Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

Technical report January 9, 2017

Kernel PCA for novelty detection

Stefano Pozza[*]

Abstract

Novelty detection indexes are used in order to identify anomaly in the observation of a phenomenon. We describe the basic idea of kernel principal component analysis, a method which enlightens the existence of a novelty in a measured value comparing it with the one predicted by a model calibrated on training data. Differently from linear PCA, kernel PCA projects the data into an infinite-dimensional space in which novelty detection has usually a better performance.

Keywords: kernel principal component analysis, novelty detection, structural health monitoring.

Here we report the basic ideas behind the use of kernel principal component analysis for the detection of novelty behavior in structural health monitoring. We refer in particular to [6] where kernel PCA is introduced, [3] which proposes kernel PCA for the novelty detection, and [5] where it is used in structural health monitoring. For a discussion of the state of the art in novelty detection we refer to [4].

Assume to know n_y parameters which can be computed on n_s temporal subinterval. Then we can define the vector $\hat{\mathbf{y}}_i \in \mathbb{R}^{n_y}$ which contains the parameters at i-th time step, with $i = 1, \ldots, n_s$. However, any $\hat{\mathbf{y}}_i$ is affected by an error. Hence, we can define the actual values of the parameters as

$$\mathbf{y}_i = \hat{\mathbf{y}}_i + \mathbf{w}_i, \quad \text{for} \quad i = 1, \ldots, n_s,$$

[*] ISTI – CNR, Area della Ricerca CNR, via G. Moruzzi 1, 56124 Pisa, Italy (stefano.pozza@isti.cnr.it). This work was supported and funded by the project "MOSCARDO - ICT technologies for structural monitoring of age-old constructions based on wireless sensor networks and drones" (FAR FAS Regione Toscana, 2016-2018).

where $\mathbf{w}_i$ is the measurement error. Let

$$\mathbf{\Phi} : \mathbb{R}^{n_y} \to \mathcal{F} \subset \mathbb{C}^d$$

be a nonlinear map where $\mathcal{F}$ is a space with big (potentially infinite) dimension d. The main idea of kernel PCA is to use this kind of maps so that the images

$$\hat{\mathbf{z}}_i := \mathbf{\Phi}(\hat{\mathbf{y}}_i) \in \mathcal{F}, \qquad \text{for } i = 1, \dots, n_s,$$

live in a space where it is easier to identify anomalous behavior of the data. In this analysis we consider a set of *training data* $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_{n_t}$, with $n_t < n_s$, which we use to train our model. Then, we will estimate the error between further data and the ones predicted by the model in order to understand if an anomaly has occurred.

The model we use is based on the *kernel PCA* and we will present it from the point of view of the *singular values decomposition* (SVD). Let us consider the matrix

$$Z = \frac{1}{\sqrt{n_t}} [\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_{n_t}] \in \mathbb{C}^{d \times n_t}.$$

Naturally, for every $i = 1, \dots, n_t$ we get $\hat{\mathbf{z}}_i = Z\sqrt{n_t}\mathbb{1}_i$, where $\mathbb{1}_i$ is a vector equal to 1 at the i-th element and zero elsewhere. Moreover, we can estimate the data outside the training set looking for some $\mathbf{x}_i \in \mathbb{C}^{n_t}$ which minimize the euclidean norm of the error $\hat{\mathbf{e}}_i$ in the approximation

$$\hat{\mathbf{z}}_i = Z\mathbf{x}_i + \hat{\mathbf{e}}_i, \qquad \text{for} \qquad i = n_t + 1, \dots, n_s.$$

Of course, this is the *least square* solution of the system, which implies that $\hat{\mathbf{e}}_\mathbf{i}$ is in the null space of Z. If we consider the singular values decomposition of Z we get

$$Z = U\Sigma V^* = \sum_{j=1}^{n_t} \sigma_j \mathbf{u}_j \mathbf{v}_j^*, \tag{1}$$

with $U = [\mathbf{u}_1, \dots, \mathbf{u}_d] \in \mathbb{C}^{d \times d}$, $V = [\mathbf{v}_1, \dots, \mathbf{v}_{n_t}] \in \mathbb{C}^{n_t \times n_t}$ unitary matrices, and $\Sigma \in \mathbb{R}^{d \times n_t}$ a diagonal matrix with non-negative diagonal $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{n_t}$, the *singular values*. We can truncate the sum in (1) up to the first m elements, hence considering only the biggest singular values of the decomposition. Therefore, we get the approximation

