

Kernel-Based Contrast Functions for Sufficient Dimension Reduction

K. Fukumizu, F. Bach, & M. I. Jordan, (2009).
Annals of Statistics, 37, 1871-1905.

Outline

- Introduction
 - dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Sufficient Dimension Reduction

- Regression setting: observe (X, Y) pairs, where the covariate X is high-dimensional
- Find a (hopefully small) subspace S of the covariate space that retains the information pertinent to the response Y
- *Semiparametric formulation*: treat the conditional distribution $p(Y | X)$ nonparametrically, and estimate the parameter S

Perspectives

- Classically the covariate vector X has been treated as ancillary in regression
- The sufficient dimension reduction (SDR) literature has aimed at making use of the randomness in X (in settings where this is reasonable)
- This has generally been achieved via inverse regression
 - at the cost of introducing strong assumptions on the distribution of the covariate X
- We'll make use of the randomness in X without employing inverse regression

Dimension Reduction for Regression

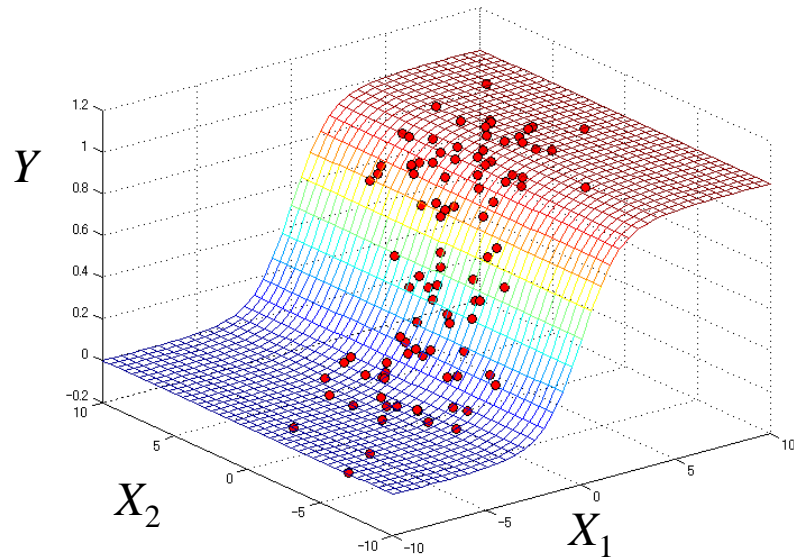
- Regression: $p(Y | X)$

Y : response variable,

$X = (X_1, \dots, X_m)$: m -dimensional covariate

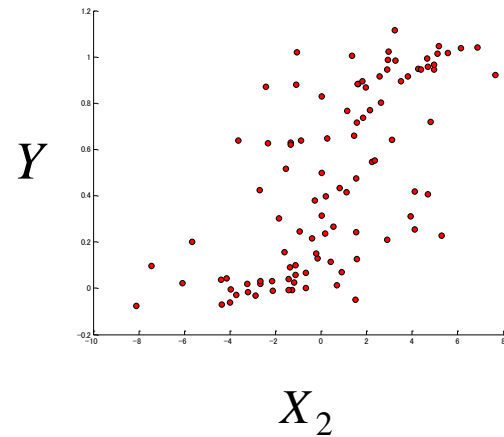
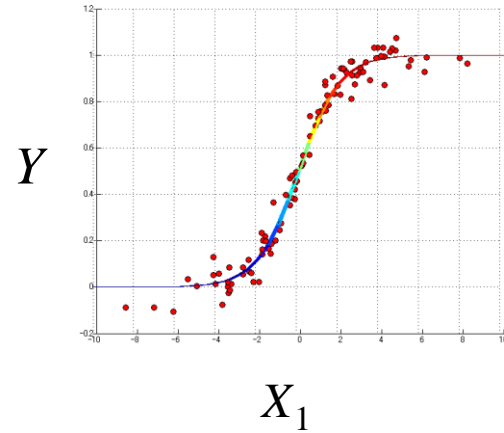
- Goal: Find the **central subspace**, which is defined via:

$$p(Y | X) = \tilde{p}(Y | b_1^T X, \dots, b_d^T X) \quad \left(= \tilde{p}(Y | B^T X) \right)$$



$$Y = \frac{1}{1 + \exp(-X_1)} + N(0; 0.1^2)$$

central subspace = X_1 axis



Some Existing Methods

- Sliced Inverse Regression (SIR, Li 1991)
 - PCA of $E[X|Y]$ → use slice of Y
 - Elliptic assumption on the distribution of X
- Principal Hessian Directions (pHd, Li 1992)
 - Average Hessian $\Sigma_{yxx} \equiv E[(Y - \bar{Y})(X - \bar{X})(X - \bar{X})^T]$ is used
 - If X is Gaussian, eigenvectors gives the central subspace
 - Gaussian assumption on X . Y must be one-dimensional
- Projection pursuit approach (e.g., Friedman et al. 1981)
 - Additive model $E[Y|X] = g_1(b_1^T X) + \dots + g_d(b_d^T X)$ is used
- Canonical Correlation Analysis (CCA) / Partial Least Squares (PLS)
 - Linear assumption on the regression
- Contour Regression (Li, Zha & Chiaromonte, 2004)
 - Elliptic assumption on the distribution of X

Dimension Reduction and Conditional Independence

- $(U, V) = (B^T X, C^T X)$

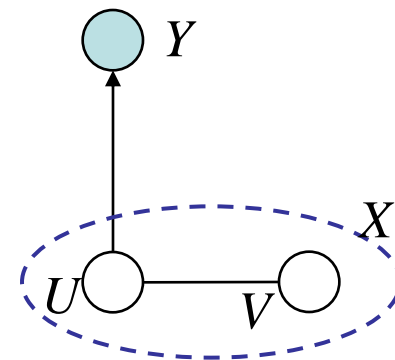
where $C: m \times (m-d)$ with columns orthogonal to B

- B gives the projector onto the central subspace

$$\Leftrightarrow p_{Y|X}(y|x) = p_{Y|U}(y|B^T x)$$

$$\Leftrightarrow p_{Y|U,V}(y|u,v) = p_{Y|U}(y|u) \quad \text{for all } y, u, v$$

$$\Leftrightarrow \text{Conditional independence} \quad Y \perp\!\!\!\perp V | U$$



- Our approach: *Characterize conditional independence*

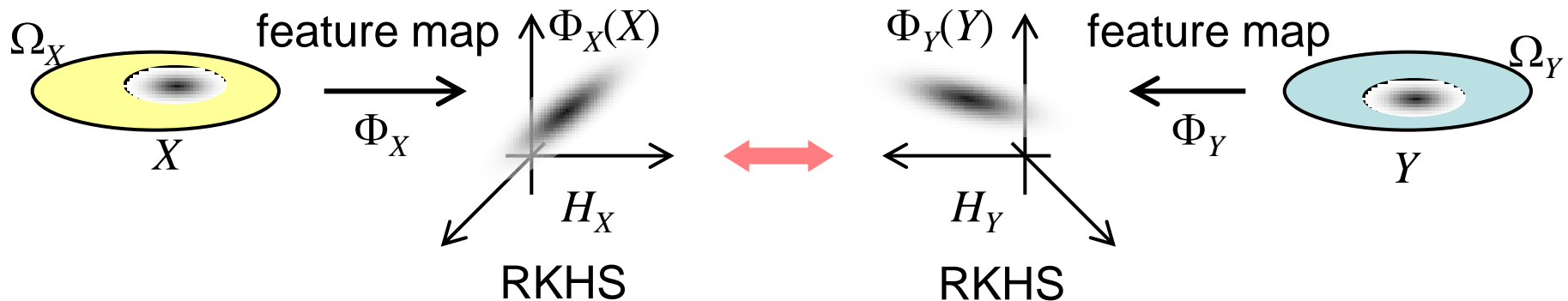
Outline

- Introduction
 - dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Reproducing Kernel Hilbert Spaces

