Subspace Learning

Alessandro Rudi, Guille D. Canas, Lorenzo Rosasco

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

ROKS 2013
Louvain, 09th of July 2013
1 Introduction
2 Main results
3 Numerics
4 Conclusions
Subspace Learning

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

Problem: How to “find” $V_\rho$ given the examples $x_1, \ldots, x_n \sim \rho$?
Subspace Learning

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

Problem

How to “find” $V_\rho$ 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

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

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

Kernel PCA [Schölkopf 1997]

performs PCA on the data embedded in $\mathcal{H}$ by a feature map $\phi$
Example 1: PCA - Kernel PCA

PCA

$V_\rho$ the smallest linear subspace of $\mathcal{H}$ that contains all the distribution $\mathcal{V}$ such that $\text{dim}(V) = \text{dim}(\mathcal{H})$ such that $\text{var}(V) = \text{var}(\mathcal{H})$

Kernel PCA [Schölkopf 1997]

performs PCA on the data embedded in $\mathcal{H}$ by a feature map $\phi$
Example 2: Kernel Support Estimation

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

If \(\phi\) is separating [De Vito 2010]

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

Examples separating \(\phi\)s 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

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

If $\phi$ is separating [De Vito 2010]

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

Examples separating $\phi$s 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

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

If \(\phi\) is \textit{separating} [De Vito 2010]

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

Examples separating \(\phi\)s 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 $\rho$, 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 $\rho$, 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

\[ V_{\rho} = \text{span} \{u_i \mid i \geq 1\} \]

where \( C u_i = \sigma_i u_i \) 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

\[ \hat{V}^k = \text{span} \{ \hat{u}_i \mid 1 \leq i \leq k \} \]

where \( \hat{C} u_i = \hat{\sigma}_i \hat{u}_i \) with \( \hat{C} : \mathcal{H} \rightarrow \mathcal{H} \) 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

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

where $\hat{C} u_i = \sigma_i \hat{u}_i$ 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 $\rho$.

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

- $C$ is the covariance operator of $\rho$
- $P_U$ is the projection operator associated to the subspace $U$
- $\|\cdot\|_p$ 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 $\rho$.

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

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

It generalizes many commonly used subspace distances
Metric for Kernel PCA

Reconstruction error:

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

- Commonly used in literature [Shawe-Taylor 2005, Blanchard 2007]
- \[ R(V) = d_{1/2,2,\rho}(V, V_\rho) \]

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

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

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

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

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

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

where $\alpha$ depends on the eigenvalue decay of $C$. 
Metric for Support Estimation

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

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

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

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

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

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

where $\alpha$ depends on the eigenvalue decay of $C$. 
Metric for Support Estimation

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

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

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

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

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 $\alpha$ 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 $V_\rho$, where $0 \leq \alpha \leq 1$ and $1 \leq p \leq \infty$
- each $\hat{V}^k$ is a subspace of $V_\rho$ thus $\hat{V}^k \in \Lambda(V_\rho)$
- $d_{\alpha,p,\rho}(V, W) \leq d_{\alpha,p,\rho}(U, W)$ \hspace{1cm} $U \subseteq V \subseteq W$

the metric $d_{\alpha,p,\rho}$ 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 $V_\rho$, where $0 \leq \alpha \leq 1$ and $1 \leq p \leq \infty$

- Each $\hat{V}^k$ is a subspace of $V_\rho$ thus $\hat{V}^k \in \Lambda(V_\rho)$

- $d_{\alpha,p,\rho}(V, W) \leq d_{\alpha,p,\rho}(U, W)$ where $U \subseteq V \subseteq W$

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

1. Introduction
2. Main results
3. Numerics
4. Conclusions
Learning rate for the general metric

**Theorem 1 (Rudi, Canas, Rosasco 2013)**

*With probability $1 - \delta$*

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

- $t = \max\{\sigma_k, \frac{9}{n} \log \frac{n}{\delta}\}$
- $\sigma_k$ the $k$-th eigenvalue of $C$
- $N_{\alpha,p}(t) = \|C(C + tI)^{-1}\|_{\alpha,p}$ a generalization of the effective dimension [Caponnetto 2005] (that is $N(t) = N_2(t)$)

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

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

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

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

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

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 $1 - \delta$

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

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 $\alpha = \frac{r-1}{2r}$
1. Introduction
2. Main results
3. Numerics
4. Conclusions
Experiments: Simulation on Kernel PCA (1)

- $\mu$ 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 $\mu$

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 components $k$
Experiments: Numerical tradeoff in Kernel PCA (3)

- $\mu$ uniform distribution on $[0, 1]$ with $Z = \mathbb{R}^2$ with Gaussian kernel
- 1000 points independently drawn from $\mu$
- 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?


