Kernel Constrained Covariance for Dependence Measurement

Arthur Gretton
arthur@tuebingen.mpg.de

MPI for Biological Cybernetics, Tübingen

with Alexander Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Bernhard Schölkopf, Nikos Logothetis

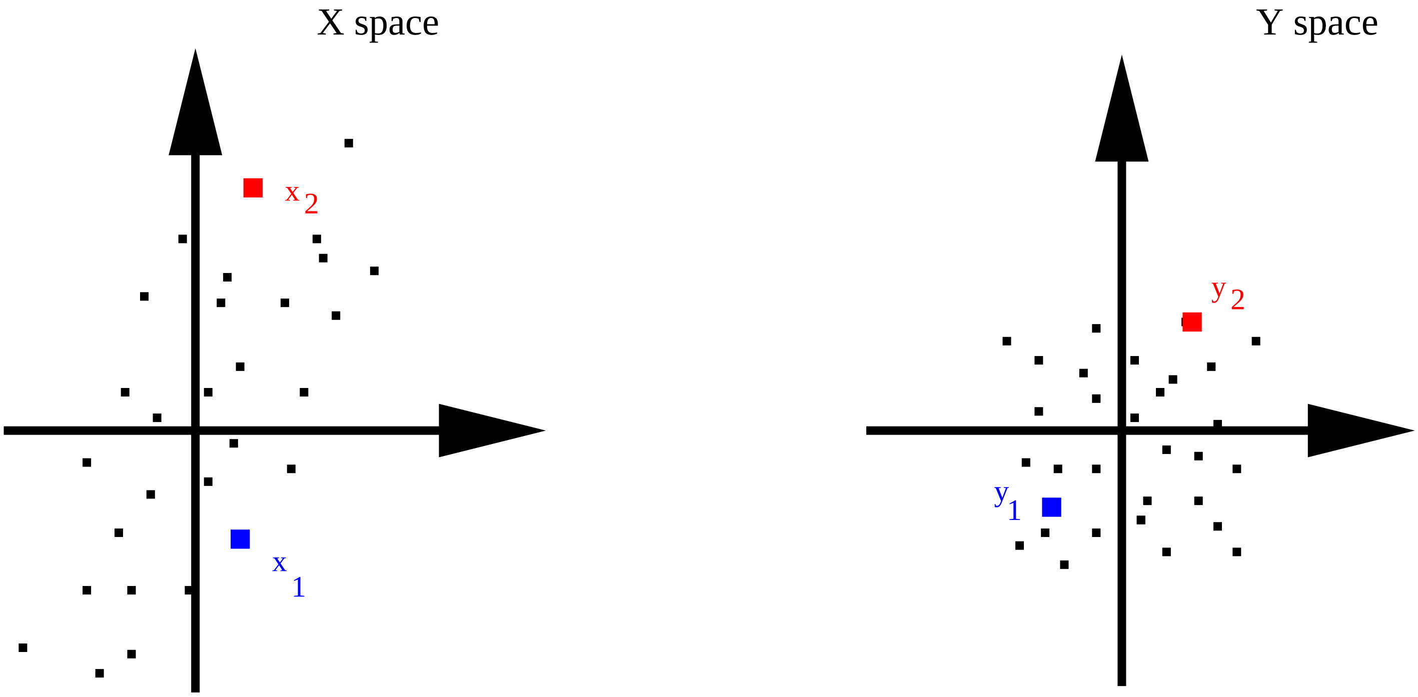
Kernel dependence measures

- Kernel dependence measures
  - Tests for independence
  - Take into account high order moments
  - Make “sensible” assumptions about smoothness
- Applications
  - Independent component analysis (ICA)
  - Feature selection (Fukumizu *et al.*)
  - Dependence detection between voxel activity in Macaque visual cortex (V1)
    - Effect of removing breathing artifacts

