

A Kernel Between Sets of Vectors

Risi Kondor

Tony Jebara

Columbia University, New York, USA.

A Kernel between Sets of Vectors

In SVM, Gaussian Processes, Kernel PCA, kernel K defines feature map $\Phi : \mathcal{X} \mapsto \mathcal{H}$ such that

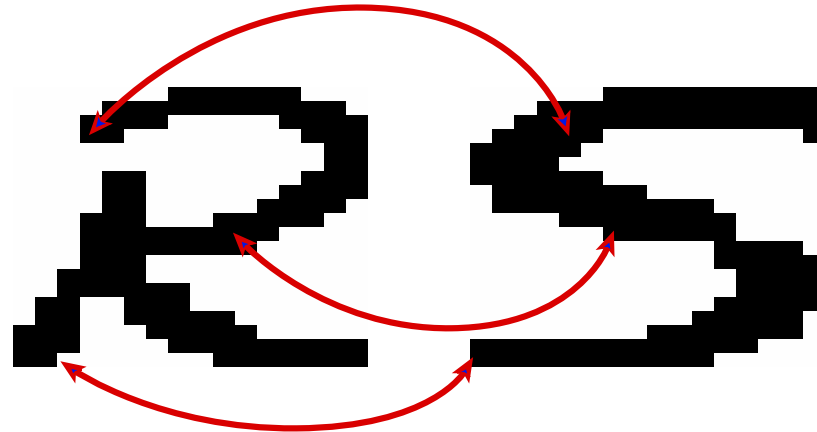
$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

- algorithm becomes linear in \mathcal{H}
- captures prior knowledge about domain
- crucial role in performance
- “kernel engineering”

A new kernel between composite objects

Conventional Kernels between Images

Representing $N \times N$ images as vectors in \mathbb{R}^{N^2} .

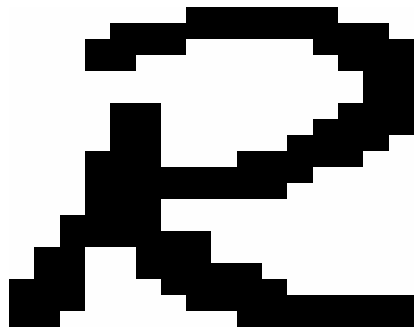


e.g.
$$K(x, x') = e^{-\|x - x'\|^2 / (2\sigma^2)} = \exp \left[- \sum \frac{(x_i - x'_i)^2}{2\sigma^2} \right]$$

Only sensitive to similarity between matching pixels, no sense of distance within image, sensitive to translations, rotations, etc..

The “Bag of Tuples” Representation

e.g. for images: set of (x, y) pairs for each foreground pixel, or set of $(x, y, \text{intensity})$ triplets