■ “Kernel methods”

- RKHS’s have generally been used to provide basis expansions for regression and classification (e.g., support vector machine)
- *Kernelization*: map data into the RKHS and apply linear or second-order methods in the RKHS
- But RKHS’s can also be used to characterize independence and conditional independence



Positive Definite Kernels and RKHS

■ Positive definite kernel (p.d. kernel)

$$k: \Omega \times \Omega \rightarrow \mathbf{R}$$

k is **positive definite** if $k(x,y) = k(y,x)$ and for any $n \in \mathbf{N}$, $x_1, \dots, x_n \in \Omega$ the matrix $\left(k(x_i, x_j)\right)_{i,j}$ (Gram matrix) is positive semidefinite.

– Example: Gaussian RBF kernel $k(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right)$

■ Reproducing kernel Hilbert space (RKHS)

k : p.d. kernel on Ω

$\Rightarrow \exists H$: reproducing kernel Hilbert space (RKHS)

1) $k(\cdot, x) \in H$ for all $x \in \Omega$.

2) $\text{Span} \{k(\cdot, x) \mid x \in \Omega\}$ is dense in H .

3) $\langle k(\cdot, x), f \rangle_H = f(x)$ (reproducing property)

■ Functional data

$$\Phi : \Omega \rightarrow H, \quad x \mapsto k(\cdot, x) \quad \text{i.e.} \quad \Phi(x) = k(\cdot, x)$$

Data: $X_1, \dots, X_N \rightarrow \Phi_X(X_1), \dots, \Phi_X(X_N) : \text{functional data}$

■ Why RKHS?

- By the reproducing property, computing the inner product on RKHS is easy:

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

$$f = \sum_{i=1}^N a_i \Phi(x_i) = \sum_i a_i k(\cdot, x_i), \quad g = \sum_{j=1}^N b_j \Phi(x_j) = \sum_j b_j k(\cdot, x_j)$$

$$\Leftrightarrow \langle f, g \rangle = \sum_{i,j} a_i b_j k(x_i, x_j)$$

- The computational cost essentially depends on the sample size. Advantageous for high-dimensional data of small sample size.

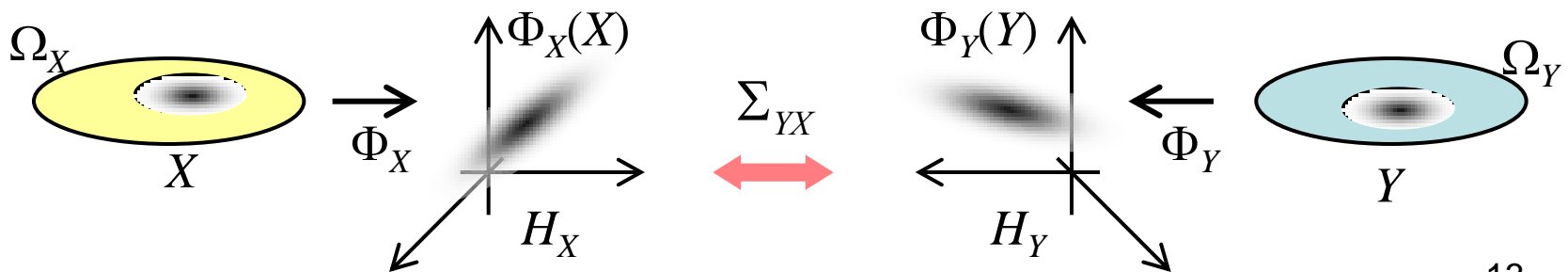
Covariance Operators on RKHS

- X, Y : random variables on Ω_X and Ω_Y , resp.
- Prepare RKHS (H_X, k_X) and (H_Y, k_Y) defined on Ω_X and Ω_Y , resp.
- Define **random variables on the RKHS** H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

