Subspace Learning

Alessandro Rudi, Guille D. Canas, Lorenzo Rosasco

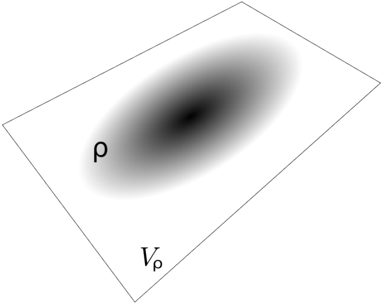
Universitá di Genova, Italy
Massachusetts Institute of Technology
Istituto Italiano di Tecnologia

ROKS 2013
Louvain, 09th of July 2013

Subspace Learning

Subspace Learning

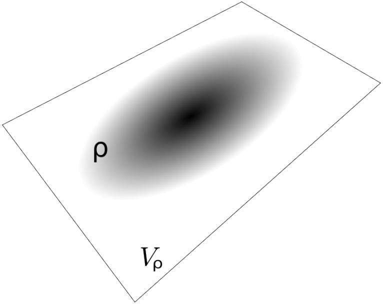
- $\mathcal{H}$: Hilbert Space
- ρ: probability distribution on $\mathcal{H}$
- $\operatorname{supp}\rho$: is the support of ρ
- $V_\rho = \overline{\operatorname{span}\{x \mid x \in \operatorname{supp}\rho\}}$
  "smallest" linear subspace containing $\operatorname{supp}\rho$



Problem How to "find" V_ρ given the examples $x_1, \ldots, x_n \sim \rho$?

Subspace Learning

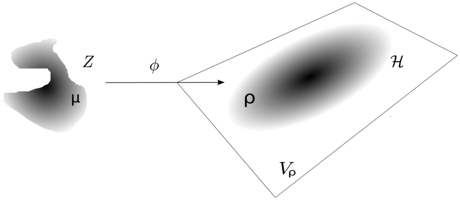
- $\mathcal{H}$: Hilbert Space
- ρ: probability distribution on $\mathcal{H}$
- $\operatorname{supp}\rho$: is the support of ρ
- $V_\rho = \overline{\operatorname{span}\{x \mid x \in \operatorname{supp}\rho\}}$
  "smallest" linear subspace
  containing $\operatorname{supp}\rho$



Problem How to "find" V_ρ given the examples $x_1, \ldots, x_n \sim \rho$?

Setting: Why a Hilbert Space $\mathcal{H}$

- limit for high dimensional data
- embedded data $(Z, \mu) \xrightarrow{\phi} \mathcal{H}$



Example 1: PCA - Kernel PCA

PCA

V_ρ the smallest linear subspace of $\mathcal{H}$ that contains all the distribution

$$V_\rho = \operatorname*{argmin}_V dim(V) \quad \text{such that} \;\; var(V) = var(\mathcal{H})$$

Kernel PCA [Schölkopf 1997]

performs PCA on the data embedded in $\mathcal{H}$ by a feature map ϕ

Example 1: PCA - Kernel PCA

PCA

V_ρ the smallest linear subspace of $\mathcal{H}$ that contains all the distribution

$$V_\rho = \underset{V}{\operatorname{argmin}}\, dim(V) \quad \text{such that} \;\; var(V) = var(\mathcal{H})$$

Kernel PCA [Schölkopf 1997]

performs PCA on the data embedded in $\mathcal{H}$ by a feature map ϕ

Example 2: Kernel Support Estimation

- $(Z, \mu),\ M = \operatorname{supp}\mu$
- $\phi : Z \to \mathcal{H},\ V_\rho = \overline{\operatorname{span}\{\phi(z) \mid Z \in M\}}$

If ϕ is *separating* [De Vito 2010]

$$M = \{z \in Z \mid \phi(z) \in V_\rho\}$$

Examples separating ϕs on $\mathbb{R}^d$

- Abel kernel, $\langle \phi(z), \phi(z') \rangle = \exp(-\gamma \|z - z'\|_{\ell_2})$
- the convex combination or the product of two separating kernels
- Gaussian kernel is NOT separating

Example 2: Kernel Support Estimation

- $(Z, \mu),\ M = \operatorname{supp} \mu$
- $\phi : Z \to \mathcal{H},\ V_\rho = \overline{\operatorname{span} \{\phi(z) \mid Z \in M\}}$

If ϕ is *separating* [De Vito 2010]

$$M = \{z \in Z \mid \phi(z) \in V_\rho\}$$

Examples separating ϕs on $\mathbb{R}^d$

- Abel kernel, $\langle \phi(z), \phi(z') \rangle = \exp(-\gamma \|z - z'\|_{\ell_2})$
- the convex combination or the product of two separating kernels
- Gaussian kernel is NOT separating