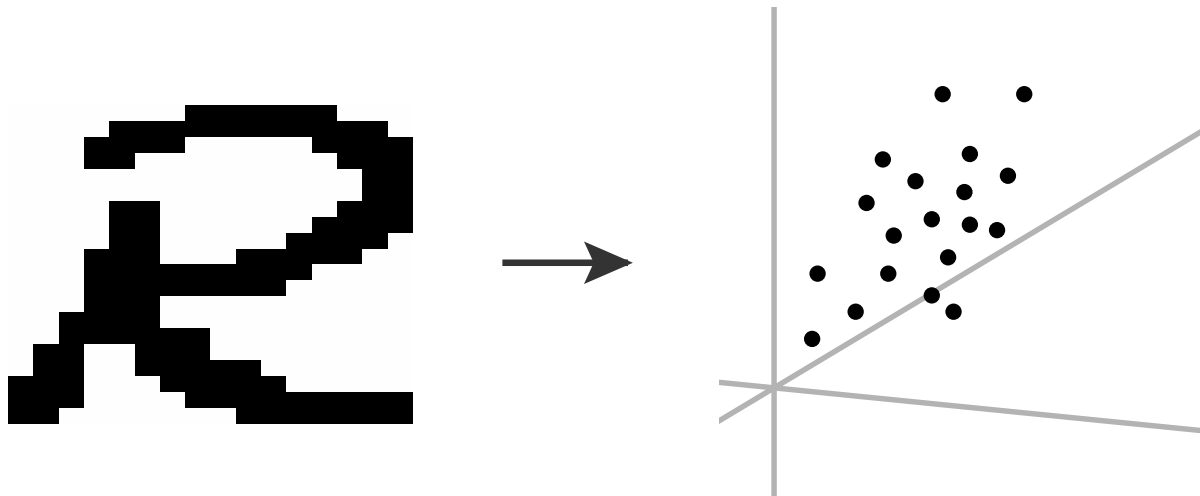


$$\mathcal{X} = \{(3, 8), (2, 15), (6, 9), (14, 8), \dots\}$$

Similar natural “bag of vectors” representations exist for sequences, time series, etc..

Cloud of Points in Feature Space

Take a “base kernel” κ between tuples, and consider the feature map $\Phi : \mathbb{R}^N \mapsto \mathcal{H}$ satisfying $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$.

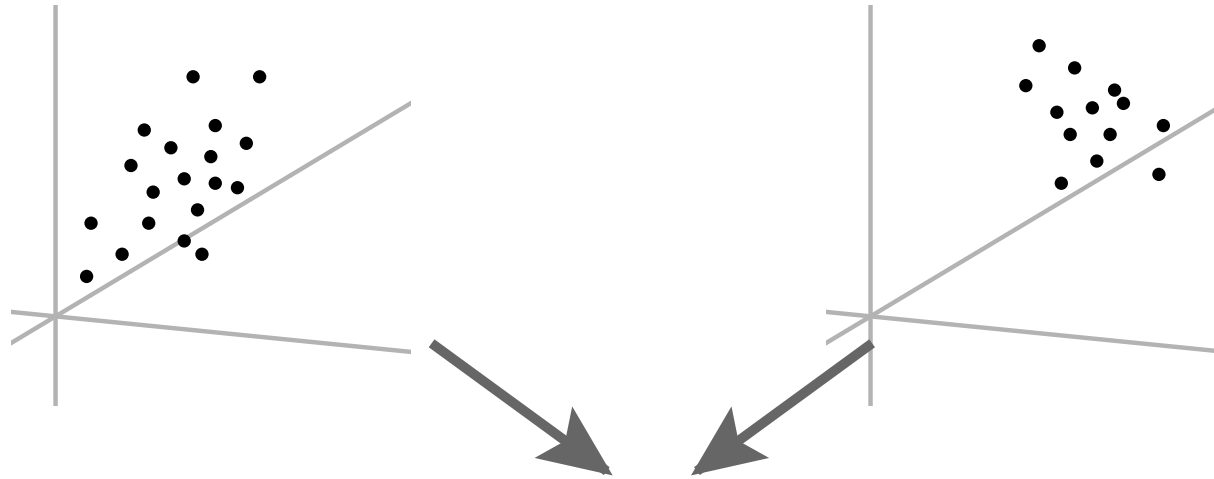


e.g. $\kappa = e^{-\|\mathbf{x} - \mathbf{x}'\|^2 / (2\sigma^2)}$

Producing a Kernel Between Examples

R

S



$$K(x, x') = ?$$

Fit distributions p and p' to x and x' and define

$$K(x, x') = K(p, p')$$

The Bhattacharyya Kernel

between distributions p and p' :

$$K(p, p') = \int \sqrt{p(x)} \sqrt{p'(x)} dx$$

related by $H = \sqrt{2 - 2K}$ to Hellinger distance

$$H(p, p') = \left[\int \left(\sqrt{p(x)} - \sqrt{p'(x)} \right)^2 dx \right]^{1/2}.$$

Positive definite and symmetric (Mercer) by construction. Also $K(x, x') = 1$. Invariant to permutations of vectors.