Talk outline

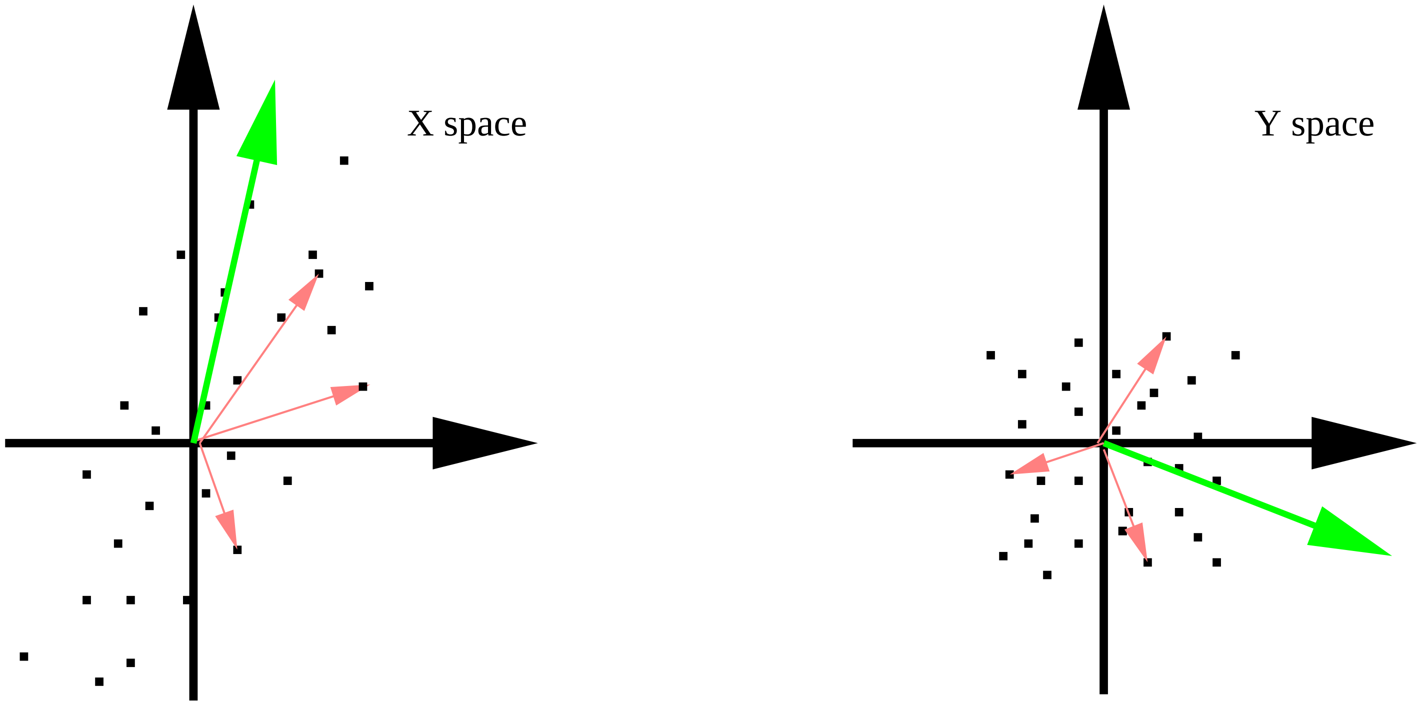
- Constrained covariance (COCO)
  - Covariance in RKHSs
  - Three useful properties of COCO
    - Independence test when kernels universal
    - For finite samples, no “good” dependence tests...
    - ...but no slow learning rates as samples increase
  - The third fact is not proved for other kernel dependence tests!

Dependence detection



- Get m pairs of points in different spaces
- Are the RVs x and y dependent?

A second order method



- Choose directions, get dot product with all points.
- Directions chosen such that the vectors of projections have biggest covariance. Is covariance 0?