- Define the **covariance operator** Σ_{YX}

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$



Covariance Operators on RKHS

- Definition

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle\Phi_X(X), \cdot\rangle] - E[\Phi_Y(Y)]E[\langle\Phi_X(X), \cdot\rangle]$$

Σ_{YX} is an **operator** from H_X to H_Y such that

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y)f(X)] - E[g(Y)]E[f(X)] \quad (= \text{Cov}[f(X), g(Y)])$$

for all $f \in H_X, g \in H_Y$

- *cf.* Euclidean case

$$V_{YX} = E[YX^T] - E[Y]E[X]^T \quad : \text{covariance matrix}$$

$$(b, V_{YX} a) = \text{Cov}[(b, Y), (a, X)]$$

Characterization of Independence

- Independence and cross-covariance operators

If the RKHS's are "rich enough":

$$X \perp\!\!\!\perp Y \iff \Sigma_{XY} = O$$



$$\text{Cov}[f(X), g(Y)] = 0$$

or

$$E[g(Y)f(X)] = E[g(Y)]E[f(X)]$$

for all $f \in H_X, g \in H_Y$

\Rightarrow is always true

\Leftarrow requires an assumption
on the kernel (universality)

e.g., Gaussian RBF kernels are
universal

$$k(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$$

– cf. for Gaussian variables,

$$X \text{ and } Y \text{ are independent} \iff V_{XY} = O \quad \text{i.e. uncorrelated}$$

- Independence and characteristic functions

Random variables X and Y are independent

$$\Leftrightarrow E_{XY} \left[e^{i\omega^T X} e^{i\eta^T Y} \right] = E_X \left[e^{i\omega^T X} \right] E_Y \left[e^{i\eta^T Y} \right] \quad \text{for all } \omega \text{ and } \eta$$

I.e., $e^{i\omega^T x}$ and $e^{i\eta^T y}$ work as test functions

- RKHS characterization

Random variables $X \in \Omega_X$ and $Y \in \Omega_Y$ are independent

$$\Leftrightarrow E_{XY} [f(X)g(Y)] = E_X [f(X)] E_Y [g(Y)] \quad \text{for all } f \in \mathcal{H}_X, g \in \mathcal{H}_Y$$

– RKHS approach is a generalization of the characteristic-function approach

RKHS and Conditional Independence

- **Conditional covariance operator**

X and Y are random vectors. $\mathcal{H}_X, \mathcal{H}_Y$: RKHS with kernel k_X, k_Y , resp.

Def. $\Sigma_{YY|X} \equiv \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$: **conditional covariance operator**

– Under a universality assumption on the kernel

$$\langle g, \Sigma_{YY|X} g \rangle = E[\text{Var}[g(Y) | X]]$$

cf. For Gaussian $\text{Var}_{Y|X}[a^T Y | X = x] = a^T (V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}) a$

– Monotonicity of conditional covariance operators

$X = (U, V)$: random vectors

$$\Sigma_{YY|U} \geq \Sigma_{YY|X}$$

\geq : in the sense of self-adjoint operators

- Conditional independence

Theorem

$X = (U, V)$ and Y are random vectors.

$\mathcal{H}_X, \mathcal{H}_U, \mathcal{H}_Y$: RKHS with **Gaussian kernel** k_X, k_U, k_Y , resp.