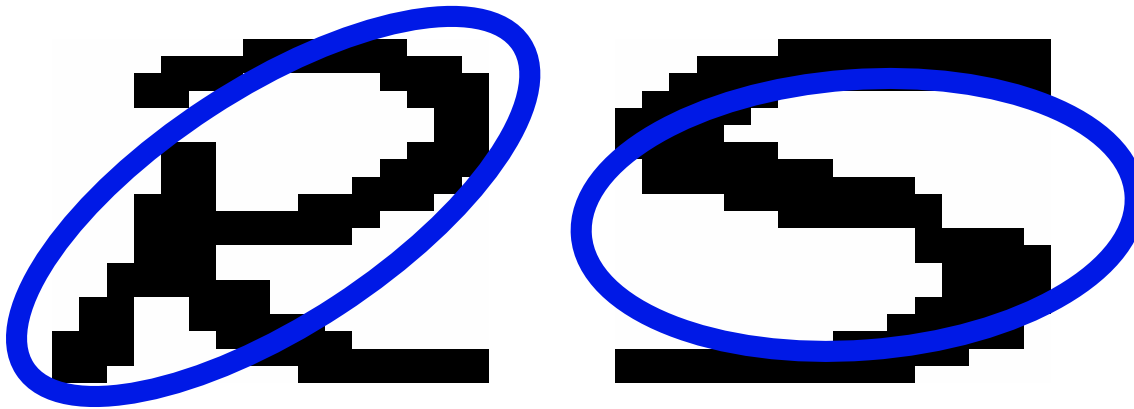
The Bhattacharyya Kernel between Normal Distributions

$$p = \mathcal{N}(\mu, \Sigma) \quad p' = \mathcal{N}(\mu', \Sigma')$$

$$K(p, p') = \int \sqrt{p(x)} \sqrt{p'(x)} dx = \left[\frac{|\Sigma^\dagger|}{|\Sigma|^{1/2} |\Sigma'|^{-1/2}} \right]^{1/2} \exp \left(-\frac{1}{4} \mu^\top \Sigma^{-1} \mu - \frac{1}{4} \mu'^\top \Sigma'^{-1} \mu' + \frac{1}{2} \mu^\dagger \Sigma^\dagger \mu^\dagger \right)$$

where $\Sigma^\dagger = \left(\frac{1}{2} \Sigma^{-1} + \frac{1}{2} \Sigma'^{-1} \right)^{-1}$ and $\mu^\dagger = \frac{1}{2} \Sigma^{-1} \mu + \frac{1}{2} \Sigma'^{-1} \mu'$.

Fitting Normal Distributions in the Original Image Space



Limited representational power

Fitting Normal Distributions in Feature Space

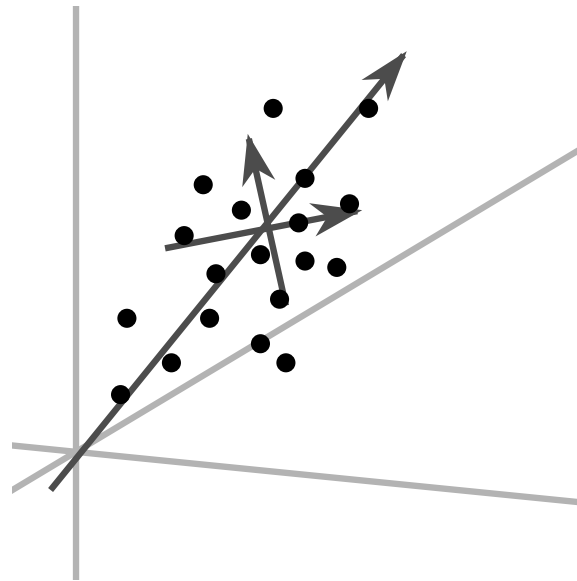
Regularized estimators:

$$\hat{\boldsymbol{\mu}} = \frac{1}{k} \sum_{i=1}^k \boldsymbol{\Phi}(x_i)$$

$$\hat{\boldsymbol{\Sigma}}_{\text{reg}} = \sum_{l=1}^r \mathbf{v}_l \lambda_l \mathbf{v}_l^\top + \eta \sum_i \boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^\top$$

where $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \dots$ is a basis and $\mathbf{v}_1, \dots, \mathbf{v}_r$ are first r eigenvectors of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{k} \sum_{i=1}^k (\boldsymbol{\Phi}(x_i) - \hat{\boldsymbol{\mu}})(\boldsymbol{\Phi}(x_i) - \hat{\boldsymbol{\mu}})^\top.$$