Take *nonlinear* features

- Points in each space mapped to vectors of nonlinear features:
  - $x \to \mathbf{x} := \left[\sqrt{\lambda_1}\varphi_1(x) \ \ \sqrt{\lambda_2}\varphi_2(x) \ \ \ldots \ \ \sqrt{\lambda_n}\varphi_n(x) \ \ \ldots\right]$
  - $y \to \mathbf{y} := \left[\sqrt{\lambda_1}\varphi_1(y) \ \ \sqrt{\lambda_2}\varphi_2(y) \ \ \ldots \ \ \sqrt{\lambda_n}\varphi_n(y) \ \ \ldots\right]$
  - $\mathbf{x} \in \mathcal{H}_\mathcal{X}$ and $\mathbf{y} \in \mathcal{H}_\mathcal{Y}$, can be infinite dimensional
  - As n increases, λ_n smaller and φ_n less smooth
- Define projection vectors in each space: $\mathbf{f} \in \mathcal{H}_\mathcal{X}$, $\mathbf{g} \in \mathcal{H}_\mathcal{Y}$.
- Formal definition of COCO:

$$\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}) := \sup_{f \in \mathcal{H}_\mathcal{X},\, g \in \mathcal{H}_\mathcal{Y}} \frac{\mathrm{cov}\left(\mathbf{f}^\top \mathbf{x}, \mathbf{g}^\top \mathbf{y}\right)}{\|\mathbf{f}\|_{\mathcal{H}_\mathcal{X}} \|\mathbf{g}\|_{\mathcal{H}_\mathcal{Y}}}$$

The kernel trick (1)

- Must we really consider infinite dimensional vectors?
- Differentiating COCO wrt $\mathbf{f}$ and $\mathbf{g}$, want biggest eigenvalue

$$\begin{bmatrix} \mathbf{0} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{f}_i \\ \mathbf{g}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \mathbf{f}_i \\ \mathbf{g}_i \end{bmatrix}$$

- When we rely on a finite sample,

$$\widehat{\mathbf{C}}_{xy} = \begin{bmatrix} \mathbf{x}_1 & \ldots & \mathbf{x}_m \end{bmatrix} \mathbf{H} \begin{bmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_m^\top \end{bmatrix}$$

The kernel trick (2)

- This means:

$$\mathbf{f} = \sum_{l=1}^{m} c_l \mathbf{x}_l,$$

$$\mathbf{g} = \sum_{l=1}^{m} d_l \mathbf{y}_l.$$

- Inner product in reproducing kernel Hilbert spaces given by kernel

$$\mathbf{x}_1^\top \mathbf{x}_2 = k\left(x_1 - x_2\right)$$

$$\mathbf{y}_1^\top \mathbf{y}_2 = k\left(y_1 - y_2\right)$$

An empirical estimate

- Constrained covariance then **largest eigenvalue** γ_i **of**

$$\begin{bmatrix} \mathbf{0} & \widetilde{\mathbf{K}}_{mm}^{(x)}\widetilde{\mathbf{K}}_{mm}^{(y)} \\ \widetilde{\mathbf{K}}_{mm}^{(y)}\widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \gamma_i \begin{bmatrix} \widetilde{\mathbf{K}}_{mm}^{(x)} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{K}}_{mm}^{(y)} \end{bmatrix} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix}.$$

