

WGAN-GP

Wasserstein GAN Arjovsky et al. [ICML 2017]
WGAN-GP Gukrajani et al. [NIPS 2017]

- Given a generator G_θ with parameters θ to be trained. Samples $Y \sim G_\theta(Z)$ where $Z \sim R$

- Given critic features h_ψ with parameters ψ to be trained. f_ψ a linear function of h_ψ.

WGAN-GP gradient penalty:

$$\max_{\psi} \mathbf{E}_{X\sim P} f_\psi(X) - \mathbf{E}_{Z\sim R} f_\psi(G_\theta(Z)) + \lambda \mathbf{E}_{\widetilde{X}} \left(\left\| \nabla_{\widetilde{X}} f_\theta(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$\widetilde{X} = \gamma x_i + (1-\gamma) G_\psi(z_j)$$

$$\gamma \sim \mathcal{U}([0,1]) \quad x_i \in \{x_\ell\}_{\ell=1}^m \quad z_j \in \{z_\ell\}_{\ell=1}^n$$

The (W)MMD

Train MMD critic features with the witness function gradient penalty

Binkowski, Sutherland, Arbel, G. [ICLR 2018], Bellemare et al. [2017] for energy distance:

$$\max_{\psi} MMD^2(h_\psi(X), h_\psi(G_\theta(Z))) + \lambda \mathbf{E}_{\widetilde{X}} \left(\left\| \nabla_{\widetilde{X}} f_\psi(\widetilde{X}) \right\| - 1 \right)^2$$

where

$$f_\psi(\cdot) = \frac{1}{m} \sum_{i=1}^{m} k(h_\psi(x_i), \cdot) - \frac{1}{n} \sum_{j=1}^{n} k(h_\psi(G_\theta(z_j)), \cdot)$$

New

$$\widetilde{X} = \gamma x_i + (1-\gamma) G_\psi(z_j)$$

$$\gamma \sim \mathcal{U}([0,1]) \quad x_i \in \{x_\ell\}_{\ell=1}^{m} \quad z_j \in \{z_\ell\}_{\ell=1}^{n}$$

Remark by Bottou et al. (2017): gradient penalty modifies the function class. So critic is not an MMD in RKHS $\mathcal{F}$.

MMD for GAN critic: revisited

From ICLR 2018:

DEMYSTIFYING MMD GANS

Mikołaj Bińkowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Dougal J. Sutherland,* Michael Arbel & Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
{dougal,michael.n.arbel,arthur.gretton}@gmail.com

MMD for GAN critic: revisited

Samples are better!

MMD for GAN critic: revisited

Samples are better!

Can we do better still?

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$

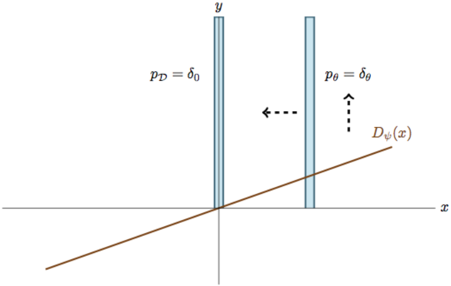


Figure from Mescheder et al. [ICML 2018]

Convergence issues for WGAN-GP penalty

WGAN-GP style gradient penalty may not converge near solution

Nagarajan and Kolter [NIPS 2017], Mescheder et al. [ICML 2018], Balduzzi et al. [ICML 2018]

The Dirac-GAN

$$P = \delta_0 \qquad Q = \delta_\theta \qquad f_\psi(x) = \psi \cdot x$$

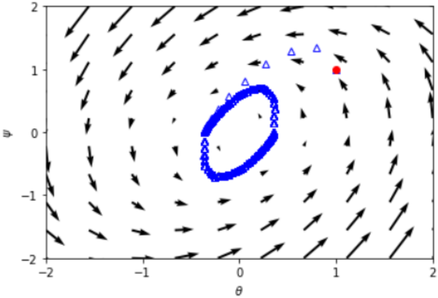


Figure from Mescheder et al. [ICML 2018]

A better gradient penalty

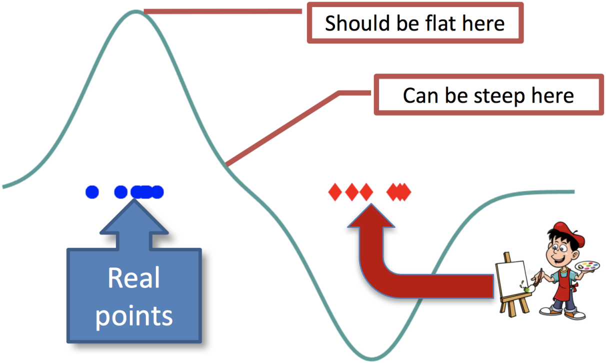
- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]



A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\|f\|_S \leq 1} \left[\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y) \right]$$

where

$$\|f\|_S^2 = \|f\|_{L_2(P)}^2 + \|\nabla f\|_{L_2(P)}^2 + \lambda \|f\|_k^2$$

L_2 norm control

Gradient control

RKHS smoothness

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

Modified witness function:

$$\widetilde{MMD} := \sup_{\|f\|_S \leq 1} \left[\mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)\right]$$

where

$$\|f\|_S^2 = \|f\|_{L_2(P)}^2 + \|\nabla f\|_{L_2(P)}^2 + \lambda \|f\|_k^2$$

L_2 norm control

Gradient control

RKHS smoothness

Problem: not computationally feasible: $O(n^3)$ per iteration.

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda}\ MMD$$

where

$$\sigma_{k,P,\lambda} = \left(\lambda + \int k(x,x)\,dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,x)\ dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

$$\|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1}\ \|f\|_k^2$$

A better gradient penalty

- New MMD GAN witness regulariser (just accepted, NIPS 2018)

  Arbel, Sutherland, Binkowski, G. [NIPS 2018]
- Based on semi-supervised learning regulariser Bousquet et al. [NIPS 2004]
- Related to Sobolev GAN Mroueh et al. [ICLR 2018]

The scaled MMD:

$$SMMD = \sigma_{k,P,\lambda} \; MMD$$

where

$$\sigma_{k,P,\lambda} = \left(\lambda + \int k(x,x) dP(x) + \sum_{i=1}^{d} \int \partial_i \partial_{i+d} k(x,x) \; dP(x) \right)^{-1/2}$$

Replace expensive constraint with cheap upper bound:

$$\|f\|_S^2 \leq \sigma_{k,P,\lambda}^{-1} \; \|f\|_k^2$$

Idea: rather than regularise the critic or witness function, regularise features directly

Evaluation and experiments

Evaluation of GANs

The inception score? Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Evaluation of GANs

The inception score? Salimans et al. [NIPS 2016]

Based on the classification output $p(y|x)$ of the inception model Szegedy et al. [ICLR 2014],

$$E_X \exp KL(P(y|X) \| P(y)).$$

High when:

- predictive label distribution $P(y|x)$ has low entropy (good quality images)
- label entropy $P(y)$ is high (good variety).

Problem: relies on a trained classifier! Can't be used on new categories (celeb, bedroom...)

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NIPS 2017]

Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \mathrm{tr}(\Sigma_P) + \mathrm{tr}(\Sigma_Q) - 2\mathrm{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where μ_P and Σ_P are the feature mean and covariance of P

Evaluation of GANs

The Frechet inception distance? Heusel et al. [NIPS 2017]

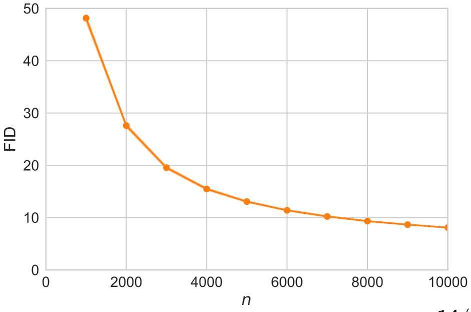
Fits Gaussians to features in the inception architecture (pool3 layer):

$$FID(P, Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}(\Sigma_P) + \text{tr}(\Sigma_Q) - 2\text{tr}\left((\Sigma_P \Sigma_Q)^{\frac{1}{2}}\right)$$

where μ_P and Σ_P are the feature mean and covariance of P

Problem: bias. For finite samples can consistently give incorrect answer.

- Bias demo, CIFAR-10 train vs test



Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \to \infty$ samples from Q.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \to \infty$ samples from Q.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \to \infty$ samples from Q.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

Evaluation of GANs

The FID can give the wrong answer in theory.

Assume m samples from P and $n \to \infty$ samples from Q.

Given two alternatives:

$$P_1 \sim \mathcal{N}(0, (1 - m^{-1})^2) \qquad P_2 \sim \mathcal{N}(0, 1) \qquad Q \sim \mathcal{N}(0, 1).$$

Clearly,

$$FID(P_1, Q) = \frac{1}{m^2} > FID(P_2, Q) = 0$$

Given m samples from P_1 and P_2,

$$FID(\widehat{P_1}, Q) < FID(\widehat{P_2}, Q).$$

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \qquad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C.

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \qquad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C.

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \text{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \text{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2I_d)) \qquad Q = \text{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d}CC^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C.

Evaluation of GANs

The FID can give the wrong answer in practice.

Let $d = 2048$, and define

$$P_1 = \mathrm{relu}(\mathcal{N}(\mathbf{0}, I_d)) \quad P_2 = \mathrm{relu}(\mathcal{N}(\mathbf{1}, .8\Sigma + .2 I_d)) \qquad Q = \mathrm{relu}(\mathcal{N}(\mathbf{1}, I_d))$$

where $\Sigma = \frac{4}{d} C C^T$, with C a $d \times d$ matrix with iid standard normal entries.

For a random draw of C:

$$FID(P_1, Q) \approx 1123.0 > 1114.8 \approx FID(P_2, Q)$$

With $m = 50\,000$ samples,

$$FID(\widehat{P_1}, Q) \approx 1133.7 < 1136.2 \approx FID(\widehat{P_2}, Q)$$

At $m = 100\,000$ samples, the ordering of the estimates is correct.