Dirac bra-ket notation

$$\langle x| = \Phi(x)^\top \quad (\text{bra})$$

$$|x\rangle = \Phi(x) \quad (\text{ket})$$

Inner product:

$$\langle x|x'\rangle = \langle \Phi(x), \Phi(x') \rangle = \kappa(x, x')$$

Bilinear forms:

$$\sum_i |x_i\rangle a_i \langle x_i|$$

Finding $\mathbf{v}_1, \dots, \mathbf{v}_r$ with Kernel PCA

Assume $|\mu\rangle = 0$. Want to solve:

$$\hat{\Sigma} |v_l\rangle = \lambda |v_l\rangle \quad \hat{\Sigma} = \frac{1}{k} \sum_{j=1}^k |x_j\rangle \langle x_j| \quad [\infty \times \infty \text{ matrix}].$$

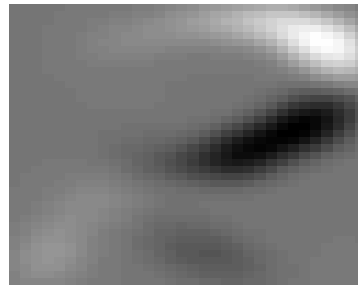
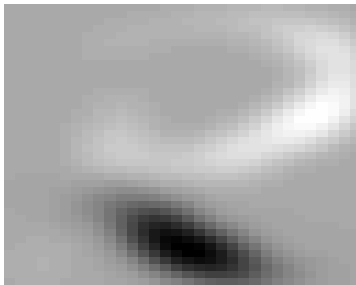
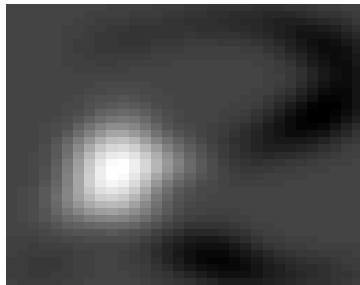
Observation: $|v\rangle = \sum_{i=1}^k \alpha_i |x_i\rangle$. Multiplying by $\langle x_l|$:

$$\frac{1}{k} \sum_{j=1}^k \sum_{i=1}^k \langle x_l | x_j \rangle \langle x_j | x_i \rangle \alpha_i = \lambda \sum_{i=1}^k \langle x_l | x_i \rangle \alpha_i.$$

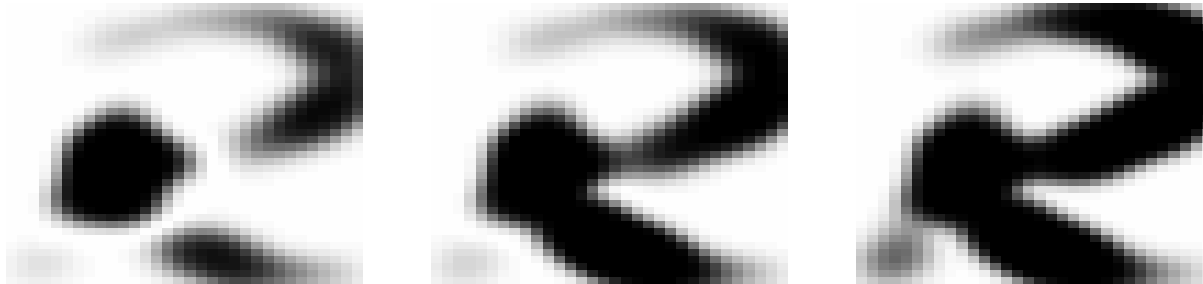
Reduces to

$$K\alpha = k\lambda\alpha \quad \text{with} \quad K_{i,j} = \langle x_i | x_j \rangle \quad [k \times k \text{ matrix}].$$

The First Three Principal Components of “R”