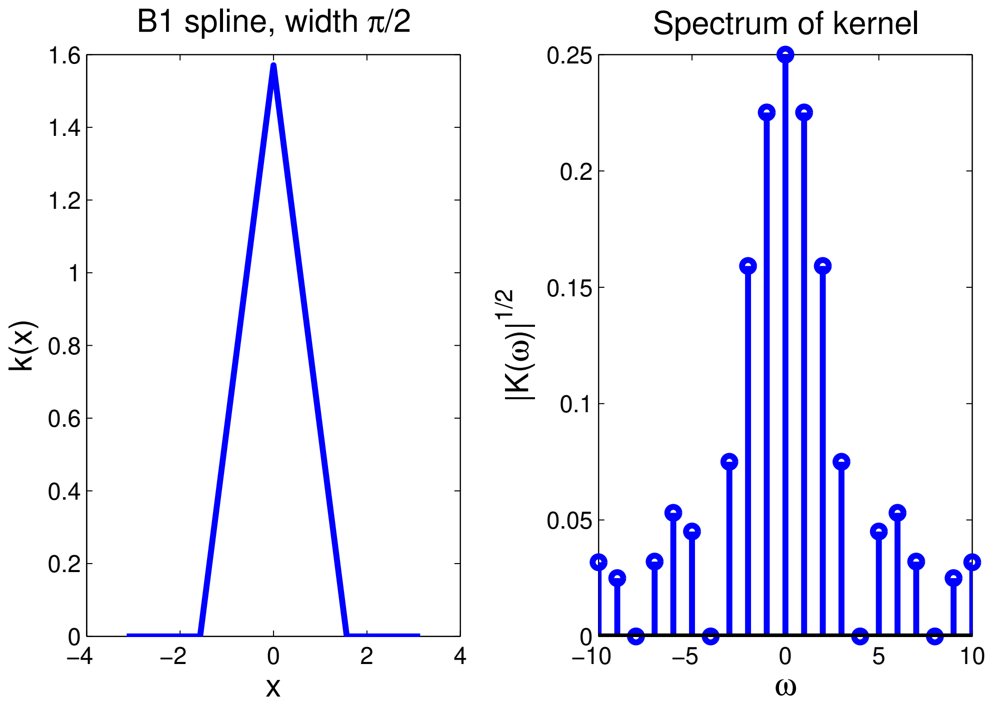
- $\widetilde{\mathbf{K}}_{mm}^{(x)}$ is matrix of inner products between centred observations in feature space.

Fact 1: COCO tests independence

- $\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) = 0$ iff x, y independent, when $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are RKHS induced by universal kernels (eg. Gaussian kernels, Laplace kernels, ...)
- Also true of
  - Kernel canonical correlation: as above, but normalising by the variance in the RKHS (Bach & Jordan)
  - Kernel mutual information: an upper bound on the MI near independence (Gretton *et al.*)
  - Kernel generalised variance: a looser upper bound on the MI near independence (Bach & Jordan)

Why universal?

- What happens when kernel is not universal?
- Example: spline kernel



Fact 2: Limitations of dep. tests

- NO test can detect all dependence for finite samples.
- Example: Set $\mathcal{P}$ of prob. distrib. $\mathsf{P}_{\mathbf{x}}$ over n variables
  - $\mathcal{P}_i$ generates independent random variables,
  - $\mathcal{P}_d$ gives dependent RVs
- Test: $\Delta(\mathbf{X})$ takes m i.i.d. samples, returns

$$\Delta(\mathbf{X}) = 1 \ : \ \mathbf{X} \sim \mathsf{P}^{(d)}_{\mathbf{x}^m}, \qquad \Delta(\mathbf{X}) = 0 \ : \ \mathbf{X} \sim \mathsf{P}^{(i)}_{\mathbf{x}^m}$$

- Uncertainty due to empirical estimate: α-test

$$\sup_{\mathsf{P}^{(i)}_{\mathbf{x}} \in \mathcal{P}_i} \mathsf{E}_{\mathbf{X} \sim \mathsf{P}^{(i)}_{\mathbf{x}^m}} (\Delta(\mathbf{X}) = 1) \leq \alpha;$$

- There exists $\mathsf{P}_{\mathbf{x}} \notin \mathcal{P}_i$ such that for small ϵ,

$$\mathsf{P}_{\mathbf{X} \sim \mathsf{P}_{\mathbf{x}^m}} (\Delta(\mathbf{X}) = 0) \geq 1 - \alpha - \epsilon;$$

Lim. of COCO as dep. test (2)

- COCO can be ≈ 0 for dependent RVs with highly non-smooth densities:

$$\mathsf{P}_{\mathsf{x},\mathsf{y}} = \alpha + \beta \varphi_l(\mathsf{x}) \varphi_l(\mathsf{y}),$$

