The Kernelized Taylor Diagram

Kristoffer Wickstrøm[1(✉)] , J. Emmanuel Johnson[2] , Sigurd Løkse[1] ,
Gustau Camps-Valls[2] , Karl Øyvind Mikalsen[1,4] ,
Michael Kampffmeyer[1,3] , and Robert Jenssen[1,3]

[1] UiT the Arctic University of Norway, Tromsø, Norway
kristoffer.k.wickstrom@uit.no
[2] Universitat de València, València, Spain
[3] Norwegian Computing Center, Oslo, Norway
[4] University Hospital of North Norway, Tromsø, Norway

Abstract. This paper presents the kernelized Taylor diagram, a graphical framework for visualizing similarities between data populations. The kernelized Taylor diagram builds on the widely used Taylor diagram, which is used to visualize similarities between populations. However, the Taylor diagram has several limitations such as not capturing non-linear relationships and sensitivity to outliers. To address such limitations, we propose the kernelized Taylor diagram. Our proposed kernelized Taylor diagram is capable of visualizing similarities between populations with minimal assumptions of the data distributions. The kernelized Taylor diagram relates the maximum mean discrepancy and the kernel mean embedding in a single diagram, a construction that, to the best of our knowledge, have not been devised prior to this work. We believe that the kernelized Taylor diagram can be a valuable tool in data visualization.

Keywords: Kernel methods · Taylor diagram · Data visualization

1 Introduction

Clear and informative visualization of similarities between populations is a key component both in the development of methodology and in scientific publications. Depending on the particular use case, a wide range of techniques are available. One such visualization technique is the Taylor diagram (TD) [10], which was devised to relate several statistical quantities and allow for comparison of numerous data points in a single diagram. The TD has been frequently used in numerous application, and particularly in climate sciences [6,8]. However, the statistical quantities displayed in the TD does have some weaknesses that limit the usability of the diagram. For instance, one quantity in the diagram is the Pearson correlation coefficient, which only models linear relationship and can be sensitive to outliers. This curtails the TD, as many real-world applications use data with outliers and that are connected through non-linear relationships.

One of the most well-known and widely used approaches for measuring similarity in machine learning is through kernel methods [3,4]. At its core, a kernel function corresponds to a dot product in a high-dimensional feature space, where non-linear relationship between data in the input space can be linearly related in the new feature space. As long as the kernel is positive definite, the mapping to the feature space does not have to be computed explicitly.

In this paper we propose the kernelized Taylor diagram (KTD), which is illustrated in Fig. 1. This diagram relates well-known quantities from the kernel literature [9], namely the maximum mean discrepancy (MMD) and the kernel mean embedding in a single

Fig. 1. KTD: The radial distance from the origin to each point is proportional to the length of kernel mean embedding. The distance between the points is the maximum mean discrepancy.

figure. To the best of our knowledge, such a diagram has never been devised prior to this work. The KTD makes no assumptions on the distributions of the populations and can model a rich family of relationships between populations. The functionality of the proposed diagram is demonstrated on synthetic data. Code: https://github.com/Wickstrom/KernelizedTaylorDiagram.
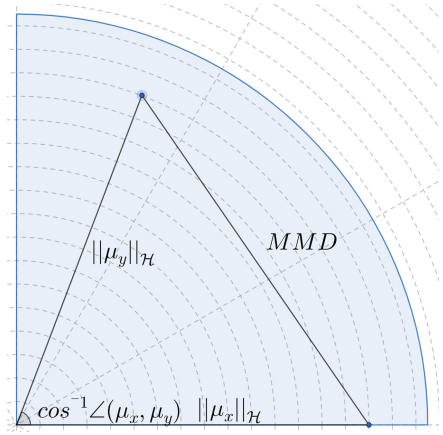
2   The Kernelized Taylor Diagram