 $Y \perp\!\!\!\perp V | U \iff \Sigma_{YY|U} = \Sigma_{YY|X}$

This theorem provides a new methodology for solving the sufficient dimension reduction problem

Outline

- Introduction
 - dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Kernel Dimension Reduction

- Use a **universal kernel** for $B^T X$ and Y

$$\Sigma_{YY|B^T X} \geq \Sigma_{YY|X}$$

(\geq : the partial order of self-adjoint operators)

$$\Sigma_{YY|B^T X} = \Sigma_{YY|X} \iff X \perp\!\!\!\perp Y \mid B^T X$$

- KDR objective function:

$$\min_{B: B^T B = I_d} \text{Tr} \left[\Sigma_{YY|B^T X} \right]$$

which is an optimization over the Stiefel manifold

Estimator

- Empirical cross-covariance operator

$$\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N} \sum_{i=1}^N \{k_Y(\cdot, Y_i) - \hat{m}_Y\} \otimes \{k_X(\cdot, X_i) - \hat{m}_X\}$$

$$\hat{m}_X = \frac{1}{N} \sum_{i=1}^N k_X(\cdot, X_i) \quad \hat{m}_Y = \frac{1}{N} \sum_{i=1}^N k_Y(\cdot, Y_i)$$

$\hat{\Sigma}_{YX}^{(N)}$ gives the empirical covariance:

$$\left\langle g, \hat{\Sigma}_{YX}^{(N)} f \right\rangle = \frac{1}{N} \sum_{i=1}^N f(X_i) g(Y_i) - \frac{1}{N} \sum_{i=1}^N f(X_i) \frac{1}{N} \sum_{i=1}^N g(Y_i)$$

- Empirical conditional covariance operator

$$\hat{\Sigma}_{YY|X}^{(N)} = \hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YX}^{(N)} \left(\hat{\Sigma}_{XX}^{(N)} + \varepsilon_N I \right)^{-1} \hat{\Sigma}_{XY}^{(N)}$$

ε_N : regularization coefficient

- Estimating function for KDR:

$$\begin{aligned} \text{Tr} \left[\hat{\Sigma}_{YY|U}^{(N)} \right] &= \text{Tr} \left[\hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YU}^{(N)} \left(\hat{\Sigma}_{UU}^{(N)} + \varepsilon_N I \right)^{-1} \hat{\Sigma}_{UY}^{(N)} \right] & U = B^T X \\ &= \text{Tr} \left[G_Y - G_Y G_U \left(G_U + N \varepsilon_N I_N \right)^{-1} \right] \end{aligned}$$

where

$$G_U = \left(I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) K_U \left(I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) : \text{centered Gram matrix}$$

$$K_U = k(B^T X_i, B^T X_j)$$

- Optimization problem:

$$\min_{B: B^T B = I_d} \text{Tr} \left[G_Y \left(G_{B^T X} + N \varepsilon_N I_N \right)^{-1} \right]$$