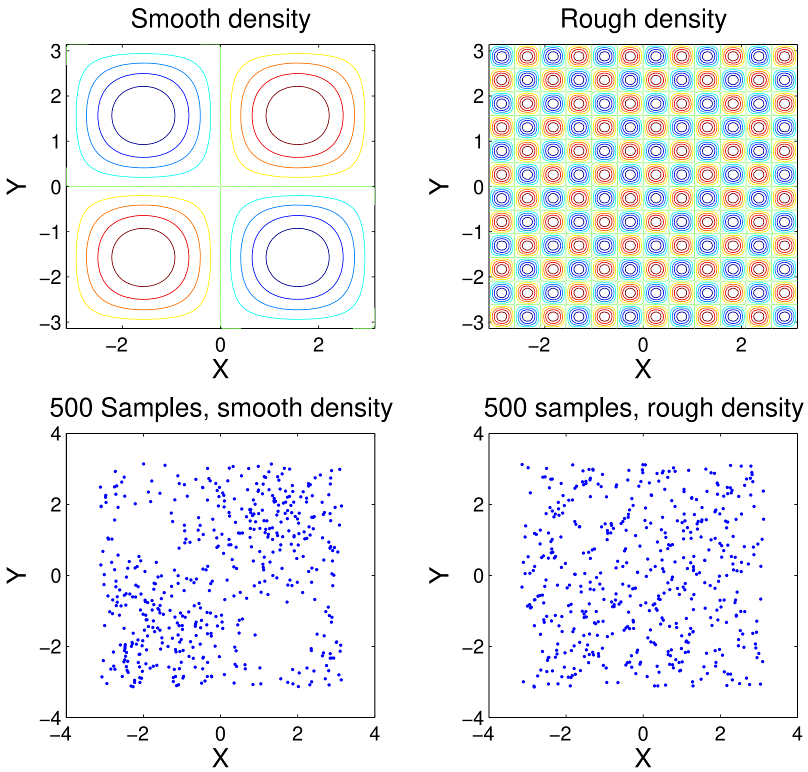
  - l large
  - β non-trivial

- COCO "as small as you want" (depends on l)
- Reason: norms in the denominator

$$\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) := \sup_{f \in \mathcal{H}_{\mathcal{X}},\, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\mathrm{cov}\left(\mathbf{f}^\top \mathbf{x}, \mathbf{g}^\top \mathbf{y}\right)}{\|\mathbf{f}\|_{\mathcal{H}_{\mathcal{X}}} \|\mathbf{g}\|_{\mathcal{H}_{\mathcal{Y}}}}$$

- RESULT: not detectable with finite sample size

Lim. of COCO as dep. test (3)