This behavior is similar for other random draws of C.

The kernel inception distance (KID)

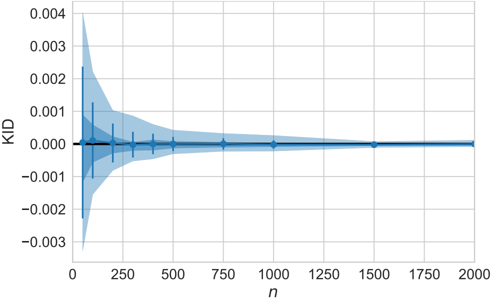
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3.$$

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test



The kernel inception distance (KID)

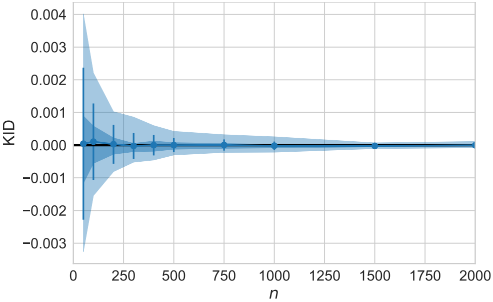
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3.$$

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test



..."but isn't KID is computationally costly?"

The kernel inception distance (KID)

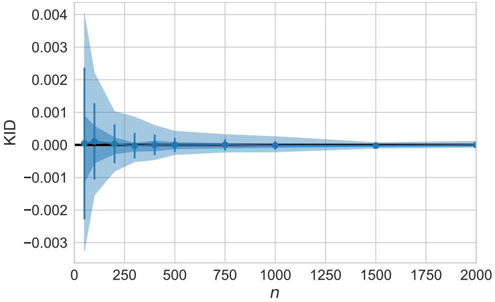
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x,y) = \left(\frac{1}{d}x^\top y + 1\right)^3.$$

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test



..."but isn't KID is computationally costly?"

"Block" KID implementation is cheaper than FID: see paper (or use Tensorflow implementation)!

The kernel inception distance (KID)

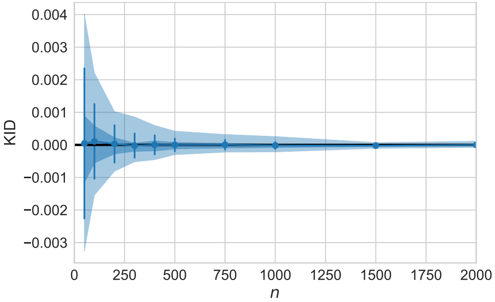
The Kernel inception distance Binkowski, Sutherland, Arbel, G. [ICLR 2018]

Measures similarity of the samples' representations in the inception architecture (pool3 layer)

MMD with kernel

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3.$$

- Checks match for feature means, variances, skewness
- Unbiased : eg CIFAR-10 train/test



Also used for automatic learning rate adjustment: if $KID(\widehat{P}_{t+1}, Q)$ not significantly better than $KID(\widehat{P}_t, Q)$ then reduce learning rate.

[Bounliphone et al. ICLR 2016]

Related: "An empirical study on evaluation metrics of generative adversarial networks", Xu et al. [arxiv, June 2018]

Benchmarks for comparison (all from ICLR 2018)

SPECTRAL NORMALIZATION FOR GENERATIVE ADVERSARIAL NETWORKS

Takeru Miyato[1], Toshiki Kataoka[1], Masanori Koyama[2], Yuichi Yoshida[3]
{miyato, kataoka}@preferred.jp
oyama asanori@gmail.com
yos i.ac.jp
works, Inc. [2]Ritsumeikan University [3]National Institute of Informatics

We combine with scaled MMD

SOBOLEV GAN

Youssef Mroueh[†], Chun-Liang Li[◦,⋆], Tom Sercu[†,⋆], Anant Raj[◊,⋆] & Yu Cheng[†]
† IBM Research AI
◦ Carnegie Mellon University
◊ Max Planck Institute for Intelligent Systems
⋆ denotes Equal Contribution
{mroueh,chengyu}@us.ibm.com, chunlial@cs.cmu.edu,
tom.sercu1@ibm.com,anant.raj@tuebingen.mpg.de

DEMYSTIFYING MMD GANS

Mikołaj Bińkowski*
Department of Mathematics
Imperial College London
mikbinkowski@gmail.com

Dougal J. Sutherland,* Michael Arbel & Arthur Gretton
Gatsby Computational Neuroscience Unit
Uni College London
,michael.n.arbel,arthur.gretton}@gmail.com

Our ICLR 2018 paper

BOUNDARY-SEEKING GENERATIVE ADVERSARIAL NETWORKS

R Devon Hjelm*
MILA, University of Montréal, IVADO
erroneus@gmail.com

Athul Paul Jacob*
MILA, MSR, University of Waterloo
apjacob@edu.uwaterloo.ca

Tong Che
MILA, University of Montréal
tong.che@umontreal.ca

Adam Trischler
MSR
adam.trischler@microsoft.com

Kyunghyun Cho
New York University,
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Yoshua Bengio
MILA, University of Montréal, CIFAR, IVADO
yoshua.bengio@umontreal.ca

Results: what does MMD buy you?

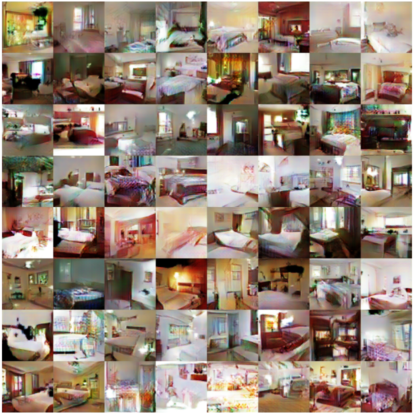
- Critic features from DCGAN: an f-filter critic has f, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64.

MMD GAN samples, $f = 64$,
FID=32, KID=3

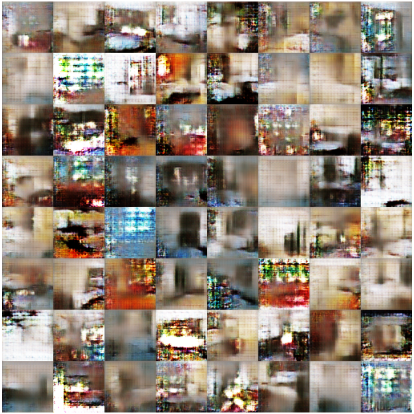
WGAN samples, $f = 64$,
FID=41, KID=4

Results: what does MMD buy you?

- Critic features from DCGAN: an f-filter critic has f, $2f$, $4f$ and $8f$ convolutional filters in layers 1-4. LSUN 64×64.



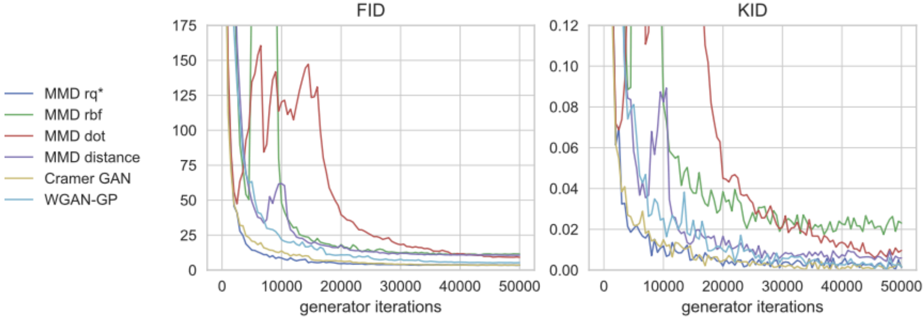
MMD GAN samples, $f = 16$, FID=86, KID=9



WGAN samples, $f = 16$, $f = 64$, FID=293, KID=37

The kernel inception distance (KID)

Faster training: performance scores vs generator iterations on MNIST



Results: celebrity faces 160×160

KID (FID) scores:

- Sobolev GAN: 14 (20)
- SN-GAN: 18 (28)
- Old MMD GAN: 13 (21)
- SMMD GAN: 6 (12)

202 599 face images, resized and cropped to 160 × 160

Results: imagenet 64×64

KID (FID) scores:

- BGAN: 47 (44)
- SN-GAN: 44 (48)
- SMMD GAN: 35 (37)

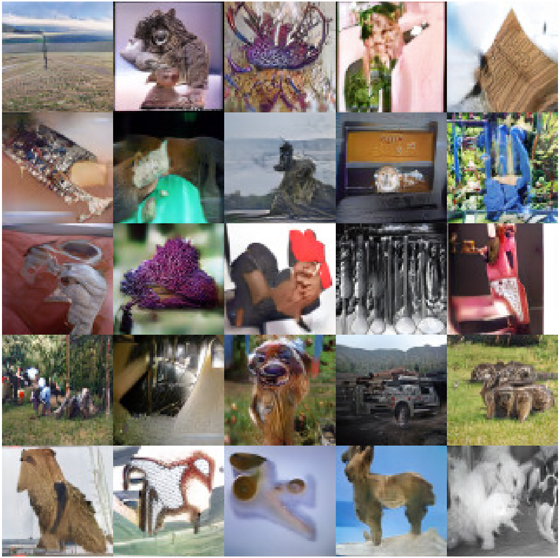
ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.

Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
  47 (44)
- SN-GAN:
  44 (48)
- SMMD GAN:
  35 (37)

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.



Results: imagenet 64×64

KID (FID)
scores:

- BGAN:
  47 (44)
- SN-GAN:
  44 (48)
- SMMD GAN:
  35 (37)

ILSVRC2012 (ImageNet) dataset, 1 281 167 images, resized to 64 × 64. Around 20 000 classes.