$$Z\mathbf{x}_i = \tilde{\mathbf{z}}_i + \tilde{\mathbf{e}}_i = \sum_{j=1}^{m} \sigma_j \mathbf{u}_j \left(\mathbf{v}_j^* \mathbf{x}_i\right) + \tilde{\mathbf{e}}_i, \qquad \text{for } i = 1, \dots, n_s,$$

Notice that $\tilde{\mathbf{e}}_i = \sum_{j=m+1}^{m_0} \sigma_j \mathbf{u}_j \left(\mathbf{v}_j^* \mathbf{x}_i\right)$, where m_0 is the index of the last nonzero singular value (notice that we are assuming $m < m_0$ here and in the following).

Summarizing, the approximation $\tilde{\mathbf{z}}_i$ of $\mathbf{\Phi}(\hat{\mathbf{y}}_i)$ is obtained by the training set $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_{n_t}$ and by the principal component approximation given by the truncation up to the first m biggest singular values in (1). Notice that the columns of U and $\sigma_1^2, \dots, \sigma_{n_t}^2$ are respectively the eigenvectors and the eigenvalues of the *covariance matrix* $ZZ^* \in \mathbb{C}^{d \times d}$. Moreover, we underline that $\tilde{\mathbf{z}}_i \in \mathcal{U}_m = \operatorname{span}(\mathbf{u}_1, \dots, \mathbf{u}_m)$.

If an anomaly appears after the training interval at the i-th temporal interval, we expect $\hat{\mathbf{z}}_i$ to lie outside the space generated by $\mathcal{U}_m$. For this reason, as done in

[3], we can define the *novelty index* $\|\hat{\mathbf{r}}_i\|$ as the norm of the *reconstruction error* [2], which is the projection of the error $\hat{\mathbf{z}}_i - \tilde{\mathbf{z}}_i$ onto the subspace orthogonal to $\mathcal{U}_m$, i.e.,

$$\hat{\mathbf{r}}_i = (I - U_m U_m^*)(\hat{\mathbf{z}}_i - \tilde{\mathbf{z}}_i) = \hat{\mathbf{z}}_i - U_m U_m^* \hat{\mathbf{z}}_i, \tag{2}$$

where $U_m = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ and I is the identity.

However, explicit computations of U_m and $\hat{\mathbf{z}}_i$ are impossible to give since d is very big or infinite. For this reason we are going to reformulate the problem so that we only need to compute inner products between vectors $\hat{\mathbf{z}}_i \in \mathcal{F}$. By (1) we have

$$ZZ^* = U\Sigma^2 U^*,$$

from which we obtain

$$ZZ^*\mathbf{u}_i = \sigma_i^2 \mathbf{u}_i \quad \text{for } i = 1, \ldots, m_0. \tag{3}$$

Therefore, since $\sigma_i \neq 0$, $\mathbf{u}_i$ is a linear combination of the columns of ZZ^*. Hence, there exist vectors $\mathbf{a}_i \in \mathbb{C}^{n_t}$ for which

$$\mathbf{u}_i = Z\mathbf{a}_i \quad \text{for } i = 1, \ldots, m_0.$$

Using this last relation with (3) and pre-multiplying by Z^* give

$$(Z^*Z)^2\mathbf{a}_i = (Z^*Z)\mathbf{a}_i\sigma_i^2 \quad \text{for } i = 1, \ldots, m_0.$$

Setting $K = Z^*Z \in \mathbb{C}^{n_t,n_t}$, $\mathbf{a}_i$ can be computed as a solution of the eigenvector problem

$$K\mathbf{a}_i = \mathbf{a}_i\sigma_i^2 \quad \text{for } i = 1, \ldots, m_0. \tag{4}$$