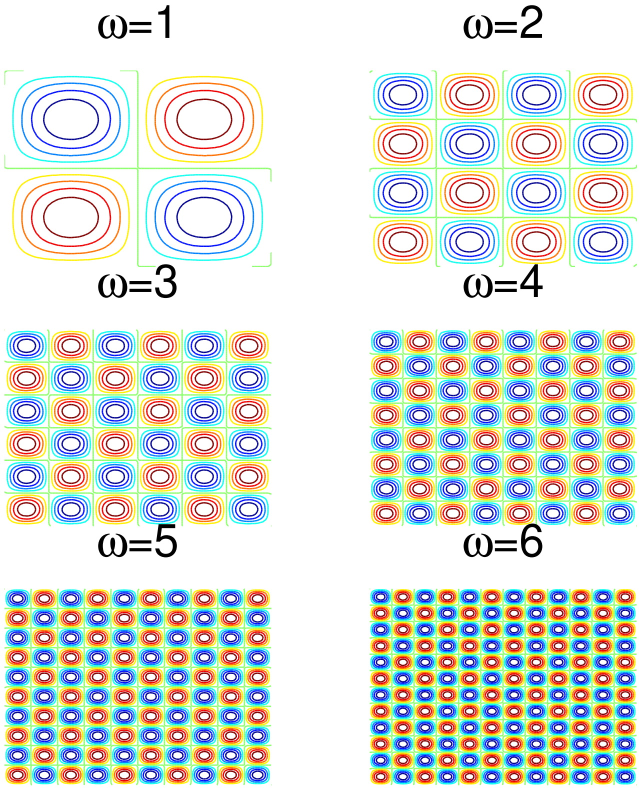


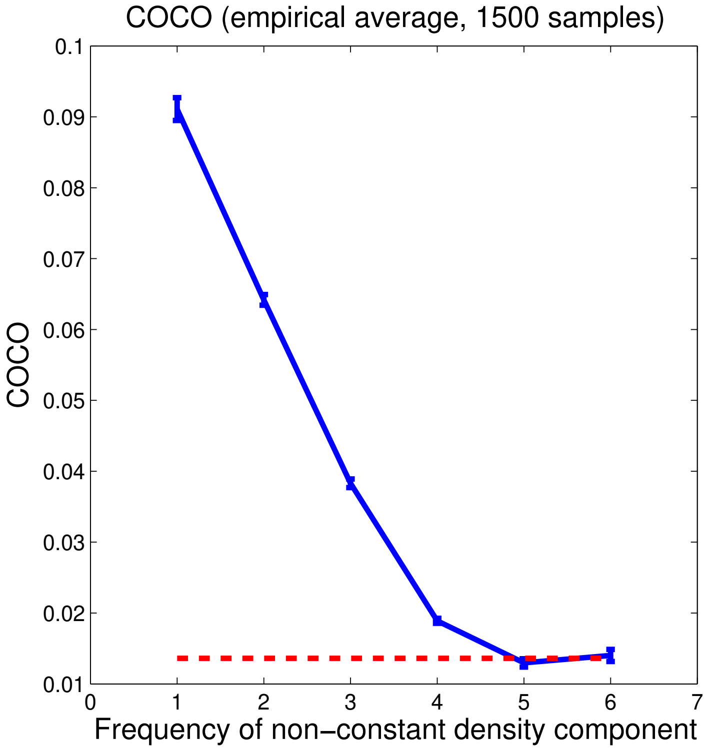
Density takes the form:

$$P_{x,y} = 1 + \sin(\omega x)\sin(\omega y)$$

Lim. of COCO as dep. test (4)

- Example: sinusoids of increasing frequency





Fact 3: No slow learning rates

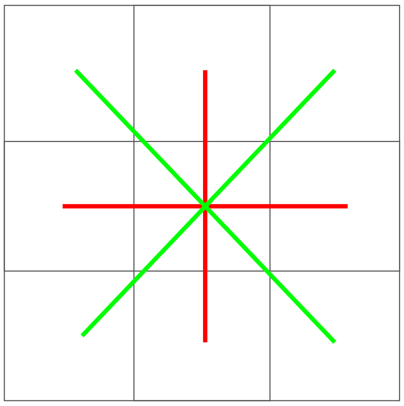
- Empirical COCO converges to the population COCO at speed $1/\sqrt{n}$.
- A dependence test: $\Delta(\mathbf{z})$ is the indicator that COCO larger than $C\sqrt{\log(1/\alpha)/n}$
- $\Delta(\mathbf{z})$ is an α-test
  - Reminder: α upper bounds prob. that test returns dependence when random variables independent
- Type II approaches zero as $1/\sqrt{n}$.
  - Reminder: Type II error is prob. that test returns independence when random variables dependent
- No slow learning rates for dependence tests!
- Finite sample results!