Summary

- MMD critic gives state-of-the-art performance for GAN training (FID and KID)
  - use convolutional input features
  - train with new gradient regulariser
- Faster training, simpler critic network
- Reasons for good performance:
  - Unlike WGAN-GP, MMD loss still a valid critic when features not optimal
  - Kernel features do some of the "work", so simpler h_ψ features possible.
  - Better gradient/feature regulariser gives better critic

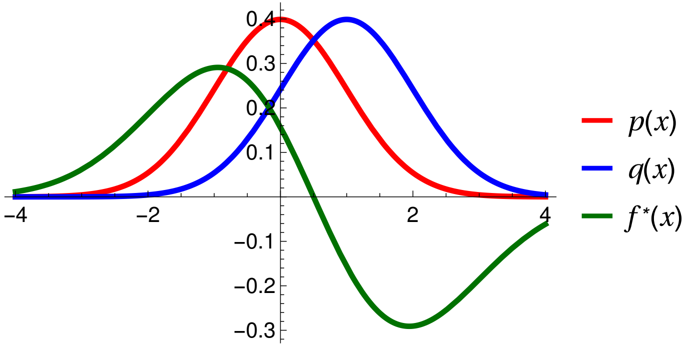
Code for "Demystifying MMD GANs," ICLR 2018, including KID score: https://github.com/mbinkowski/MMD-GAN

Code for new SMMD: https://github.com/MichaelArbel/Scaled-MMD-GAN

Testing against a probabilistic model

Statistical model criticism

$$MMD(P, Q) = \|f^*\|^2 = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_Q f - E_p f]$$



$f^*(x)$ is the witness function

Can we compute MMD with samples from Q and a **model** P?

Problem: usualy can't compute $E_p f$ in closed form.

Stein idea

To get rid of $E_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} [E_q f - E_p f]$$

we define the **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Then

$$E_P T_P f = 0$$

subject to appropriate boundary conditions. (Oates, Girolami, Chopin, 2016)

Stein idea: proof

$$E_p[T_p f] = \int \left[\frac{1}{p(x)} \frac{d}{dx}(f(x)p(x))\right] p(x) dx$$

$$\int \left[\frac{d}{dx}(f(x)p(x))\right] dx$$

$$= [f(x)p(x)]_{-\infty}^{\infty}$$

$$= 0$$

Stein idea: proof

$$E_{p}\left[T_{p}f\right] = \int \left[\frac{1}{\cancel{p(x)}}\frac{d}{dx}\left(f(x)p(x)\right)\right]\cancel{p(x)}dx$$

$$\int \left[\frac{d}{dx}\left(f(x)p(x)\right)\right]dx$$

$$= \left[f(x)p(x)\right]_{-\infty}^{\infty}$$

$$= 0$$

Stein idea: proof

$$E_p\left[T_p f\right] = \int \left[\frac{1}{\cancel{p(x)}}\frac{d}{dx}\left(f(x)p(x)\right)\right]\cancel{p(x)}dx$$

$$\int \left[\frac{d}{dx}\left(f(x)p(x)\right)\right] dx$$

$$= \left[f(x)p(x)\right]_{-\infty}^{\infty}$$

$$= 0$$

Stein idea: proof

$$E_p [T_p f] = \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx$$

$$\int \left[\frac{d}{dx} (f(x)p(x)) \right] dx$$

$$= [f(x)p(x)]_{-\infty}^{\infty}$$

$$= 0$$

Stein idea: proof

$$E_{p}\left[T_{p}f\right] = \int \left[\frac{1}{\cancel{p(x)}}\frac{d}{dx}\left(f(x)p(x)\right)\right]\cancel{p(x)}dx$$

$$\int \left[\frac{d}{dx}\left(f(x)p(x)\right)\right]dx$$

$$= \left[f(x)p(x)\right]_{-\infty}^{\infty}$$

$$= 0$$

Kernel Stein Discrepancy

Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right)$$

Kernel Stein Discrepancy (KSD)

$$KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g$$

Kernel Stein Discrepancy

Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right)$$

Kernel Stein Discrepancy (KSD)

$$KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$

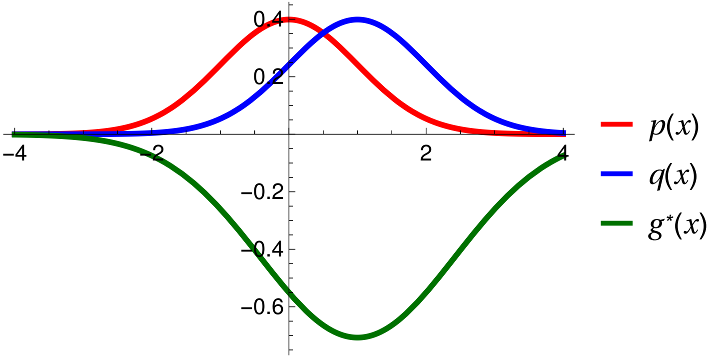
Kernel Stein Discrepancy

Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Kernel Stein Discrepancy (KSD)

$$KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



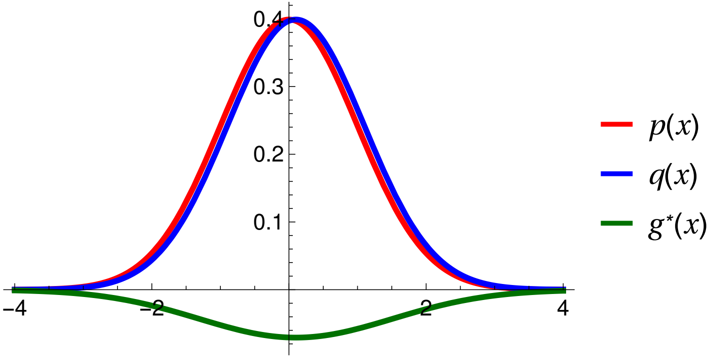
Kernel Stein Discrepancy

Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx}\left(f(x)p(x)\right)$$

Kernel Stein Discrepancy (KSD)

$$KSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



Kernel stein discrepancy

Closed-form expression for KSD: given $Z, Z' \sim q$, then
(Chwialkowski, Strathmann, G., ICML 2016) (Liu, Lee, Jordan ICML 2016)

$$\mathrm{KSD}(p, q, \mathcal{F}) = E_q h_p(Z, Z')$$

where

$$\begin{aligned} h_p(x, y) := {} & \partial_x \log p(x) \partial_x \log p(y) k(x, y) \\ & + \partial_y \log p(y) \partial_x k(x, y) \\ & + \partial_x \log p(x) \partial_y k(x, y) \\ & + \partial_x \partial_y k(x, y) \end{aligned}$$

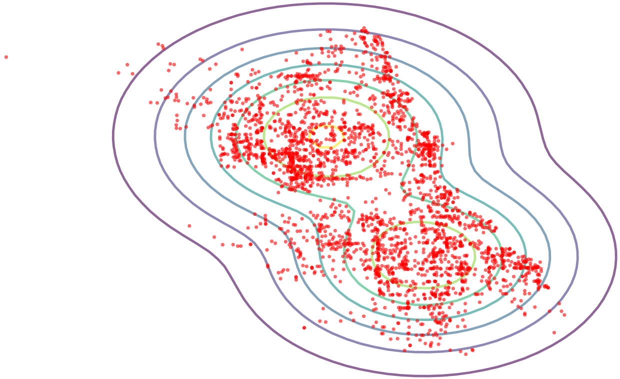
and k is RKHS kernel for $\mathcal{F}$

Only depends on kernel and $\partial_x \log p(x)$. Do not need to normalize p, or sample from it.

Statistical model criticism

Chicago crime data

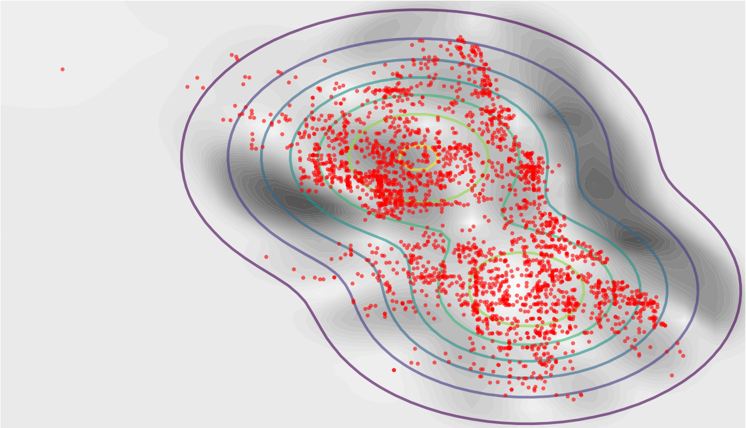
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with two components.

Statistical model criticism

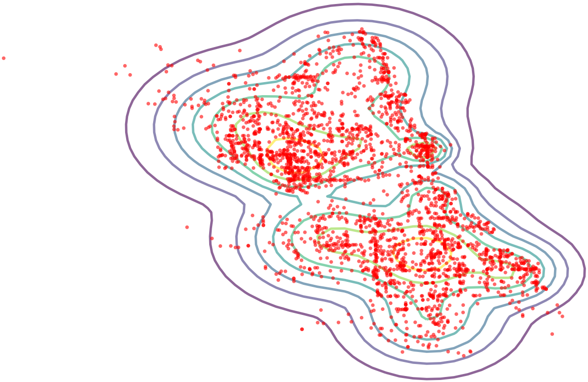


Chicago crime data

Model is Gaussian mixture with two components

Stein **witness function**

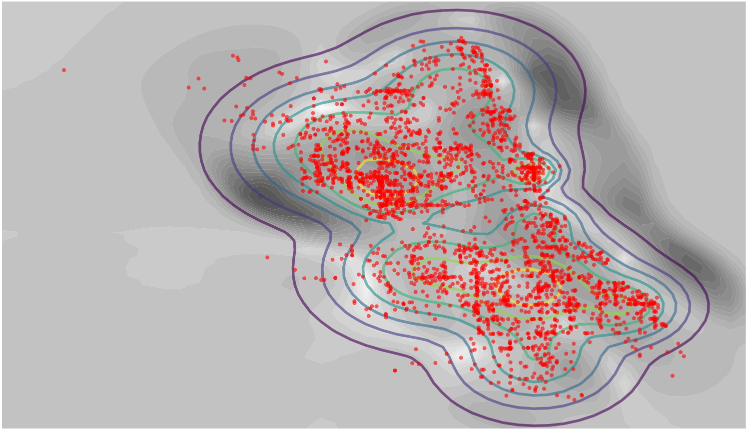
Statistical model criticism



Chicago crime data

Model is Gaussian mixture with ten components.