Experiments with KDR

■ Wine data

Data

13 dim. 178 data

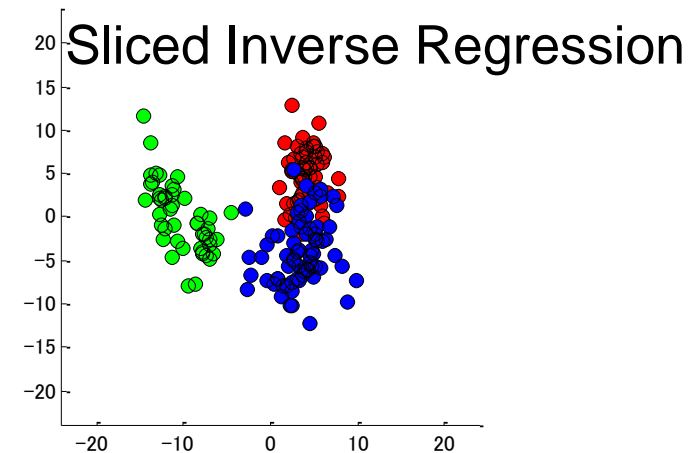
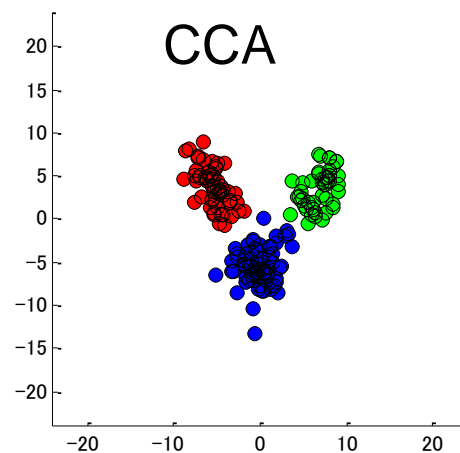
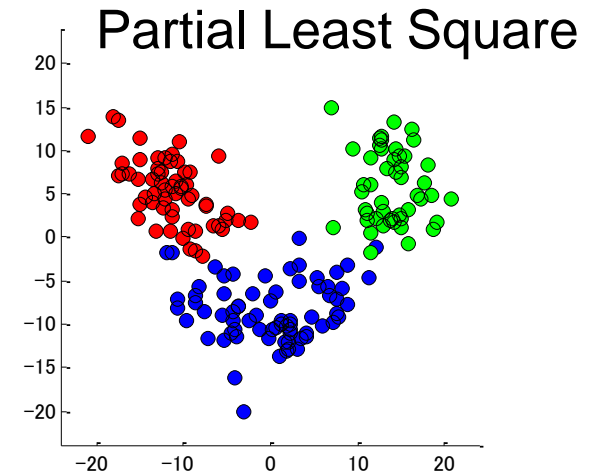
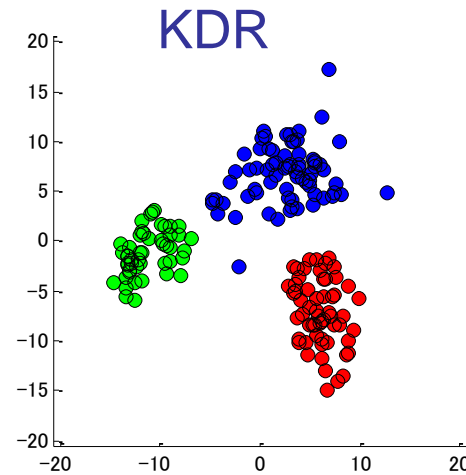
3 classes

2 dim. projection

$$k(z_1, z_2)$$

$$= \exp\left(-\frac{\|z_1 - z_2\|^2}{\sigma^2}\right)$$

$$\sigma = 30$$



Consistency of KDR

Theorem

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \rightarrow 0, \quad N^{1/2} \varepsilon_N \rightarrow \infty \quad (N \rightarrow \infty).$$

Let S_0 be the set of optimal parameters:

$$S_0 = \left\{ B \mid B^T B = I_d, \operatorname{Tr} \left[\Sigma_{YY|X}^B \right] = \min_{B'} \operatorname{Tr} \left[\Sigma_{YY|X}^{B'} \right] \right\}$$

Then, under some conditions, for any open set $U \supset S_0$

$$\Pr \left(\hat{B}^{(N)} \in U \right) \rightarrow 1 \quad (N \rightarrow \infty).$$

Lemma

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \rightarrow 0, \quad N^{1/2} \varepsilon_N \rightarrow \infty \quad (N \rightarrow \infty).$$

Then, under some conditions,

$$\sup_{B: B^T B = I_d} \left| \text{Tr} \left[\ddot{\Sigma}_{YY|X}^{B(N)} \right] - \text{Tr} \left[\Sigma_{YY|X}^B \right] \right| \rightarrow 0 \quad (N \rightarrow \infty)$$

in probability.

Conclusions

- Are you a Bayesian or a frequentist?
- My own answer is “both,” but there are days where I'm much more clearly one than the other
 - and it is an ongoing intellectual challenge to try to understand the ramifications of this distinction
- I view them as complementary perspectives, but there is a wave/particle uncomfortableness at times
- A main conclusion: machine learning is a part of statistics; don't just read the machine learning literature---read, ponder and contribute to the broad statistical literature