Example 2: Kernel Support Estimation

- $(Z, \mu),\ M = \operatorname{supp}\mu$
- $\phi : Z \to \mathcal{H},\ V_\rho = \overline{\operatorname{span}\{\phi(z) \mid Z \in M\}}$

If ϕ is *separating* [De Vito 2010]

$$M = \{z \in Z \mid \phi(z) \in V_\rho\}$$

Examples separating ϕs on $\mathbb{R}^d$

- Abel kernel, $\langle \phi(z), \phi(z') \rangle = \exp(-\gamma \|z - z'\|_{\ell_2})$
- the convex combination or the product of two separating kernels
- Gaussian kernel is NOT separating

Problem definition

Given $x_1, \ldots, x_n$ drawn independently from ρ, find $\hat{V}$ such that

$$P\left(d(\hat{V}, V_\rho) > \epsilon\right) \leq \delta(\epsilon, n)$$

How to build $\hat{V}$?
Which distance d on linear subspaces?

Problem definition

Given $x_1, \ldots, x_n$ drawn independently from ρ, find $\hat{V}$ such that

$$P\left(d(\hat{V}, V_\rho) > \epsilon\right) \leq \delta(\epsilon, n)$$

How to build $\hat{V}$?
Which distance d on linear subspaces?

Covariance Lemma in the continuous case

$$V_\rho = \overline{\operatorname{span}\{u_i \mid i \geq 1\}}$$

where $Cu_i = \sigma_i u_i$ with $C : \mathcal{H} \to \mathcal{H}$ the covariance operator

$$C = \mathbb{E}_{x \sim \rho}[x \otimes x] - \mu \otimes \mu$$

Truncated estimator

Analogously we can define

$$\hat{V}^k = \operatorname{span}\{\hat{u}_i \mid 1 \leq i \leq k\}$$

where $\hat{C} u_i = \hat{\sigma}_i \hat{u}_i$ with $\hat{C} : \mathcal{H} \to \mathcal{H}$ the empirical covariance operator

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i \otimes x_i - \hat{\mu} \otimes \hat{\mu}$$

What is a good value of k?
Shall we simply take $k = n$?

Truncated estimator

Analogously we can define

$$\hat{V}^k = \operatorname{span}\{\hat{u}_i \mid 1 \leq i \leq k\}$$

where $\hat{C} u_i = \hat{\sigma}_i \hat{u}_i$ with $\hat{C} : \mathcal{H} \to \mathcal{H}$ the empirical covariance operator

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i \otimes x_i - \hat{\mu} \otimes \hat{\mu}$$

What is a good value of k?
Shall we simply take $k = n$?

Which metric?

Let C be the covariance operator associated to the distribution ρ.

$$d_{\alpha,p,\rho}(U,V) = \|(P_U - P_V)C^{\alpha}\|_p$$

- C is the covariance operator of ρ
- P_U is the projection operator associated to the subspace U
- $\|\|_p$ is the p-Schatten norm, $\|A\|_p^p = \sum_{i \geq 1} \sigma_i^p$

It generalizes many commonly used subspace distances

Which metric?

Let C be the covariance operator associated to the distribution ρ.

$$d_{\alpha,p,\rho}(U, V) = \|(P_U - P_V)C^{\alpha}\|_p$$

- C is the covariance operator of ρ
- P_U is the projection operator associated to the subspace U
- $\|\|_p$ is the p-Schatten norm, $\|A\|_p^p = \sum_{i \geq 1} \sigma_i^p$

It generalizes many commonly used subspace distances

Metric for Kernel PCA

Reconstruction error:

$$R(V) = \mathbb{E}_{x\sim\rho}\left[\|x - P_V x\|^2_{\mathcal{H}}\right]$$

- Commonly used in literature [Shawe-Taylor 2005, Blanchard 2007]
- $R(V) = d^2_{\frac{1}{2},2,\rho}(V, V_\rho)$

note that $R(W) \leq R(V)$ when $V \subseteq W$

Metric for Support Estimation

When the feature map is separating, the support M is defined as

$$M = \{z \in Z \mid F_{V_\rho}(z) = 0\} \text{ with } F_{V_\rho}(z) = \operatorname{dist}_{V_\rho}(\phi(z))$$

The natural estimator studied in [De Vito 2010, De Vito 2012] is defined as

$$\hat{M} = \{z \in Z \mid F_{\hat{V}^k}(z) \leq \tau\} \text{ with } F_{\hat{V}^k}(z) = \operatorname{dist}_{\hat{V}^k}(\phi(z))$$

In order to study the convergence of the set $\hat{M}$ to M is of interest to bound the quantity

$$\sup_{z \in Z} |F_{V_\rho}(z) - F_{\hat{V}^k}(z)| \leq \left\|(P_{\hat{V}^k} - P_{V_\rho})C^\alpha\right\|_\infty = d_{\alpha,\infty,\rho}(\hat{V}^k, V_\rho)$$

where α depends on the eigenvalue decay of C.