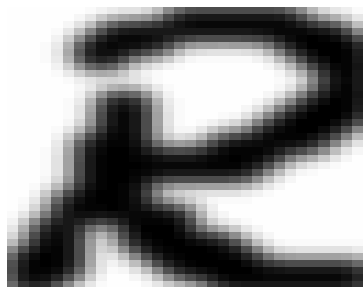


Reconstruction from the first 1,2 and 3 components

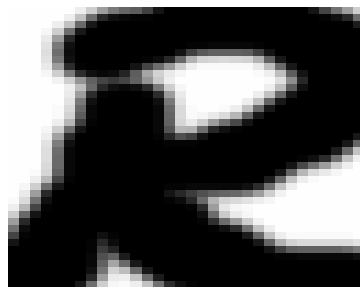


$$\text{intensity}(x) \propto e^{-\langle x|\Sigma|x\rangle}$$

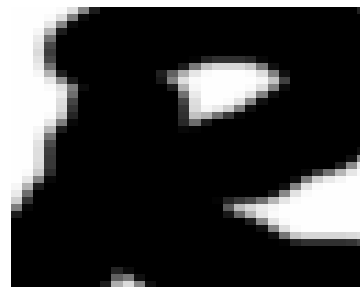
Reconstruction from the first 3 components with regularization



$\eta = 0.01$



$\eta = 0.1$

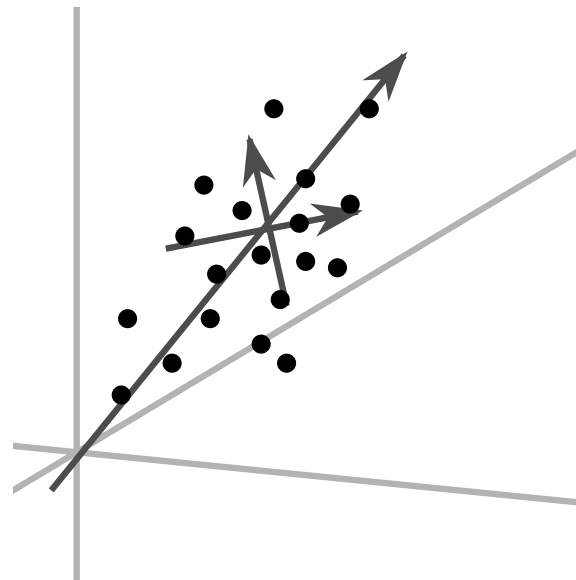


$\eta = 1$

$$\text{intensity}(x) \propto e^{-\langle x | \Sigma_{\text{reg}}^{-1} | x \rangle}$$

Properties of Bhattacharyya Kernels with Regularization

- Smoothing $\leftrightarrow \eta$
- Graceful behavior under natural transformations such as translations/rotations; just rotates cloud in \mathcal{H}



Relationship to Gaussian Processes

What are $|\xi\rangle \in \mathcal{H}$? By $f(x) = \langle x|\xi\rangle$ they are really images themselves (RKHS view).

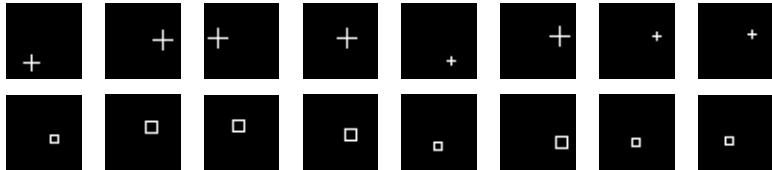
$\mathcal{N}(\hat{\mu}, \hat{\Sigma})$ defines a distribution over such functions (images), in fact, a Gaussian Process with

$$\mathbf{E}[f(x)] = \frac{1}{k} \sum_i \kappa(x_i, x)$$

$$\mathbf{Cov}(f(x), f(x')) = \langle \kappa(x, \cdot) | \hat{\Sigma} | \kappa(x', \cdot) \rangle .$$