Statistical model criticism



Chicago crime data

Model is Gaussian mixture with ten components

Stein **witness function**

Code: https://github.com/karlnapf/kernel_goodness_of_fit

Kernel stein discrepancy

Further applications:

- Evaluation of approximate MCMC methods.
(Chwialkowski, Strathmann, G., ICML 2016; Gorham, Mackey, ICML 2017)

What kernel to use?

- The inverse multiquadric kernel,

$$k(x, y) = \left(c + \|x - y\|_2^2\right)^{\beta}$$

for $\beta \in (-1, 0)$.


arXiv.org > stat > arXiv:1703.01717

Statistics > Machine Learning

Measuring Sample Quality with Kernels

Jackson Gorham, Lester Mackey ICML 2017

(Submitted on 6 Mar 2017 (v1), last revised 3 Aug 2017 (this version, v6))

Testing statistical dependence

Dependence testing

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y independent?

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Text from dogtime.com and petfinder.com

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_{P}, \underbrace{P_X P_Y}_{Q}, \mathcal{H}_\kappa)$$

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_{P}, \underbrace{P_X P_Y}_{Q}, \mathcal{H}_\kappa)$$

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

MMD as a dependence measure?

Could we use MMD?

$$MMD(\underbrace{P_{XY}}_{P}, \underbrace{P_X P_Y}_{Q}, \mathcal{H}_\kappa)$$

- We don't have samples from $Q := P_X P_Y$, only pairs $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$
  - Solution: simulate Q with pairs (x_i, y_j) for $j \neq i$.
- What kernel κ to use for the RKHS $\mathcal{H}_\kappa$?

MMD as a dependence measure

Kernel k on images with feature space $\mathcal{F}$,

$k\big($  ,  $\big)$

Kernel l on captions with feature space $\mathcal{G}$,

$l\big($ A large animal who slings slobber, … , A responsive, interactive pet … $\big)$

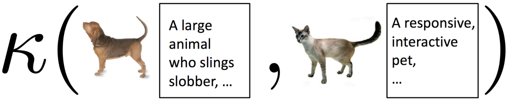
MMD as a dependence measure

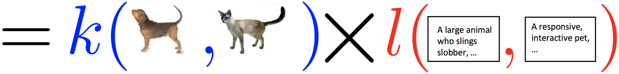
Kernel k on images with feature space $\mathcal{F}$,

Kernel l on captions with feature space $\mathcal{G}$,

$l($ A large animal who slings slobber, … $,$ A responsive, interactive pet … $)$

Kernel κ on image-text pairs: are images and captions similar?





MMD as a dependence measure

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y independent?

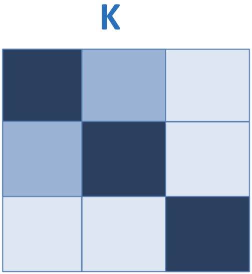
$$MMD^2(\widehat{P}_{XY}, \widehat{P}_X \widehat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2}\text{trace}(K L)$$

(K, L column centered)

MMD as a dependence measure

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y independent?

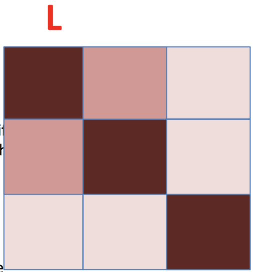
$$MMD^2(\widehat{P}_{XY}, \widehat{P}_X \widehat{P}_Y, \mathcal{H}_\kappa) := \frac{1}{n^2}\text{trace}(K\,L)$$



A large animal who slings slobber, exudes a distinctive houndy odor, ...

Their noses guide them through li
and they're never happier than wh
following an interesting scent.

A responsive, interactive pet, one that will blow in your ear and follow you everywhere.



Text from dogtime.com and petfinder.com

MMD as a dependence measure

Two questions:

- Why the product kernel? Many ways to combine kernels - why not eg a sum?
- Is there a more interpretable way of defining this dependence measure?

Illustration: dependence $\neq$correlation

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y dependent?

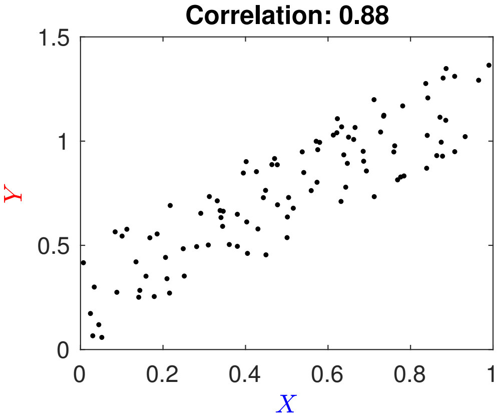


Illustration: dependence $\neq$correlation

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y dependent?

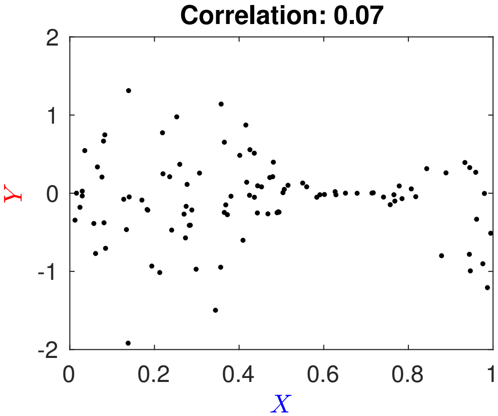
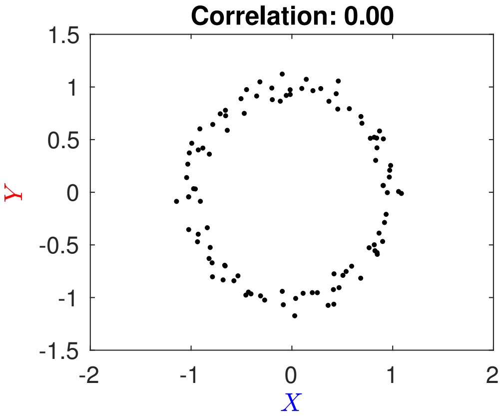


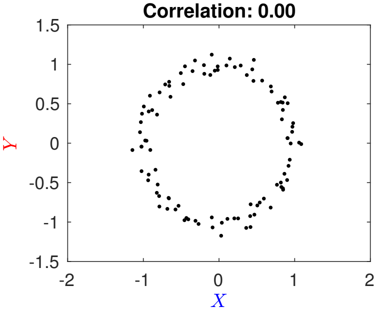
Illustration: dependence $\neq$correlation

- **Given:** Samples from a distribution P_{XY}
- **Goal:** Are X and Y dependent?



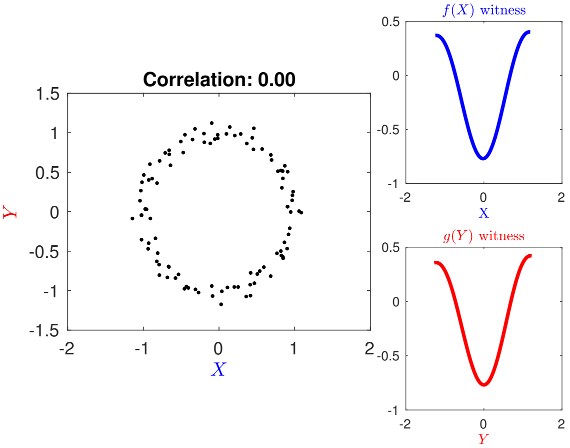
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



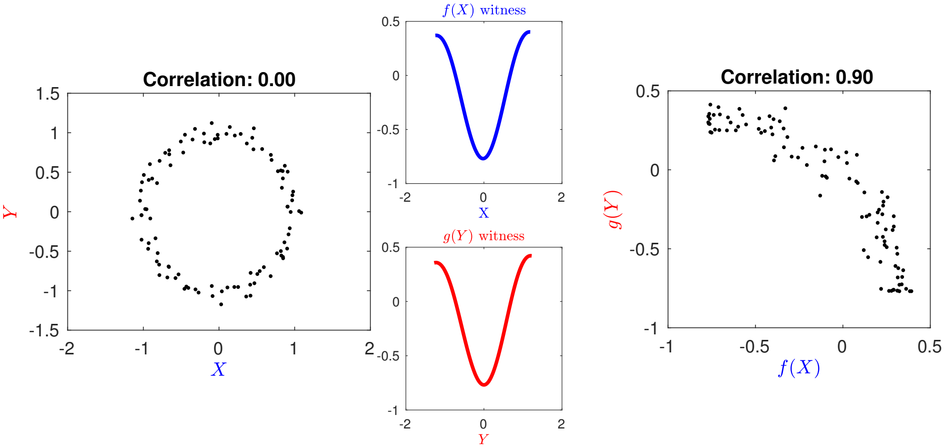
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.