Metric for Support Estimation

When the feature map is separating, the support M is defined as

$$M = \{z \in Z \mid F_{V_\rho}(z) = 0\} \text{ with } F_{V_\rho}(z) = \operatorname{dist}_{V_\rho}(\phi(z))$$

The natural estimator studied in [De Vito 2010, De Vito 2012] is defined as

$$\hat{M} = \{z \in Z \mid F_{\hat{V}^k}(z) \leq \tau\} \text{ with } F_{\hat{V}^k}(z) = \operatorname{dist}_{\hat{V}^k}(\phi(z))$$

In order to study the convergence of the set $\hat{M}$ to M is of interest to bound the quantity

$$\sup_{z \in Z} |F_{V_\rho}(z) - F_{\hat{V}^k}(z)| \leq \left\|(P_{\hat{V}^k} - P_{V_\rho})C^\alpha\right\|_\infty = d_{\alpha,\infty,\rho}(\hat{V}^k, V_\rho)$$

where α depends on the eigenvalue decay of C.

Metric for Support Estimation

When the feature map is separating, the support M is defined as

$$M = \{z \in Z \mid F_{V_\rho}(z) = 0\} \text{ with } F_{V_\rho}(z) = \operatorname{dist}_{V_\rho}(\phi(z))$$

The natural estimator studied in [De Vito 2010, De Vito 2012] is defined as

$$\hat{M} = \{z \in Z \mid F_{\hat{V}^k}(z) \leq \tau\} \text{ with } F_{\hat{V}^k}(z) = \operatorname{dist}_{\hat{V}^k}(\phi(z))$$

In order to study the convergence of the set $\hat{M}$ to M is of interest to bound the quantity

$$\sup_{z \in Z} |F_{V_\rho}(z) - F_{\hat{V}^k}(z)| \leq \left\| (P_{\hat{V}^k} - P_{V_\rho}) C^\alpha \right\|_\infty = d_{\alpha,\infty,\rho}(\hat{V}^k, V_\rho)$$

where α depends on the eigenvalue decay of C.

More on General metric

- $d_{\alpha,p,\rho}$ is a metric for $\Lambda(V_\rho)$, the collection of subspaces of V_ρ, where $0 \leq \alpha \leq 1$ and $1 \leq p \leq \infty$
- each $\hat{V}^k$ is a subspace of V_ρ thus $\hat{V}^k \in \Lambda(V_\rho)$
- $d_{\alpha,p,\rho}(V, W) \leq d_{\alpha,p,\rho}(U, W) \qquad U \subseteq V \subseteq W$

the metric $d_{a,p,\rho}$ allows to control a variety of metrics classically used to measure distance between sets [Beer 1993]

More on General metric

- $d_{\alpha,p,\rho}$ is a metric for $\Lambda(V_\rho)$, the collection of subspaces of V_ρ, where $0 \le \alpha \le 1$ and $1 \le p \le \infty$
- each $\hat{V}^k$ is a subspace of V_ρ thus $\hat{V}^k \in \Lambda(V_\rho)$
- $d_{\alpha,p,\rho}(V, W) \le d_{\alpha,p,\rho}(U, W) \qquad U \subseteq V \subseteq W$

the metric $d_{a,p,\rho}$ allows to control a variety of metrics classically used to measure distance between sets [Beer 1993]

Subspace Learning

Learning rate for the general metric

Theorem 1 (Rudi, Canas, Rosasco 2013)

With probability $1-\delta$

$$d_{\alpha,p,\rho}(\hat{V}^k, V_\rho) \leq 4t^\alpha \mathcal{N}_{\alpha p}(t)^\alpha$$

- $t = \max\{\sigma_k, \frac{9}{n}\log\frac{n}{\delta}\}$
- σ_k the k-th eigenvalue of C
- $\mathcal{N}_{\alpha p}(t) = \left\|C(C+tI)^{-1}\right\|_{\alpha p}$ a generalization of the effective dimension [Caponnetto 2005] (that is $\mathcal{N}(t) = \mathcal{N}_2(t)$)

tools from: spectral theory, Löwner partial orderings, concentrations bounds on operators [Tropp 2012]