Taylor Diagram. The TD was introduced as a tool that could relate several statistical quantities in a single figure [10]. It strength lies in the ability to compare numerous data points where it would otherwise be necessary to utilized several figures and/or tables. The theoretical starting point of the TD is the Pearson correlation coefficient ρ and the root-mean-squared-error E between two data points. [10] argued that neither are sufficient to capture potential similarities on their own, but in the aggregate the they are capable of detecting a wide range of differences between data points. Let $\mathbf{x}$ and $\mathbf{z}$ represent two D-dimensional vectors representing two data points. The correlation coefficient between $\mathbf{x}$ and $\mathbf{z}$ is defined as:

$$\rho = \frac{1}{D} \sum_{d=1}^{D} \frac{(x_d - \bar{x})(y_d - \bar{z})}{\sigma_x \sigma_y}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the mean values and σ_x and σ_y are the standard deviations. The root-mean-squared-error for mean centered data points is defined as:

$$E^2 = \mathbb{E}\left[\frac{1}{D}\sum_{d=1}^{D}\left((x_d - \bar{x}) - (z_d - \bar{z})\right)^2\right]$$

$$= \underbrace{\frac{1}{D^2}\mathbb{E}\left[\sum_{d=1}^{D}(x_d - \bar{x})^2\right]}_{\sigma_x^2} + \underbrace{\frac{1}{D^2}\mathbb{E}\left[\sum_{d=1}^{D}(y_d - \bar{y})^2\right]}_{\sigma_y^2} - \underbrace{\frac{1}{D^2}\mathbb{E}\left[\sum_{d=1}^{D}(x_d - \bar{x})(y_d - \bar{y})\right]}_{\sigma_{xy}}$$

$$= \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\rho. \tag{2}$$

The key point of the TD is recognize the relationship between the statistical quantities in Eq. 2 and the law of cosines:

$$c^2 = a^2 + b^2 - 2ab\cos(\theta). \tag{3}$$

Here, a and b are the lengths of two sides of a triangle with angle θ between each other and an opposite side of length c. The TD has seen widespread use in several domains such as in geophysical sciences [6,8]. Nevertheless, the TD has some key weaknesses that limits it functionality in many practical applications. The Pearson correlation coefficient has a number of limitations [1]. It can only model linear relationships [2], which can be restricting in many practical application. Also, the Pearson correlation coefficient is known be sensitive to outliers [1].

The Kernelized Taylor Diagram. To address such limitations, we propose the KTD, which uses well-know measures from the kernel literature to model similarities between populations. The starting point of the KTD is one of the most widely used distance measures in the kernel literature, namely the maximum mean discrepancy (MMD) [7], which measures the distance between two distributions where each distributions is represented by a mean embedding of the data. Let $X \sim P$ and $Y \sim Q$, and $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ denoted the mean embedding vectors representing two distributions P and Q. Then, the MMD is defined as the norm between the two embeddings in a reproducing kernel Hilbert space $\mathcal{H}$:

$$\begin{aligned}
MMD^2 &= \|\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\|_{\mathcal{H}}^2 \\
&= \|\boldsymbol{\mu}_x\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_y\|_{\mathcal{H}}^2 - 2\langle\boldsymbol{\mu}_x, \boldsymbol{\mu}_y\rangle_{\mathcal{H}} \\
&= \|\boldsymbol{\mu}_x\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_y\|_{\mathcal{H}}^2 - 2\|\boldsymbol{\mu}_x\|_{\mathcal{H}}\|\boldsymbol{\mu}_y\|_{\mathcal{H}}\frac{\langle\boldsymbol{\mu}_x, \boldsymbol{\mu}_y\rangle_{\mathcal{H}}}{\|\boldsymbol{\mu}_x\|_{\mathcal{H}}\|\boldsymbol{\mu}_y\|_{\mathcal{H}}} \\
&= \|\boldsymbol{\mu}_x\|_{\mathcal{H}}^2 + \|\boldsymbol{\mu}_y\|_{\mathcal{H}}^2 - 2\|\boldsymbol{\mu}_x\|_{\mathcal{H}}\|\boldsymbol{\mu}_y\|_{\mathcal{H}}\cos\angle(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y).
\end{aligned} \tag{4}$$