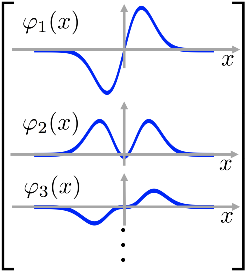
Define two spaces, one for each witness

Function in $\mathcal{F}$

$$f(x) = \sum_{j=1}^{\infty} f_j \varphi_j(x)$$

Feature map

$$\varphi(x) =$$



Kernel for RKHS $\mathcal{F}$ on $\mathcal{X}$:

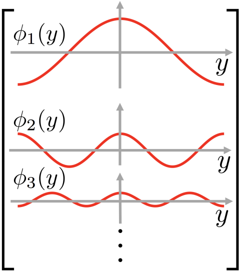
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in $\mathcal{G}$

$$g(y) = \sum_{j=1}^{\infty} g_j \phi_j(y)$$

Feature map

$$\phi(y) =$$



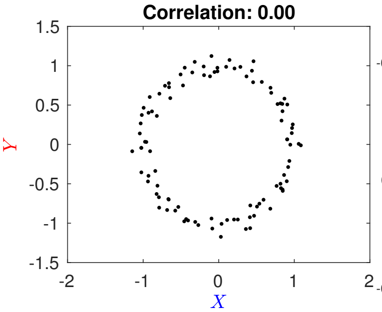
Kernel for RKHS $\mathcal{G}$ on $\mathcal{Y}$:

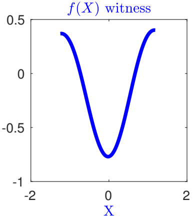
$$l(x, x') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$

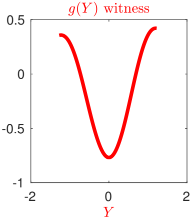
The constrained covariance

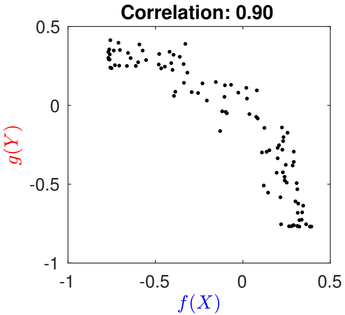
The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}[f(x)g(y)]$$









The constrained covariance

The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(x)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(y)\right)\right]$$

The constrained covariance

The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} E_{xy}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(x)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(y)\right)\right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} E_{xy}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(x)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(y)\right)\right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

$$E_{xy}[f(x)g(y)]$$

$$= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathbf{E}_{xy}\left(\begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix} \begin{bmatrix} \phi_1(y) & \phi_2(y) & \dots \end{bmatrix}\right)}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}$$

The constrained covariance

The constrained covariance is

$$\text{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} E_{xy}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(x)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(y)\right)\right]$$

Fine print: feature mappings $\varphi(x)$ and $\phi(y)$ assumed to have zero mean.

Rewriting:

$$E_{xy}[f(x)g(y)]$$

$$= \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}^\top \underbrace{\mathbf{E}_{xy}\left(\begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \vdots \end{bmatrix}\begin{bmatrix} \phi_1(y) & \phi_2(y) & \dots \end{bmatrix}\right)}_{C_{\varphi(x)\phi(y)}} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \end{bmatrix}$$

COCO: max singular value of feature covariance $C_{\varphi(x)\phi(y)}$

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{COCO}$?

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{COCO}$?

$\widehat{COCO}$ is largest eigenvalue $\gamma_{\max}$ of

$$\begin{bmatrix} 0 & \frac{1}{n}KL \\ \frac{1}{n}LK & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} K & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$.

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n}\sum_{i=1}^n \phi(y_i)$.

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS'05

Computing COCO in practice

Given sample $\{(x_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{COCO}$?

$\widehat{COCO}$ is largest eigenvalue $\gamma_{\max}$ of

$$\begin{bmatrix} 0 & \frac{1}{n}KL \\ \frac{1}{n}LK & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} K & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$.

Witness functions (singular vectors):

$$f(x) \propto \sum_{i=1}^n \alpha_i k(x_i, x) \qquad g(y) \propto \sum_{i=1}^n \beta_i l(y_i, y)$$

Fine print: kernels are computed with empirically centered features $\varphi(x) - \frac{1}{n}\sum_{i=1}^n \varphi(x_i)$ and $\phi(y) - \frac{1}{n}\sum_{i=1}^n \phi(y_i)$.

G., Smola., Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Schoelkopf, and Logothetis, AISTATS'05

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2}\left(\|f\|^2_{\mathcal{F}} - 1\right) - \frac{\gamma}{2}\left(\|g\|^2_{\mathcal{G}} - 1\right)}_{\text{smoothness constraints}}.$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}.$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \qquad g = \sum_{i=1}^{n} \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\phi(y_i)$.

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}.$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \qquad g = \sum_{i=1}^{n} \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\phi(y_i)$.

First step is smoothness constraint:

$$\|f\|_{\mathcal{F}}^2 - 1 = \langle f, f \rangle_{\mathcal{F}} - 1$$

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f,g,\lambda,\gamma) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}.$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \qquad g = \sum_{i=1}^{n} \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\phi(y_i)$.

First step is smoothness constraint:

$$\begin{aligned}\|f\|_{\mathcal{F}}^2 - 1 &= \langle f, f\rangle_{\mathcal{F}} - 1 \\ &= \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{i=1}^{n} \alpha_i \varphi(x_i)\right\rangle_{\mathcal{F}} - 1\end{aligned}$$

Empirical COCO: proof (1)

The Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)]}_{\text{covariance}} - \underbrace{\frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) - \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right)}_{\text{smoothness constraints}}.$$

Fine print: $f(x_i)g(y_i)$ centered to have zero empirical mean.

Assume (cf representer theorem):

$$f = \sum_{i=1}^{n} \alpha_i \varphi(x_i) \qquad g = \sum_{i=1}^{n} \beta_i \psi(y_i)$$

for centered $\varphi(x_i)$, $\phi(y_i)$.

First step is smoothness constraint:

$$\begin{aligned}\|f\|_{\mathcal{F}}^2 - 1 &= \langle f, f\rangle_{\mathcal{F}} - 1 \\ &= \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\rangle_{\mathcal{F}} - 1 \\ &= \alpha^\top K \alpha - 1\end{aligned}$$

Proof sketch (2)

Second step is covariance:

$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)] = \frac{1}{n}\sum_{i=1}^{n}\langle f,\varphi(x_i)\rangle_{\mathcal{F}}\langle g,\varphi(y_i)\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\langle \sum_{\ell=1}^{n}\alpha_\ell\varphi(x_\ell),\varphi(x_i)\right\rangle_{\mathcal{F}}\langle g,\varphi(y_i)\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\alpha^\top KL\beta$$

Proof sketch (2)

Second step is covariance:

$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)] = \frac{1}{n}\sum_{i=1}^{n}\langle f, \varphi(x_i)\rangle_{\mathcal{F}}\langle g, \varphi(y_i)\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\langle \sum_{\ell=1}^{n}\alpha_\ell\varphi(x_\ell), \varphi(x_i)\right\rangle_{\mathcal{F}}\langle g, \varphi(y_i)\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\alpha^\top KL\beta$$

Proof sketch (2)

Second step is covariance:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)] &= \frac{1}{n}\sum_{i=1}^{n}\langle f,\varphi(x_i)\rangle_{\mathcal{F}}\,\langle g,\varphi(y_i)\rangle_{\mathcal{G}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\langle \sum_{\ell=1}^{n}\alpha_\ell\varphi(x_\ell),\varphi(x_i)\right\rangle_{\mathcal{F}}\langle g,\varphi(y_i)\rangle_{\mathcal{G}} \\
&= \frac{1}{n}\alpha^\top KL\beta
\end{aligned}$$

where $K_{ij} = k(x_i,x_j) {=} \langle\varphi(x_i),\varphi(x_j)\rangle_{\mathcal{F}}$ $\qquad L_{ij} = l(y_i,y_j)$.

Proof sketch (2)

Second step is covariance:

$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i)g(y_i)] = \frac{1}{n}\sum_{i=1}^{n}\langle f,\varphi(x_i)\rangle_{\mathcal{F}}\langle g,\varphi(y_i)\rangle_{\mathcal{G}}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\langle \sum_{\ell=1}^{n}\alpha_\ell\varphi(x_\ell),\varphi(x_i)\right\rangle_{\mathcal{F}}\langle g,\varphi(y_i)\rangle_{\mathcal{G}}$$

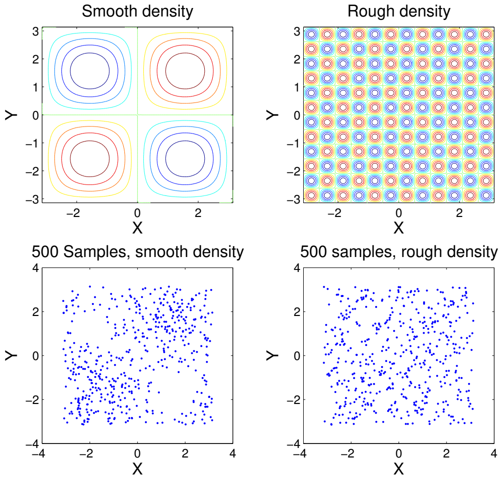
$$= \frac{1}{n}\alpha^\top KL\beta$$

where $K_{ij} = k(x_i,x_j) = \langle\varphi(x_i),\varphi(x_j)\rangle_{\mathcal{F}}$ $\qquad L_{ij} = l(y_i,y_j)$.

The Lagrangian is now:

$$\mathcal{L}(f,g,\lambda,\gamma) = \frac{1}{n}\alpha^\top KL\beta - \frac{\lambda}{2}\left(\alpha^\top K\alpha - 1\right) - \frac{\gamma}{2}\left(\beta^\top L\beta - 1\right)$$

What is a large dependence with COCO?