Learning rate for the general metric

Assumption on the eigenvalue decay of C

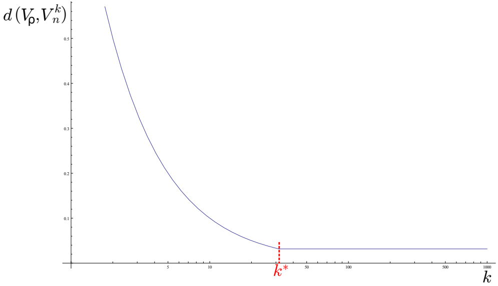
if we assume that $\sigma_m(C) \sim m^{-r}$ with $r > 1$ we have

$$d_{\alpha,p,\rho}(\hat{V}^k, V_\rho) \leq \begin{cases} Qk^{-r\alpha+\frac{1}{p}} & \text{if } k < k^* \qquad \text{(polynomial decay)} \\ Qk^{*-r\alpha+\frac{1}{p}} & \text{if } k \geq k^* \qquad \text{(plateau)} \end{cases}$$

with probability $1 - \delta$ and q, Q constants

$$k^* = \left(\frac{qn}{9 \log(n/\delta)} \right)^{\frac{1}{r}}$$

Learning Rates for Kernel PCA and Reconstruction error



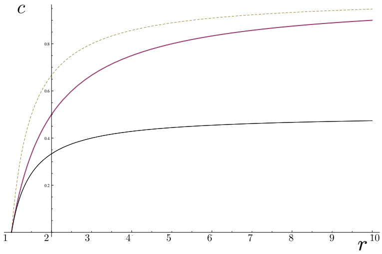
$$k^* = \left(\frac{n}{\log n}\right)^{\frac{1}{r}}$$

$$R(\hat{V}^k) = d_{\frac{1}{2},2,\rho}(\hat{V}^k, V_\rho)^2 \leq Q \begin{cases} k^{-\frac{r-1}{r}} & k < k^* \\ \left(\frac{\log n}{n}\right)^{\frac{r-1}{r}} & k \geq k^* \end{cases}$$

$$\text{where } \sigma_m(C) \sim m^{-r}, \ r > 1$$

Rates comparison on Kernel PCA

- [Blanchard 2007] (dotted line). Analysis for fixed k and reconstruction error. It makes assumptions on the fourth order. Learning rate $O(n^{-c})$ with $c = \frac{s(r-1)}{r-s+rs}$ where s is the fourth-moment eigenvalue decay.
- [Shawe-Taylor 2005] (black line) Analysis for fixed k and reconstruction error. Learning rate $O(n^{-c})$ with $c = \frac{r}{2(r-1)}$.
- *Our result for reconstruction error* (purple thick line). Learning rate $O(n^{-c})$ with $c = \frac{r}{r-1}$ where s is the fourth-moment eigenvalue decay.



Learning Rates for Kernel Support Estimation

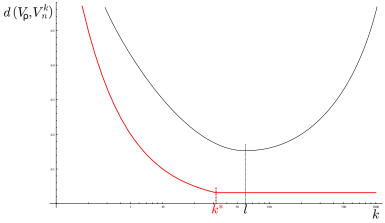
With probability $1 - \delta$

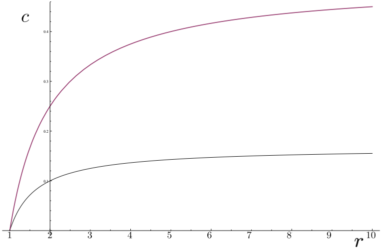
$$d_{\alpha,\infty,\rho}(\hat{V}^k, V_\rho) \leq Q \begin{cases} k^{-r\alpha} & k < k^* \\ \left(\frac{\log n}{n}\right)^{r\alpha} & k \geq k^* \end{cases}$$

where $k^* = \left(\frac{n}{\log n}\right)^{\frac{1}{r}}$ and $\sigma_m(C) \sim m^{-r},\ r > 1$

Rates comparison on Kernel Support Estimation

- [De Vito 2010, De Vito 2012] (black line on the left) It does not respect the monotonicity of the distance w.r.t. nested sets. (black line on the right) Learning rate $O(n^{-c})$ with $c = \frac{r-1}{2(3r-1)}$ with the worst case $\alpha = \frac{r-1}{2r}$
- *Our result* (red thick line). (red line on the left). It respect the monotonicity of the distance. (black line on the right) Learning rate $O(n^{-c})$ with $c = \frac{r-1}{2r}$ with the worst case $\alpha = \frac{r-1}{2r}$