Notice that since $\mathbf{u}_i^*\mathbf{u}_i = 1$ we must rescale the eigenvector so that $\mathbf{a}_i^* K\mathbf{a}_i = 1$, i.e., $\|\mathbf{a}_i\| = 1/\sigma_i$. In order to compute K we need to compute inner products of the form $\mathbf{\Phi}(\hat{\mathbf{y}}_i)^*\mathbf{\Phi}(\hat{\mathbf{y}}_j)$. Hence, we consider a kernel function representation of the inner product

$$k(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j) = \mathbf{\Phi}(\hat{\mathbf{y}}_i)^*\mathbf{\Phi}(\hat{\mathbf{y}}_j).$$

Then, instead of specifying $\mathbf{\Phi}$, we define it implicitly by its kernel. A Gaussian or a RBF kernel is usually considered since it implicitly defines an infinite dimension real map which depends on a single parameter. Hence, following what suggested in [3, 5], we choose to define for every $\mathbf{y}_i, \mathbf{y}_j \in \mathbb{R}^{n_y}$

$$k(\mathbf{y}_i, \mathbf{y}_j) = \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\sigma^2}\right), \tag{5}$$

with σ a real nonzero parameter to be determined. Now we can compute K and the matrix of its eigenvectors $A_m = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$. Thus, noticing that

$$\hat{\mathbf{z}}_i^*\hat{\mathbf{z}}_i = \mathbf{\Phi}(\hat{\mathbf{y}}_i)^*\mathbf{\Phi}(\hat{\mathbf{y}}_i) = k(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i),$$

we can express the novelty index (2) as

$$\|\hat{\mathbf{r}}_i\| = \sqrt{k(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i) - \|A_m^*\, Z^*\mathbf{\Phi}(\hat{\mathbf{y}}_i)\|^2}. \tag{6}$$

We remark that in the case of the Gaussian kernel (5) we get $k(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i) = 1$. Since (6) involves only inner products of the kind $\mathbf{\Phi}(\hat{\mathbf{y}}_i)^*\mathbf{\Phi}(\hat{\mathbf{y}}_j)$, we can obtain the novelty parameter using the kernel definition and the computed values A_m and Z.

In linear PCA usually the data sequence is centered, i.e., to each of the $\hat{\mathbf{y}}_j$ we subtract $\frac{1}{n_t}\sum_{i=1}^{n_t}\hat{\mathbf{y}}_i$, the mean of the sequence in the training set. As notice in [6] the same can be done for the kernel PCA replacing the matrix K by

$$HKH, \qquad \text{where } H := I - \frac{1}{n_t}\mathbb{1},$$

and $\mathbb{1}$ is a $n_t \times n_t$ matrix with all elements equal to 1.

We last need to determine the model parameters m and σ. We do it following the methods described in [5]. An optimal value for σ in (5) can be found by requiring the corresponding matrix $K(\sigma) = Z^*Z$ maximizing Shannon's informative entropy (notice that Z depends on σ). The values of the elements of $K(\sigma)$ lie on a certain interval $[k_{min}, k_{max}]$. Let us divide $[k_{min}, k_{max}]$ in ℓ equispaced subintervals. Then we can define the *Shannon's entropy* as

$$I_{\text{ent}}(K(\sigma)) := -\sum_{j=0}^{\ell-1} p_j \log_2(p_j),$$

with p_j the discrete probability that an element of K lie in the $(j+1)$st subinterval. In [5] the following procedure is suggested.

1. Subtract k_{min} from each element of $K(\sigma)$.
2. Rescale the values of the elements of $K(\sigma)$ so that they lie in the interval $[0, 255]$.
3. Compute h_j, the number of elements of the matrix lying in the subinterval $[j - 1/2, j + 1/2)$, with $j = 0, \ldots, 255$.
4. Compute $p_j = h_j/(n_t)^2$ and hence the Shannon's entropy $I_{\text{ent}}(K(\sigma))$.

The procedure is repeated for several values of σ so that we can find a value close to the one which maximize the entropy. Notice that if $\sigma \to 0$ then $K(\sigma)$ tends to the identical matrix, while if $\sigma \to +\infty$ then $K(\sigma)$ tends to be a matrix of all ones. Hence, both big and small values of σ make $I_{\text{ent}}(K(\sigma))$ small.