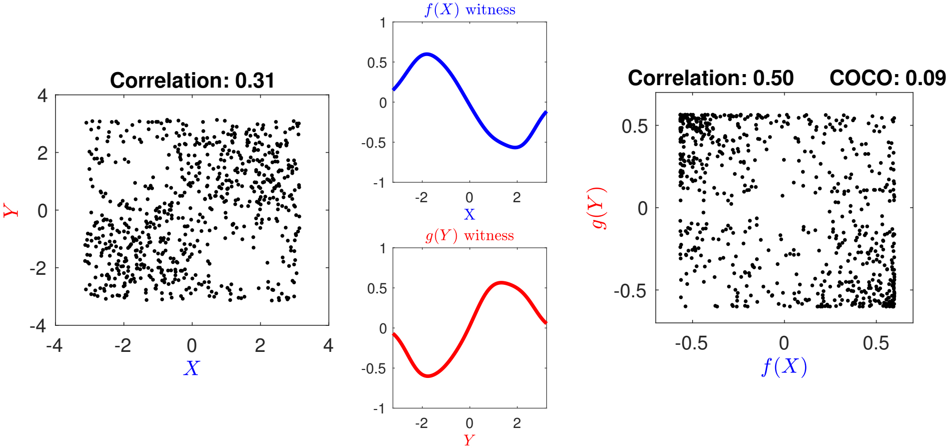
Which of these is the more “dependent”?

Density takes the form:

$$P_{XY} \propto 1 + \sin(\omega x)\sin(\omega y)$$

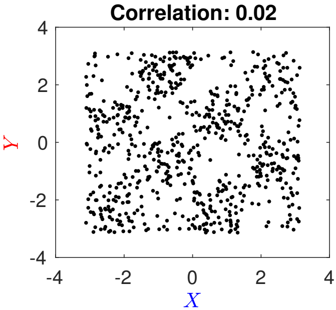
Finding covariance with smooth transformations

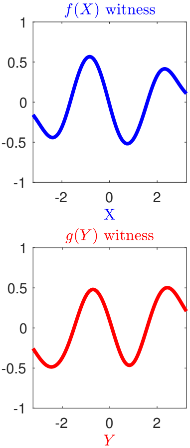
Case of $\omega = 1$:

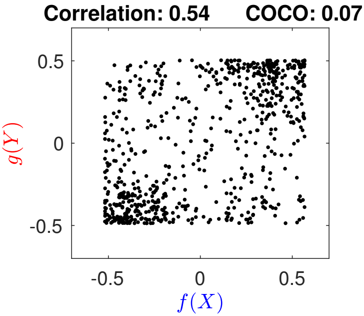


Finding covariance with smooth transformations

Case of $\omega = 2$:

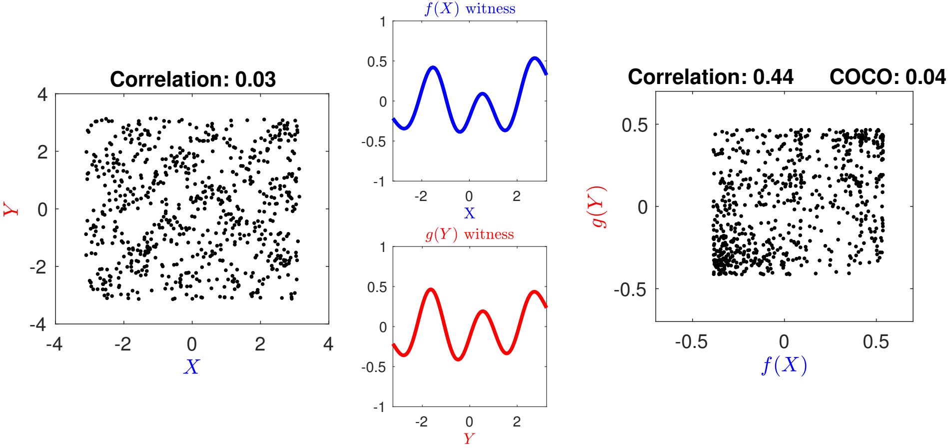






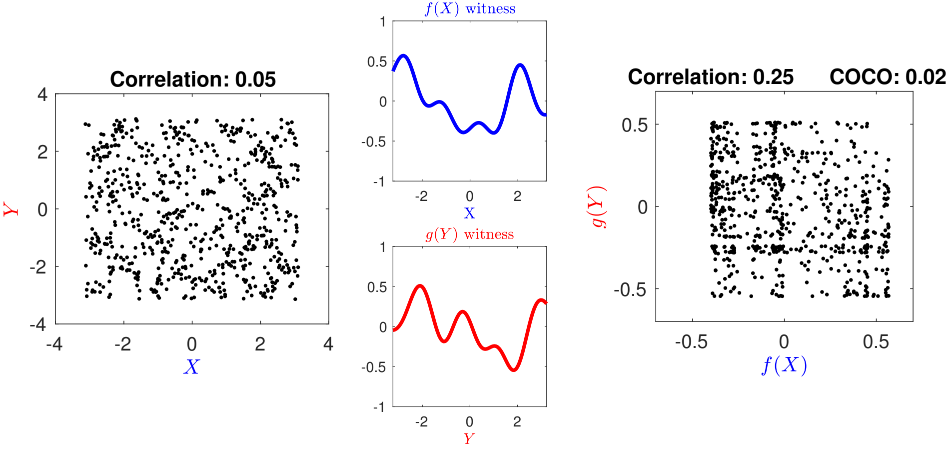
Finding covariance with smooth transformations

Case of $\omega = 3$:



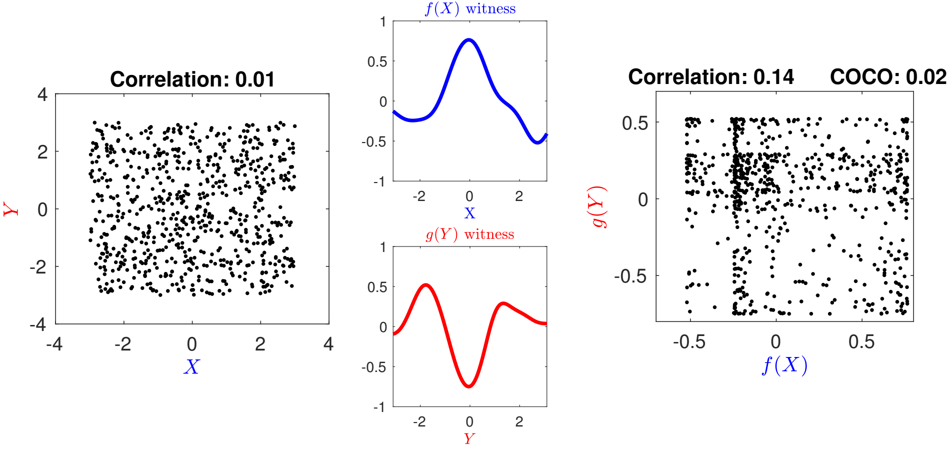
Finding covariance with smooth transformations

Case of $\omega = 4$:



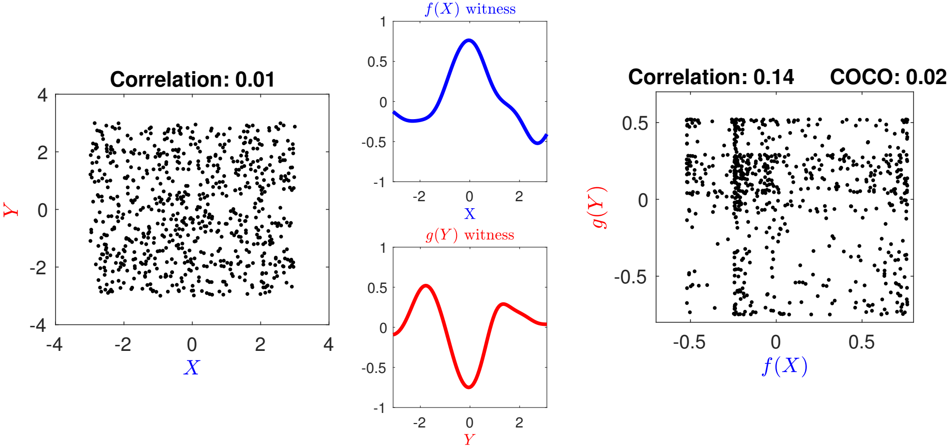
Finding covariance with smooth transformations

Case of $\omega = ??$:



Finding covariance with smooth transformations

Case of $\omega = 0$: uniform noise! (shows bias)



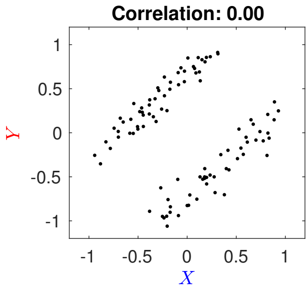
Dependence largest when at “low” frequencies

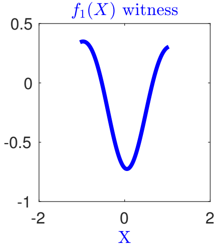
- As dependence is encoded at higher frequencies, the smooth mappings f, g achieve lower linear dependence.
- Even for independent variables, COCO will not be zero at finite sample sizes, since some mild linear dependence will be found by f,g (bias)
- This bias will decrease with increasing sample size.

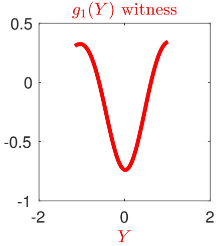
Can we do better than COCO?

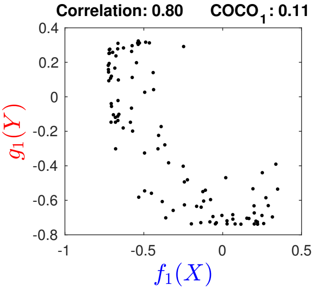
A second example with zero correlation.