In general, the true data distributions are not known, so the mean embeddings are replaced by empirical mean embeddings that are estimated based on samples from each distribution:

$$\hat{\boldsymbol{\mu}}_x = \frac{1}{N}\sum_{n=1}^{N}\kappa(\mathbf{x}_n, \cdot), \tag{5}$$

where $\kappa(\cdot, \cdot)$ is a positive definite kernel that measures similarity between data points. If the kernel is characteristic [7], MMD is a metric and is zero only if

the two distributions are equal. [5] showed that the well-known Gaussian kernel with kernel width σ, $G_\sigma(\mathbf{x}_i, \mathbf{x}_j) = \exp(||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma)$, is a characteristic kernel. Furthermore, MMD does not assume a particular distribution of the data, and can capture both non-linear and linear relationships between distributions.

Similarly as with the TD, we recognize the law of cosines in Eq. 4. The mean embeddings of the two distributions are the side lengths of a triangle with angle $\cos \angle(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)$ between each other and an opposite side with length equal to the MMD between the distributions. The KTD is shown in Fig. 1.

The length of the mean embeddings indicate the distance from the origin to each point in the KTD. For the Gaussian kernel, the kernel mean embedding captures all moments of the data population [9]. But it is not obvious how to interpret what information the kernel mean embeddings are illustrating in the diagram. However, the kernel mean embeddings can be related to uncertainty through the information potential (IP) from information theoretic learning [11], which allows for a similar interpretation of the KTD as the TD. That is, the kernel mean embeddings correspond to the σ in Eq. 2. In most applications, the IP must be estimated from data. In information theoretic learning, the IP is often estimated through the quadratic IP estimator using a Gaussian kernel [11]:

$$\hat{V}_{2,\sigma}(X) = \frac{1}{N^2} \sum_{i,j}^{N} G_\sigma(\mathbf{x}_i, \mathbf{x}_j). \tag{6}$$

Next, the squared norm terms in Eq. 4 can be expressed as:

$$\|\boldsymbol{\mu}_x\|_{\mathcal{H}}^2 = \frac{1}{N^2} \sum_{i,j}^{N} \kappa(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

If the mean embeddings are calculated using a Gaussian kernel, Eq. 6 and Eq. 7 are equivalent. Furthermore, the IP is related to entropy as follows:

$$\hat{H}_2(X) = -\log(\hat{V}_{2,\sigma}(X)). \tag{8}$$

Entropy measures the amount of information in a random variable, but can also be interpreted as a measure of uncertainty. High entropy indicates more variation in the data, while low entropy means that the data is clustered together. From Eq. 8 it is evident that when the information potential of X is high and the entropy will be low, and the opposite when the information potential of X is low. For the KTD, this means that random variables with a high value for the kernel mean embedding, and thus far from the origin, is associated with low uncertainty, and oppositely for a low value of the kernel mean embedding. This insight is important, as it allows us to relate concepts from the TD to the KTD.

3   Experiments

To illustrate the functionality of the KTD we consider the case were the true distribution of the data is known and generate 1000 samples from 5 different populations. The reference distribution X_{ref} is sampled from a standard normal distribution. The remaining populations are constructed as follows:
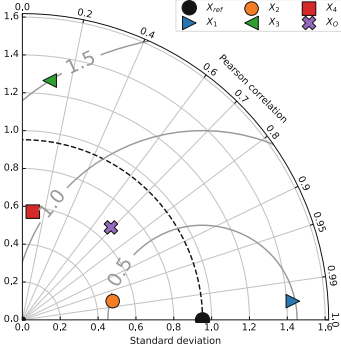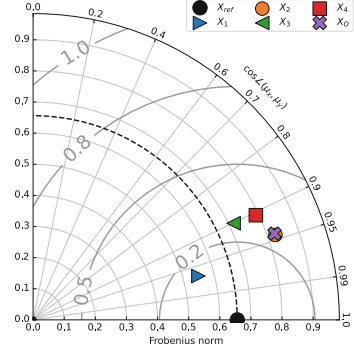
(a) Taylor diagram   (b) Kernelized Taylor diagram