We set the value of m as the first index for which

$$\frac{\sum_{j=1}^{m}\sigma_j}{\sum_{j=1}^{n_t}\sigma_j} \geq F,$$

with $F \in [0, 1]$ the fraction of the sum of singular values we are interested in. In [5] the value $F = 0.99$ is suggested. Another classical method for determining m is to look at the semi logarithmic plot of the singular values in decreasing order and look for a drop in the distribution. If the drop exists then m is chosen as the index of the singular value before it (see, e.g., [1]).

We conclude the report with two remarks. First, notice that in the framework presented until now, we can obtain linear PCA setting $\mathbf{\Phi}$ as a identical map. Hence, the feature space $\mathcal{F} = \mathbb{R}^{n_y}$ and thus $m < n_y$. However, in this case we must pay attention in the use of the described novelty indexes, as we can see by the following example. Assume that $\hat{\mathbf{y}}_i$ is a measured parameter which is α times $\hat{\mathbf{y}}_j$, one of the vectors in the training data. We expect that, for α far from 1, the novelty index (2) should be big. However, in linear PCA

$$\hat{\mathbf{r}}_i = \hat{\mathbf{y}}_i - U_m U_m^* \hat{\mathbf{y}}_i = \alpha \left(\hat{\mathbf{y}}_j - U_m U_m^* \hat{\mathbf{y}}_j \right) = \alpha \hat{\mathbf{r}}_j.$$

Hence, when $|\alpha| < 1$ we get $\|\hat{\mathbf{r}}_i\| < \|\hat{\mathbf{r}}_j\|$. Therefore, the novelty index is considering a clearly anomalous measure as one which fits the model better than the training data.

Secondly, we remark that in [5] Reynders, Wursten and De Roeck suggested a different novelty index given by the norm of

$$\hat{\mathbf{s}}_i = U_{[m+1:m_0]} U_{[m+1:m_0]}^* \hat{\mathbf{z}}_i, \tag{7}$$

with $U_{[m+1:m_0]} = [\mathbf{u}_{m+1}, \ldots, \mathbf{u}_{m_0}]$ (notice that in [5] all the singular values are implicitly assumed to be positive). In this case the index represents the projection of the prediction error $\hat{\mathbf{z}}_i - \tilde{\mathbf{z}}_i$ onto the subspace given by the basis $\mathbf{u}_{m+1}, \ldots, \mathbf{u}_{m_0}$, which is the subspace generated by the eigenvectors corresponding to the truncated singular values $\sigma_{m+1}, \ldots, \sigma_{m_0}$. This subspace is orthogonal to $\mathcal{U}_m$ but is smaller than the orthogonal complement of $\mathcal{U}_m$ we have considered in the definition (2). In our numerical examples the two indexes do not differ substantially. However, the index defined in (7) may suffer from some numerical problems. Indeed, when some of the singular values σ_j for $j = m+1, \ldots, m_0$ are close to zero we may have roundoff problems in rescaling the vectors $\mathbf{a}_j$. This may lead to errors in the computation of $\|\hat{\mathbf{s}}_i\|$. A solution to this problem is to cut from $U_{[m+1:m_0]}$ the vectors corresponding to the too small singular values.

In our opinion, the use of (7) adds some complexity and risks in the computation of the novelty index. Moreover, (2) index is able to capture anomalies appearing also in the part of the subspace orthogonal to $\mathcal{U}_m$ which is not considered by (7). For these reasons in our experiments we preferred to use the index given by (2) even though not substantial differences arose in the numerical examples we made.

References

[1] A. Deraemaeker, E. Reynders, G. De Roeck, and J. Kullaa. Vibration-based structural health monitoring using output-only measurements under changing environment. *Mechanical Systems and Signal Processing*, 22(1):34 – 56, 2008.

[2] K. I. Diamantaras and S. Y. Kung. *Principal Component Neural Networks: Theory and Applications.* John Wiley & Sons, Inc., New York, NY, USA, 1996.

[3] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863 – 874, 2007.

[4] Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. Review: A review of novelty detection. *Signal Process.*, 99:215–249, June 2014.

[5] E. Reynders, G. Wursten, and G. de Roeck. Output-only structural health monitoring in changing environmental conditions by means of nonlinear system identification. *Structural Health Monitoring*, 13(1):82–93, 2014.

[6] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.