First singular value of feature covariance $C_{\varphi(x)\phi(y)}$:





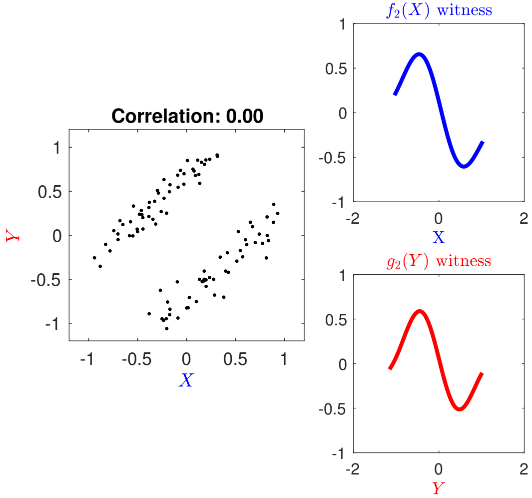




Can we do better than COCO?

A second example with zero correlation.

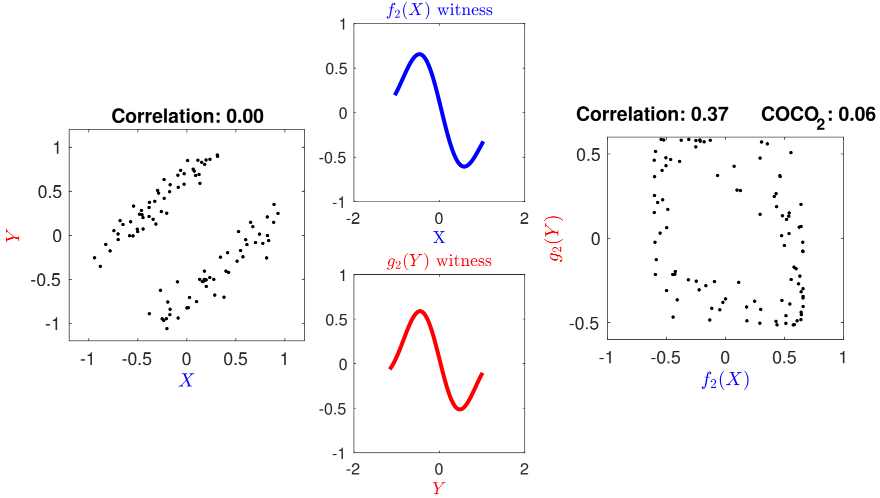
Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



Can we do better than COCO?

A second example with zero correlation.

Second singular value of feature covariance $C_{\varphi(x)\phi(y)}$:



The Hilbert-Schmidt Independence Criterion

Writing the ith singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC)

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$

G, Bousquet , Smola., and Schoelkopf, ALT05; G,., Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

The Hilbert-Schmidt Independence Criterion

Writing the ith singular value of the feature covariance $C_{\varphi(x)\phi(y)}$ as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC)

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2.$$

G, Bousquet , Smola., and Schoelkopf, ALT05; G,., Fukumizu, Teo., Song., Schoelkopf., and Smola, NIPS 2007,.

HSIC is MMD with product kernel!

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = MMD^2(P_{XY}, P_X P_Y; \mathcal{H}_\kappa)$$

where $\kappa((x,y),(x',y')) = k(x,x')l(y,y')$.

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{HSIC}$?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2}\text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α?
- Asymptotics of $\widehat{HSIC}$ when $P_{XY} = P_X P_Y$:

$$n\widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{HSIC}$?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2}\text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α?
- Asymptotics of $\widehat{HSIC}$ when $P_{XY} = P_X P_Y$:

$$n\widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{HSIC}$?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2}\text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α?
- Asymptotics of $\widehat{HSIC}$ when $P_{XY} = P_X P_Y$:

$$n\widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}, \quad h_{ijqr} = \frac{1}{4!} \sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

Asymptotics of HSIC under independence

- Given sample $\{(x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$, what is empirical $\widehat{HSIC}$?
- Empirical HSIC (biased)

$$\widehat{HSIC} = \frac{1}{n^2}\text{trace}(KL)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ (K and L computed with empirically centered features)

- Statistical testing: given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α?
- Asymptotics of $\widehat{HSIC}$ when $P_{XY} = P_X P_Y$:

$$n\widehat{HSIC} \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l z_l^2, \qquad z_l \sim \mathcal{N}(0,1)\text{i.i.d.}$$

where $\lambda_l \psi_l(z_j) = \int h_{ijqr} \psi_l(z_i) dF_{i,q,r}$, $\quad h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu} l_{tu} + k_{tu} l_{vw} - 2k_{tu} l_{tv}$

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- **Original time series:**

$$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$$
$$Y_1\ Y_2\ Y_3\ Y_4\ Y_5\ Y_6\ Y_7\ Y_8\ Y_9\ Y_{10}$$

- **Permutation:**

$$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$$
$$Y_7\ Y_3\ Y_9\ Y_2\ Y_4\ Y_8\ Y_5\ Y_1\ Y_6\ Y_{10}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?
- **Original time series:**

$$X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}$$
$$Y_1 \; Y_2 \; Y_3 \; Y_4 \; Y_5 \; Y_6 \; Y_7 \; Y_8 \; Y_9 \; Y_{10}$$

- **Permutation**:

$$X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}$$
$$Y_7 \; Y_3 \; Y_9 \; Y_2 \; Y_4 \; Y_8 \; Y_5 \; Y_1 \; Y_6 \; Y_{10}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC} > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- **Original time series:**

$$X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}$$
$$Y_1 \; Y_2 \; Y_3 \; Y_4 \; Y_5 \; Y_6 \; Y_7 \; Y_8 \; Y_9 \; Y_{10}$$

- **Permutation**:

$$X_1 \; X_2 \; X_3 \; X_4 \; X_5 \; X_6 \; X_7 \; X_8 \; X_9 \; X_{10}$$
$$Y_7 \; Y_3 \; Y_9 \; Y_2 \; Y_4 \; Y_8 \; Y_5 \; Y_1 \; Y_6 \; Y_{10}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

Application: dependence detection across languages

Testing task: detect dependence between English and French text

X	Y
Honourable senators, I have a question for the Leader of the Government in the Senate	Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat
No doubt there is great pressure on provincial and municipal governments	Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions
In fact, we have increased federal investments for early childhood development.	Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes
⋮	⋮

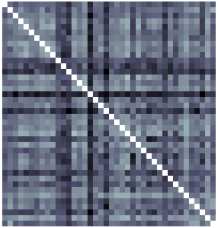
Text from the aligned hansards of the 36th parliament of canada, https://www.isi.edu/natural-language/download/hansard/

Application: dependence detection across languages

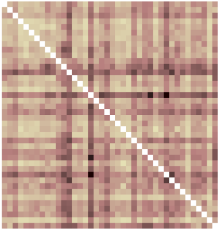
Testing task: detect dependence between English and French text

k-spectrum kernel, $k = 10$, sample size $n = 10$

X	Y
Honourable senators, I have a question for the Leader of the Government in the Senate	Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat
No doubt there is great pressure on provincial and municipal governments	Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions
In fact, we have increased federal investments for early childhood development.	Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes
⋮	⋮



K



L

$$\widehat{HSIC} = \frac{1}{n^2}\operatorname{trace}(K L)$$

(K and L column centered)

Application:Dependence detection across languages

Results (for $\alpha = 0.05$)

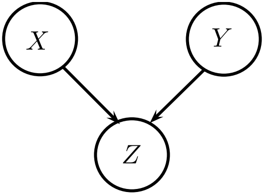
- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

Settings: Five line extracts, averaged over 300 repetitions, for "Agriculture" transcripts. Similar results for Fisheries and Immigration transcripts.

Testing higher order interactions

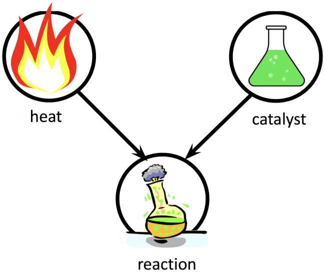
Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?



Detecting higher order interaction

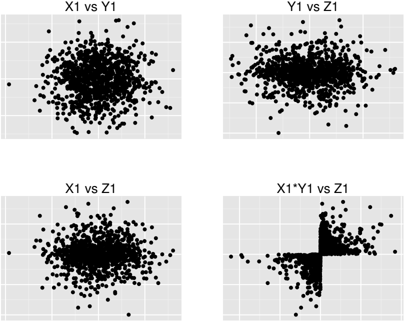
How to detect V-structures with pairwise weak individual dependence?

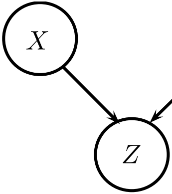


Detecting higher order interaction

How to detect V-structures with pairwise weak individual dependence?

$X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$

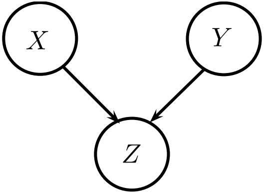




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\, X, Y \sim \text{sign}(XY) Exp(\frac{1}{\sqrt{2}})$

Fine print: Faithfulness violated here!

V-structure discovery



Assume $X \perp\!\!\!\perp Y$ has been established.

V-structure can then be detected by:

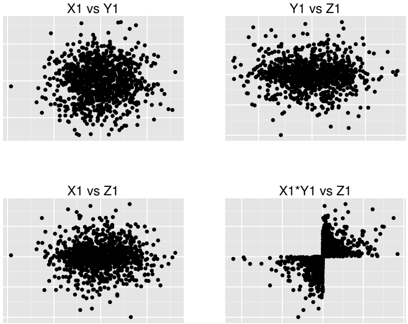
- Consistent CI test: $\mathbf{H_0} : X \perp\!\!\!\perp Y | Z$ [Fukumizu et al. 2008, Zhang et al. 2011]
- Factorisation test: $\mathbf{H_0} : (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$ (multiple standard two-variable tests)

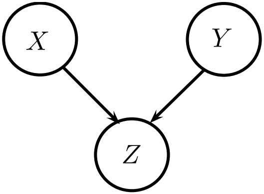
How well do these work?