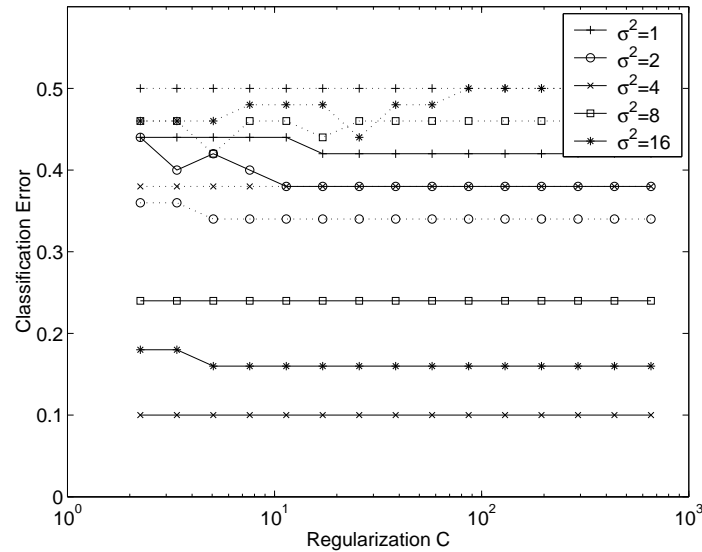
Not the usual Bayesian Gaussian Process training procedure!

Experiment: crosses and squares



SVM trained on 100 examples of
 40×40 images

Gaussian κ with $r = 4$ and
 $\eta = 0.01$

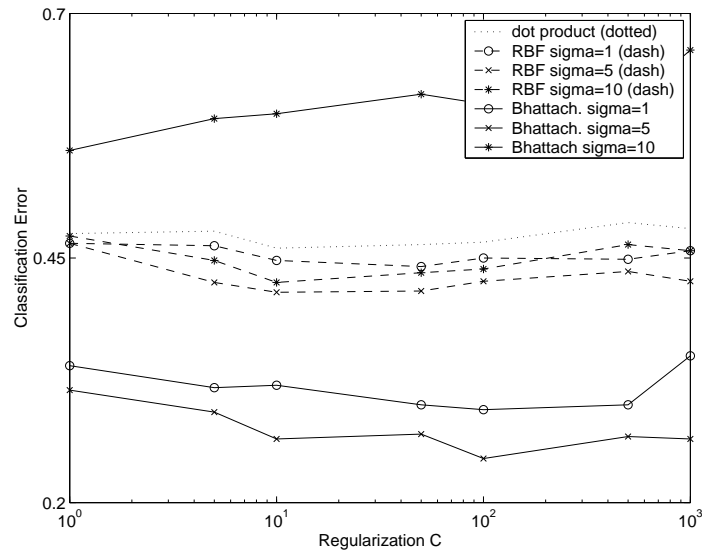


solid: Bhattacharyya kernel **dotted:** conventional RBF

Experiment: NIST digit recognition

Artificially hard problem:

SVM on 10 classes (one vs. all); only 30 pixels sampled from each image and 12 images per class in training set; $r = 10$ and $\eta = 0.1$.



solid: Bhattacharyya kernel **dashed:** conventional RBF

Summary

- Bag of vectors representation.
- Kernel trick employed on two levels (κ and K).
- Semiparametric; Bhattacharyya kernel

$$K(p, p') = \int \sqrt{p(x)} \sqrt{p'(x)} dx$$

computable in closed form for Normal distribution, even in \mathcal{H} .

- Graceful behavior under natural transformations
- Possibly applicable to many other data types, not just images: sequences, time series, 3D objects, proteins, . . .

References

- L. Wolf and A. Shashua: **Kernel Principal Angles for Classification Machines with Applications to image Sequence Interpretation** CVPR 2003 [very similar ideas developed independently].
- T. Gärtner, P. A. Flach, A. Kowalczyk and A. J. Smola **Multi-Instance Kernels** ICML 2002.
- T. Jebara and R. Kondor: **Bhattacharyya and Expected Likelihood Kernels** COLT/KW 2003.
- A. Bhattacharyya: **On a Measure of Divergence between two Statistical Populations Defined by their Probability Distributions** Bull. Calcutta Math. Soc. 35 (1943).