Conclusions

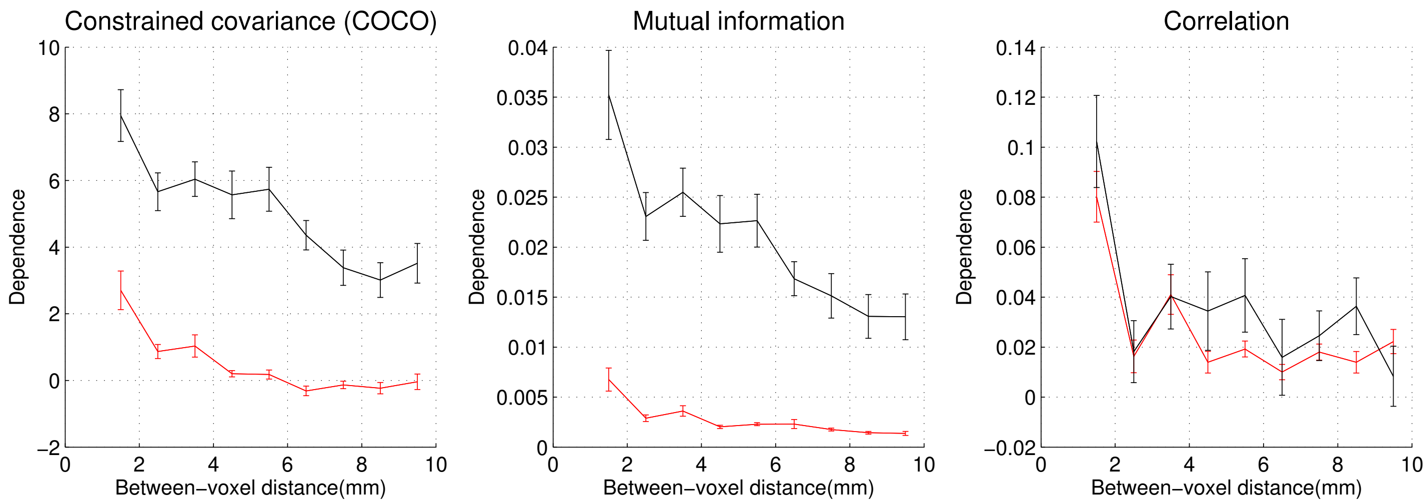
- COCO: spectral norm of covariance operator between RKHSs
- For universal RKHSs, a test of independence
- Small dependence in the case where “rougher” densities
- Fast, finite sample convergence (unique!)
- No slow learning rates for independence tests based on COCO

Breathing artifact removal

- Diverse (film-based) stimulus shown to Macaque
- MRI time series sampled at 4Hz (1200 samples $\times 2$)
- Dependence between pairs of voxels of similar distance
- Breathing component at approx 0.4Hz (also harmonics)
- Sinusoid projected directly out of time series (sidelobes)
- Goal: minimal loss of MRI signal.



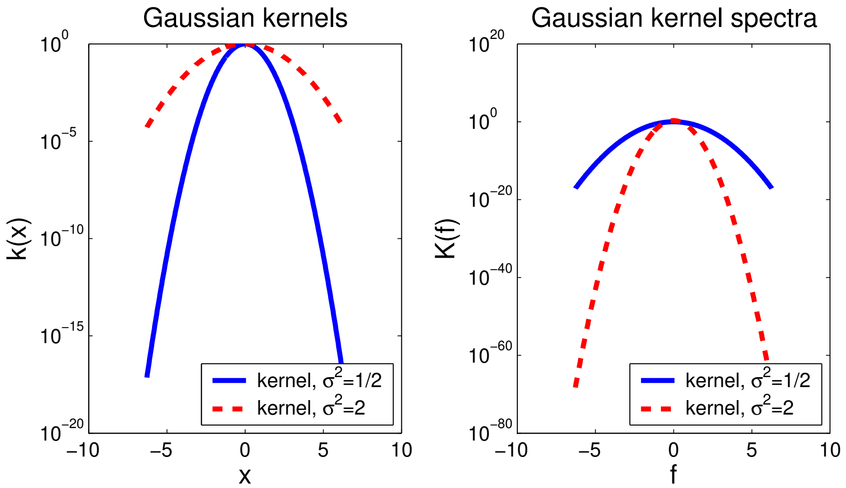
Breathing removal: results



- Red: after breathing removal. Black: before removal
- Zero point is dependence wrt test region
  - 2nd order method underestimates dependence with breathing (phase shift)
  - High order methods give 5mm dependence range
  - COCO gives greater increase wrt test region dep.

Lim. of COCO as dep. test (5)

- Reminder: the RKHS norm of a function is $\|f\|^2_{\mathcal{H}_\mathcal{X}} := \sum_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i\right)^{-1}$.
- If kernel decays quickly, its spectrum decays slowly:
  - then non-smooth functions have smaller RKHS norm
- Example: spectrum of two Gaussian kernels



Lim. of COCO as dep. test (6)

- Could we just decrease kernel size?
- Yes, but only up to a point