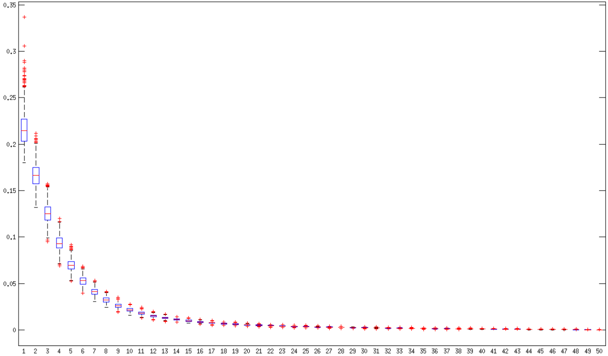




Subspace Learning

Experiments:Simulation on Kernel PCA(1)

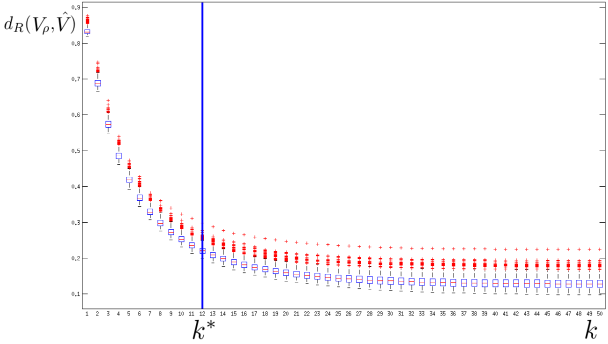
- μ uniform distribution on $[0, 1]$ with $Z = \mathbb{R}^2$
- $K(x, y) = \exp(-\gamma \|x - y\|_{\ell_1})$
- 1000 trials, each one of 1000 points independently drawn from μ



Eigenvalue decay of the associated empirical Covariance operator $\hat{C}$

Experiments:Simulation on Kernel PCA(2)

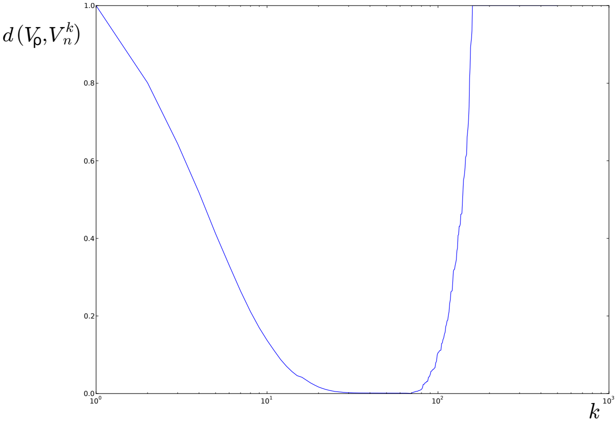
- the true covariance C can be computed analytically, it has polynomial decay $r = 2$.
- thus we can compute k^*
- the experiment shows the plateau behavior



Reconstruction error function of the number of the number of components k

Experiments: Numerical tradeoff in Kernel PCA (3)

- μ uniform distribution on $[0, 1]$ with $Z = \mathbb{R}^2$ with Gaussian kernel
- 1000 points independently drawn from μ
- computations performed on 32bits floating point precision



Reconstruction error with respect to the number of components k

Contribution

- Learning Rates for a wide range of metrics on linear subspaces
- Specific results for Kernel PCA and Spectral Support Estimation
- an optimal k^* for the truncated estimator

Future work

- Theoretical analysis on statistical/computational trade-off
- What happens with the noise?

G. Blanchard, O. Bousquet, and L. Zwald, *Statistical properties of kernel principal component analysis*, Machine Learning **66** (2007), no. 2.

G. Beer, *Topologies on closed and closed convex sets*, Springer, 1993.

A Caponnetto, L Rosasco, E De Vito, and A Verri, *Empirical effective dimensions and fast rates for regularized least-squares algorithm*, Tech. report, CBCL Paper 252/AI Memo 2005-019, MIT, 2005.

E. De Vito, L. Rosasco, and A. Toigo, *Spectral regularization for support estimation*, Neural Information Processing Systems, 2010.

——, *Learning sets with separating kernels*, arXiv preprint arXiv:1204.3573 (2012).

B. Schölkopf, A. Smola, and K.R. Müller, *Kernel principal component analysis*, Artificial Neural Networks—ICANN'97 (1997), 583–588.

J. Shawe-Taylor, C. K. Williams, N. Cristianini, and J. Kandola, *On the eigenspectrum of the gram matrix and the generalization error of kernel-pca*, IEEE Transactions on Information Theory **51** (2005), no. 7.

J.A. Tropp, *User-friendly tools for random matrices: An introduction*.