Fig. 2. Comparison of TD with the KTD on the data described in Sect. 3. The experiment illustrates how the TD is not able to capture non-linear dependencies and is sensitive to outliers, when compared with the proposed KTD.

$$X_1 \sim 2X_{\mathrm{ref}} + \epsilon, \quad X_2 \sim \frac{X_{\mathrm{ref}}}{2} + \epsilon, X_3 \sim X_{\mathrm{ref}}^2 + \epsilon,$$

$$X_4 \sim X_{\mathrm{ref}} \sin(X_{\mathrm{ref}}) + \epsilon, X_O \sim \frac{X_{\mathrm{ref}}}{2} + \epsilon \text{ (with outliers)},$$

where $\epsilon \sim \mathcal{N}(0, 0.01)$. Population X_1 and X_2 are chosen to represent a linear relationship to the reference distribution, but with different scaling such that the standard deviation is different compared to the reference. Population X_3 and X_4 are chosen to represent a non-linear relationship with the reference. Lastly, X_0 is chosen to also have a linea relationship with the reference, but with two outliers added to the population. These two outliers are samples from $\mathcal{N}(10, 1)$.

Figure 2a displays the TD for these populations in relation to the reference distribution, while Fig. 2b shows the KTD. First, we consider Fig. 2a. Note that X_1 and X_2 both have a high similarity with the reference but with different length from the origin as a result of the difference in standard deviation. Next, both X_3 and X_4 are indicated as having low similarity with the reference, which is expected since the relationship is non-linear. Lastly, X_O, which is almost identical to X_2 except for two outliers, shows a much lower similarity score. This illustrates how sensitive the TD can be to outliers.

In Fig. 2b, X_1 and X_2 also shows a related and high similarity score. However, note that compared to Fig. 2a, the distance to the origin have been changed, which is explained through the connection to the information potential described in Sect. 2. Next, both X_3 and X_4 are now indicated to have a high similarity with the reference, which illustrates that the KTD is capable of capturing non-linear similarities. Lastly, X_2 and X_O are located at almost the same point in the diagram, which shows that the KTD is robust against outliers in the data.

4   Conclusion

In this article we proposed the KTD, which relates well-known quantities from the kernel literature in a single diagram. To the best of our knowledge, such a diagram has not been devised previously. Our proposed diagram addresses some key limitation in the widely used TD, such as modeling non-linear relationships and outliers in the data. In future works, we intend to examine the usability of the diagram on real-world data such as in climate applications. We believe that the KTD can be a useful tool in many machine learning applications.

References

1. Armstrong, R.A.: Should pearson's correlation coefficient be avoided? Ophthalmic Physiol. Opt. **39**(5), 316–327 (2019)
2. Correa, C.D., Lindstrom, P.: The mutual information diagram for uncertainty visualization. Int. J. Uncertain. Quantif. **3**, 187–201 (2013)
3. Cortes, C., Mohri, M., Rostamizadeh, A.: Algorithms for learning kernels based on centered alignment. J. Mach. Learn. Res. **13**, 795–828 (2012)
4. Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J.: On kernel-target alignment. In: Neural Information Processing Systems, pp. 367–373. MIT Press (2002)
5. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: Neural Information Processing Systems, vol. 20 (2008)
6. Gleckler, P.J., Taylor, K.E., Doutriaux, C.: Performance metrics for climate models. J. Geophys. Res. Atmos. **113**(D6) (2008)
7. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**, 723–773 (2012)
8. Jakob Themeßl, M., Gobiet, A., Leuprecht, A.: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. Int. J. Climatol. **31**(10), 1530–1544 (2011)
9. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean embedding of distributions: a review and beyond. Founds. Trends Mach. Learn. **10**(1–2), 1–141 (2017)
10. Taylor, K.E.: Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. **106**(D7), 7183–7192 (2001)
11. Xu, D., Erdogmuns, D.: Renyi's entropy, divergence and their nonparametric estimators. In: Principe, J.C. (ed.) Information Theoretic Learning. ISS, pp. 47–102. Springer, New York (2010). https://doi.org/10.1007/978-1-4419-1570-2_2