Detecting higher order interaction

Generalise earlier example to p dimensions

$X \perp\!\!\!\perp Y, Y \perp\!\!\!\perp Z, X \perp\!\!\!\perp Z$

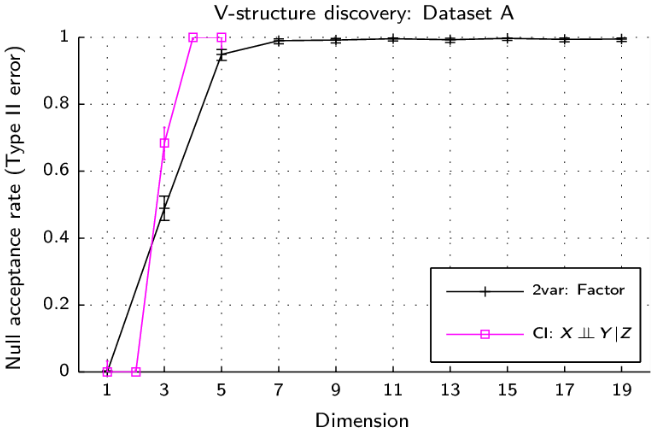




- $X, Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- $Z|\, X, Y \sim \text{sign}(XY) Exp(\frac{1}{\sqrt{2}})$
- $X_{2:p}, Y_{2:p}, Z_{2:p} \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{p-1})$

Fine print: Faithfulness violated here!

V-structure discovery



CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$$

Lancaster interaction measure

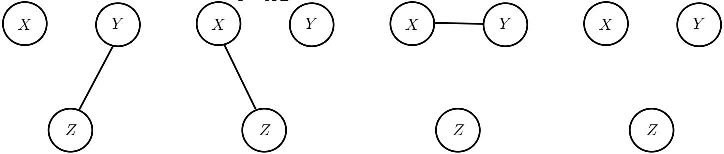
Lancaster interaction measure of $(X_1, \dots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$$

$\Delta_L P =$

P_{XYZ} $\quad -P_X P_{YZ}$ $\quad -P_Y P_{XZ}$ $\quad -P_Z P_{XY}$ $\quad +2 P_X P_Y P_Z$

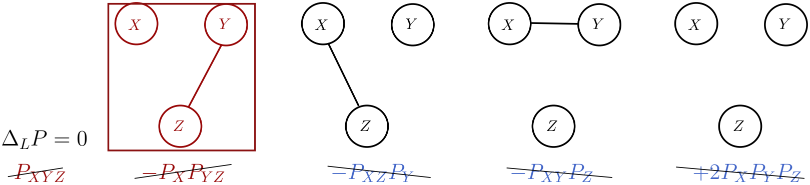


Lancaster interaction measure

Lancaster interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$$



Case of $P_X \perp\!\!\!\perp P_{YZ}$

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$$

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \;\Rightarrow\; \Delta_L P = 0.$$

...so what might be missed?

Lancaster interaction measure

Lancaster interaction measure of $(X_1, \ldots, X_D) \sim P$ is a signed measure ΔP that vanishes whenever P can be factorised non-trivially.

$$D = 2: \qquad \Delta_L P = P_{XY} - P_X P_Y$$

$$D = 3: \qquad \Delta_L P = P_{XYZ} - P_X P_{YZ} - P_Y P_{XZ} - P_Z P_{XY} + 2 P_X P_Y P_Z$$

$$\Delta_L P = 0 \not\Rightarrow (X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X$$

Example:

$P(0,0,0) = 0.2$	$P(0,0,1) = 0.1$	$P(1,0,0) = 0.1$	$P(1,0,1) = 0.1$
$P(0,1,0) = 0.1$	$P(0,1,1) = 0.1$	$P(1,1,0) = 0.1$	$P(1,1,1) = 0.2$

A kernel test statistic using Lancaster Measure

Construct a test by estimating $\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa}$, where $\kappa = k \otimes l \otimes m$:

$$\|\mu_\kappa(P_{XYZ} - P_{XY}P_Z - \cdots)\|^2_{\mathcal{H}_\kappa} =$$
$$\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XYZ} \rangle_{\mathcal{H}_\kappa} - 2\langle \mu_\kappa P_{XYZ}, \mu_\kappa P_{XY}P_Z \rangle_{\mathcal{H}_\kappa} \cdots$$

A kernel test statistic using Lancaster Measure

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_X P_Y P_Z$
P_{XYZ}	$(\mathbf{K} \circ \mathbf{L} \circ \mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{L})\,\mathbf{M})_{++}$	$((\mathbf{K} \circ \mathbf{M})\,\mathbf{L})_{++}$	$((\mathbf{M} \circ \mathbf{L})\,\mathbf{K})_{++}$	$tr(\mathbf{K}_+ \circ \mathbf{L}_+ \circ \mathbf{M}_+)$
$P_{XY}P_Z$		$(\mathbf{K} \circ \mathbf{L})_{++}\,\mathbf{M}_{++}$	$(\mathbf{MKL})_{++}$	$(\mathbf{KLM})_{++}$	$(\mathbf{KL})_{++}\mathbf{M}_{++}$
$P_{XZ}P_Y$			$(\mathbf{K} \circ \mathbf{M})_{++}\,\mathbf{L}_{++}$	$(\mathbf{KML})_{++}$	$(\mathbf{KM})_{++}\mathbf{L}_{++}$
$P_{YZ}P_X$				$(\mathbf{L} \circ \mathbf{M})_{++}\,\mathbf{K}_{++}$	$(\mathbf{LM})_{++}\mathbf{K}_{++}$
$P_X P_Y P_Z$					$\mathbf{K}_{++}\mathbf{L}_{++}\mathbf{M}_{++}$

Table: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2} \boxed{(HKH \circ HLH \circ HMH)_{++}}.$$

Empirical joint central moment in the feature space

A kernel test statistic using Lancaster Measure

$\nu \backslash \nu'$	P_{XYZ}	$P_{XY}P_Z$	$P_{XZ}P_Y$	$P_{YZ}P_X$	$P_XP_YP_Z$
P_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L) M)_{++}$	$((K \circ M) L)_{++}$	$((M \circ L) K)_{++}$	$tr(K_+ \circ L_+ \circ M_+)$
$P_{XY}P_Z$		$(K \circ L)_{++} M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++}M_{++}$
$P_{XZ}P_Y$			$(K \circ M)_{++} L_{++}$	$(KML)_{++}$	$(KM)_{++}L_{++}$
$P_{YZ}P_X$				$(L \circ M)_{++} K_{++}$	$(LM)_{++}K_{++}$
$P_XP_YP_Z$					$K_{++}L_{++}M_{++}$

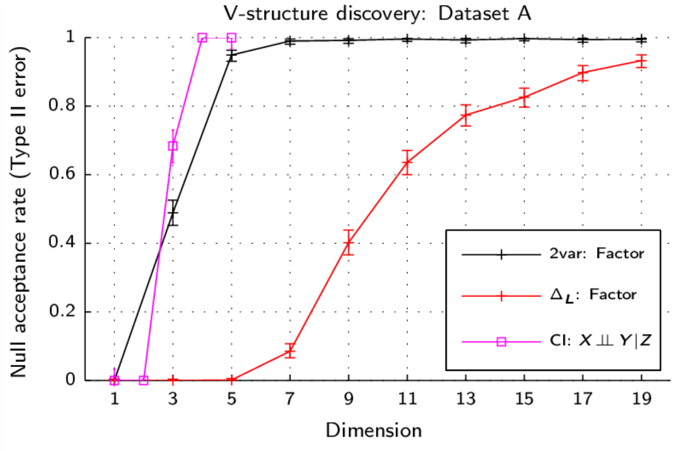
Table: V-statistic estimators of $\langle \mu_\kappa \nu, \mu_\kappa \nu' \rangle_{\mathcal{H}_\kappa}$ (without terms $P_X P_Y P_Z$). H is centering matrix $I - n^{-1}$

Lancaster interaction statistic: Sejdinovic, G, Bergsma, NIPS13

$$\|\mu_\kappa(\Delta_L P)\|^2_{\mathcal{H}_\kappa} = \frac{1}{n^2} \boxed{(HKH \circ HLH \circ HMH)_{++}}.$$

Empirical joint central moment in the feature space

V-structure discovery



Lancaster test, CI test for $X \perp\!\!\!\perp Y|Z$ from Zhang et al. (2011), and a factorisation test, $n = 500$

Interaction for $D \geq 4$

- Interaction measure valid for all D:

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! \, J_\pi P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.

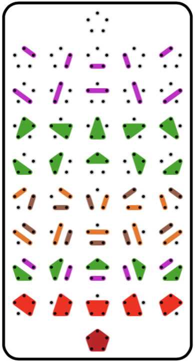
Interaction for $D \geq 4$

- Interaction measure valid for all D:

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! J_\pi P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



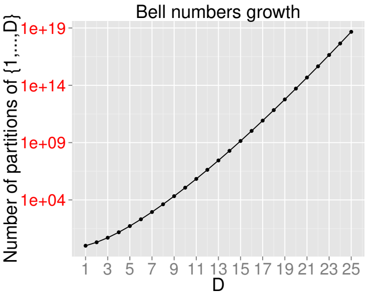
Interaction for $D \geq 4$

- Interaction measure valid for all D:

(Streitberg, 1990)

$$\Delta_S P = \sum_{\pi} (-1)^{|\pi|-1} \left(|\pi| - 1\right)! J_{\pi} P$$

  - For a partition π, J_π associates to the joint the corresponding factorisation, e.g., $J_{13|2|4} P = P_{X_1 X_3} P_{X_2} P_{X_4}$.



Co-authors

From Gatsby:

- Mikolaj Binkowski
- Kacper Chwialkowski
- Wittawat Jitkrittum
- Heiko Strathmann
- Dougal Sutherland
- Wenkai Xu

External collaborators:

- Kenji Fukumizu
- Bernhard Schoelkopf
- Dino Sejdinovic
- Bharath Sriperumbudur
- Alex Smola
- Zoltan Szabo

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT