Model Criticism for Regression on the Grassmannian

Yi Hong[1], Roland Kwitt[3], and Marc Niethammer[1,2]

[1]University of North Carolina (UNC) at Chapel Hill, NC, US
[2]Biomedical Research Imaging Center, UNC-Chapel Hill, NC, US
[3]Department of Computer Science, University of Salzburg, Austria

Abstract. Reliable estimation of model parameters from data requires a suitable model. In this work, we investigate and extend a recent *model criticism* approach to evaluate regression models on the Grassmann manifold. Model criticism allows us to check if a model fits and if the underlying model assumptions are justified by the observed data. This is a critical step to check model validity which is often neglected in practice. Using synthetic data we demonstrate that the proposed model criticism approach can indeed reject models that are improper for observed data and that the approach can guide the model selection process. We study two real applications: degeneration of corpus callosum shapes during aging and developmental shape changes in the rat calvarium. Our experimental results suggest that the three tested regression models on the Grassmannian (equivalent to linear, time-warped, and cubic-spline regression in $\mathbb{R}^n$, respectively) can all capture changes of the corpus callosum, but only the cubic-spline model is appropriate for shape changes of the rat calvarium. While our approach is developed for the Grassmannian, the principles are applicable to smooth manifolds in general.

1 Introduction

In Euclidean space, regression models, e.g., linear least squares, are commonly used to estimate the relationship between variables. Recently, they have been successfully extended to manifolds, e.g., the manifold of diffeomorphisms [9], general Riemannian manifolds [2], and in particular the Grassmannian [6]. These models capture changes of data objects (e.g., images, or shapes) with respect to an associated descriptive variable, such as age. To measure the goodness-of-fit of these regression models, one usually computes the sum of squared errors (SSE), R^2 [2], or the mean squared error (MSE) using cross-validation.

While these measures assess fitting quality, they do not directly check if the underlying model assumptions hold. *Model criticism* does exactly that: it checks if a model's assumptions (including the noise model) are consistent with the observed data. It thereby provides valuable additional information beyond the classical measures for model fit. Recently, statistical model criticism [8] using a kernel-based two-sample test [3] has been proposed and its utility to evaluate regression models in Euclidean space was demonstrated.

In this paper, we take this approach one step further and use the fact that the strategy of [8] depends on a suitable kernel function for the data and can thus be extended to manifolds given such a kernel function. Given a population of data samples on the Grassmannian, we (1) perform regression, then (2) generate samples from the regression model and (3) assess whether the observed data could have been generated by the fitted model. We demonstrate the approach by criticizing three different regression models on the Grassmannian [7] using both synthetic and real data. We argue that model criticism is complementary to traditional measures of model fit, but has the advantage of directly assessing the suitability of a statistical model and its fit for given observed data.

Contributions. We propose an extension of model criticism to regression models on the Grassmannian. In particular, we extend the approach of [8] by (1) providing a strategy to generate Gaussian-distributed samples with a specific variance on this manifold, and (2) incorporating existing kernels into the two-sample testing strategy used for model criticism. We then apply the approach to check the validity of different regression models. Our experimental results are based on both synthetic and real data, including corpus callosum and rat calvarium shapes, providing insight into the appropriateness of the regression models beyond the customary use of the R^2 statistic.

2 Model criticism for regression in Euclidean space

The objective of model criticism for regression is to test for discrepancies between observed data and a model estimated from the data [8]. We assume the data observations are independently and identically distributed (i.i.d.) and that we can draw i.i.d. samples from the model respecting the noise assumptions; then, the key ingredient of model criticism is to measure whether these two samples are drawn from the same underlying distribution. To perform this *two-sample test*, a "kernelized" variant of the maximum mean discrepancy (MMD) [3] has been proposed as one choice of the test-statistic.

Review of kernel-based two-sample testing [3]. Assume we have i.i.d. samples $X = \{x_i\}_{i=1}^m$ and $Y = \{y_i\}_{i=1}^n$, drawn randomly from distributions p, q, defined on a domain $\mathcal{X}$. The goal of two-sample testing is to assess if $p = q$. One choice of a test-statistic is the MMD, defined as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)]) \ , \tag{1}$$

where $\mathcal{F}$ is a suitable class of functions $f : \mathcal{X} \to \mathbb{R}$. To uniquely measure whether $p = q$, Gretton et al. [3] let $\mathcal{F}$ be the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, i.e., $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, with associated reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The kernel can be written as $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$ and $\phi : \mathcal{X} \to \mathcal{H}$ denotes the feature map.

According to [3], Eq. (1) can then be expressed as the RKHS distance between the mean embeddings $\mu[p]$ and $\mu[q]$ of the distributions p and q, in particular

$\mu[p] := \mathbb{E}_{x\sim p}[\phi(x)]$. Since the mean embedding satisfies $\mathbb{E}_{x\sim p}[f(x)] = \langle \mu[p], f\rangle_{\mathcal{H}}$, Eq. (1) can be written as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{\|f\|_{\mathcal{H}}\leq 1} \langle \mu[p] - \mu[q], f\rangle = \|\mu[p] - \mu[q]\|_{\mathcal{H}} \tag{2}$$

with the empirical estimate (using the kernel function k)

$$\widehat{\mathrm{MMD}}[\mathcal{F}, X, Y] = \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)\right]^{\frac{1}{2}}. \tag{3}$$

In the following, we omit $\mathcal{F}$ and let $\widehat{\mathrm{MMD}}(X, Y)$ denote the computation of the MMD statistic on two samples X and Y using a suitable kernel function k.

Model criticism using two-sample testing. Assume we have data observations $\{Y_i^{obs}\}_{i=1}^N$, drawn from some distribution p, conditioned on an associated independent value t_i. A regression model M is estimated from the tuples $\{(t_i, Y_i^{obs})\}_{i=1}^N$. If the regression model is based on a Gaussian noise assumption[1], we can generate i.i.d. samples from the model; let this distribution be denoted by q and a sample by $\{Y_i^{est} = M(t_i) + n_i\}_{i=1}^N$, where $n_i \sim N(0, \sigma^2)$ and σ is the standard deviation of the residuals. Criticizing the model M now means to perform a two-sample test between $\{(t_i, Y_i^{obs})\}_{i=1}^N$ and $\{(t_i, Y_i^{est})\}_{i=1}^N$ under the null-hypothesis $H_0 : p = q$. This is done by computing the test-statistic between data observations and samples drawn from the regression model, i.e., $T^* = \widehat{\mathrm{MMD}}(\{(t_i, Y_i^{obs})\}_{i=1}^N, \{(t_i, Y_i^{est})\}_{i=1}^N)$. Then, to obtain the distribution of T under H_0, we repeatedly draw (from q) N i.i.d. samples to form two populations, $\{(t_i, Y_i^a)\}_{i=1}^N$ and $\{(t_i, Y_i^b)\}_{i=1}^N$. For each such draw j, we compute $T_j = \widehat{\mathrm{MMD}}(\{(t_i, Y_i^a)\}_{i=1}^N, \{(t_i, Y_i^b)\}_{i=1}^N)$ and thereby obtain the empirical distribution of T under H_0. Note that in case of observations in $\mathbb{R}^n$, computing the MMD statistic for model criticism is straightforward, since we can simply add t_i as an additional dimension to our data, i.e., we obtain observations in $\mathbb{R}^{n+1}$. The well-known RBF kernel can then be used to compute Eq. (3). We will see that this needs to be handled differently on manifolds.

Finally, we count the number of times that the bootstrapped statistics under H_0 are larger than the test statistic T^*, which results in a p-value estimate. Because T^* will be affected by the added random noise, we can also sample it a large number of times, resulting in a distribution of T^* and associated p-values.

3 Model criticism for regression on the Grassmannian

We now extend the model criticism approach of the previous section to the Grassmann manifold. The test objects are regression models on the Grassmannian, i.e., generalizations of classical regression in Euclidean space which minimize the sum of squared (geodesic) distances to the regression curves. Noise in these models is assumed to be Gaussian. To generalize the model criticism idea, one

[1] Other noise models can also be used as long as one can sample from them.

key ingredient is to draw random samples on the Grassmannian at each point on the regression curves. The other key ingredient is a suitable kernel k for Eq. (3) and a strategy to include the independent value into the kernel.

Drawing random samples on the Grassmannian. Similar to the Euclidean case, we assume we have N data observations on the Grassmannian $\mathcal{G}(r,s)$[2] with associated independent values, i.e., $\{(t_i, \mathbf{Y}_i^{obs})\}_{i=1}^N$. Using a regression model M estimated from this data, we can compute the corresponding data points on the regressed curve for each t_i as $\bar{\mathbf{Y}}_i^{est} = M(t_i)$. To draw sample points at each t_i under a Gaussian noise model, we adhere to the following strategy (although, other approaches such as the one outlined in [11] are possible). First, we compute the empirical standard deviation of the residuals as

$$\sigma = \sqrt{\frac{\sum_{i=1}^N d^2(\mathbf{Y}_i^{obs}, \bar{\mathbf{Y}}_i^{est})}{N-1}} \ , \tag{4}$$

where $d(\cdot,\cdot)$ denotes the geodesic distance on $\mathcal{G}(r,s)$. For each estimated data point $\bar{\mathbf{Y}}_i^{est}$, we then generate a tangent vector $\dot{\mathbf{Y}}_i^{est}$ as the projection of an $s \times r$ random matrix $\mathbf{Z}_i = [z_{uv}], z_{uv} \sim \mathcal{N}(0, \hat{\sigma}^2)$ onto the tangent space at $\bar{\mathbf{Y}}_i^{est}$; this is done via $\dot{\mathbf{Y}}_i^{est} = (\mathbf{I}_s - \bar{\mathbf{Y}}_i^{est}(\bar{\mathbf{Y}}_i^{est})^\top)\mathbf{Z}_i$ where $\mathbf{I}_s$ is the $s \times s$ identity matrix. The random point $\mathbf{Y}_i^{est}$ (at t_i), is eventually computed via the Riemannian exponential map as

$$\mathbf{Y}_i^{est} = \mathrm{Exp}(\bar{\mathbf{Y}}_i^{est}, \dot{\mathbf{Y}}_i^{est}) \ . \tag{5}$$

We note that the standard deviation $\hat{\sigma}$ of the samples in $\mathbf{Z}_i$ is proportional to the standard deviation σ as computed by Eq. (4). In fact, it can be shown[3] that the resulting geodesic distance between $\bar{\mathbf{Y}}_i^{est}$ and $\mathbf{Y}_i^{est}$ has standard deviation $\hat{\sigma}\sqrt{rs}$. Consequently, we set $\hat{\sigma} = \sigma/\sqrt{rs}$ when creating the $\mathbf{Z}_i$.

Kernels for model criticism on the Grassmannian. The next step is to adjust the kernel-based two-sample test of [3] for model criticism on the $\mathcal{G}(r,s)$ by selecting a suitable kernel. In [5], several positive definite (and universal) kernels on $\mathcal{G}(r,s)$ have been proposed, e.g., RBF, Laplace and Binomial kernels. These kernels are constructed by using the Binet-Cauchy kernel k_{bc} [10] and the projection kernel k_p [4]. We selected the kernel $k_{l,p}(\mathbf{X},\mathbf{Y}) = \exp\left(-\beta\sqrt{r - \|\mathbf{X}^\top\mathbf{Y}\|_F^2}\right), \beta > 0$ from [5] for our experiments[4]. However, for model criticism as proposed in [8] to be used on regression models, we need to be able to compute the MMD test for $(t_i, \mathbf{Y}_i)$ (i.e., including the information about the independent variable). While this is simple in the Euclidean case (cf.

---

[2] The Grassmannian $\mathcal{G}(r,s)$ is the manifold of r-dimensional subspaces of $\mathbb{R}^s$. A point on $\mathcal{G}(r,s)$ is identified by an $s \times r$ matrix $\mathbf{Y}$ with $\mathbf{Y}^\top\mathbf{Y} = \mathbf{I}_r$.

[3] Say we have samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ with Fréchet mean $\bar{\mathbf{Y}}$. The Fréchet variance is defined as $\sigma^2 = \min_{\mathbf{Y} \in \mathcal{G}(r,s)} {}^1/_n \sum_{i=1}^n d^2(\mathbf{Y}, \mathbf{Y}_i)$; this can equivalently be written as $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(\dot{\mathbf{Y}}_i^\top \dot{\mathbf{Y}}_i)$ with $\dot{\mathbf{Y}}_i = \mathrm{Exp}^{-1}(\bar{\mathbf{Y}}, \mathbf{Y}_i)$. Expanding the trace yields $\sigma^2 = \sum_{j=1}^s \sum_{k=1}^r [{}^1/_n \sum_{i=1}^n y_{i,j,k}^2]$ where the term in brackets is $\hat{\sigma}^2$ since $y_{i,j,k}$ is i.i.d. as $\mathcal{N}(0, \hat{\sigma}^2)$. We finally obtain $\sigma^2 = rs\hat{\sigma}^2$.

[4] Here, r denotes the subspace dimension in $\mathcal{G}(r,s)$, as defined before.

Section 2), the situation on manifolds is more complicated since we cannot simply add t_i to our observations. Our strategy is to use an RBF kernel for the t_i, i.e., $k_{rbf}(t, t') = \exp(-(t-t')^2/(2\gamma^2))$ and then leverage the closure properties of positive definite kernels, which allow multiplication of positive definite kernels. This yields our *combined kernel* as

$$k((t_i, \mathbf{X}), (t_j, \mathbf{Y})) = \exp\left(-\frac{(t_i - t_j)^2}{2\gamma^2}\right) \cdot \exp\left(-\beta\sqrt{r - \|\mathbf{X}^\top \mathbf{Y}\|_F^2}\right) \quad . \tag{6}$$

In all the experiments, we set both β and γ to 1 for simplicity.

Model criticism on the Grassmannian. The computational steps for model criticism on the Grassmannian are listed below:

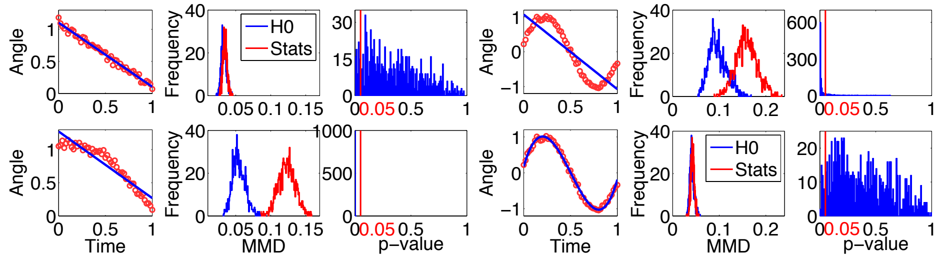
(1) Compute the points $\{\bar{\mathbf{Y}}_i^{est} = M(t_i)\}_{i=1}^N$ for each data observation $\mathbf{Y}_i^{obs}$ on $\mathcal{G}(r, s)$ from the estimated regression model M.
(2) Estimate the standard deviation of residuals, σ, using Eq. (4).
(3) Generate noisy samples, $\mathbf{Y}_i^{est}$ at each t_i using Eq. (5) and $\hat{\sigma}^2 = \sigma^2/(rs)$.
(4) Compute $T^* = \widehat{\text{MMD}}(\{(t_i, \mathbf{Y}_i^{est})\}_{i=1}^N, \{(t_i, \mathbf{Y}_i^{obs})\}_{i=1}^N)$ using Eqs. (3) & (6).
(5) Repeat (3) and (4) many times to obtain a population of T^*.
(6) Generate two groups of samples using (3), $\{(t_i, \mathbf{Y}_i^a)\}_{i=1}^N$, and $\{(t_i, \mathbf{Y}_i^b)\}_{i=1}^N$, and compute $T = \widehat{\text{MMD}}(\{(t_i, \mathbf{Y}_i^a)\}_{i=1}^N, \{(t_i, \mathbf{Y}_i^b)\}_{i=1}^N)$.
(7) Repeat (6) many times to obtain a distribution of T under H_0.
(8) Compute a p-value for each T^* in (5) with respect to the distribution of T from (7), resulting in a population of p-values. This allows to reject the null-hypothesis that the observed data distribution is the same as the sampled data distribution of the model at a chosen significance level α (e.g., 0.05).

4 Experimental results

We criticize three different regression models on the Grassmannian [7]: Std-GGR, TW-GGR, and CS-GGR, which are generalizations of linear least squares, time-warped, and cubic-spline regression, respectively. These three models are estimated by energy minimization from an optimal-control point of view.

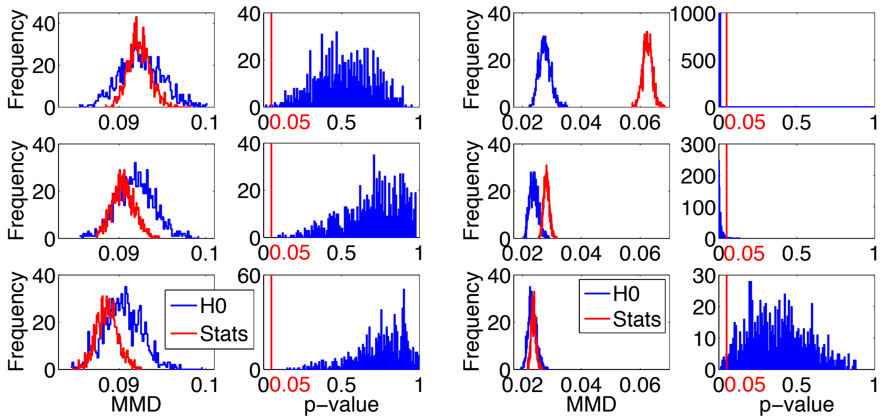
Synthetic data. For easy visualization, we synthesize data on the Grassmannian $\mathcal{G}(1, 2)$, i.e., the space of lines through the origin in $\mathbb{R}^2$. Each point uniquely corresponds to an angle in $[-\pi/2, \pi/2]$ with respect to the horizontal axis.

(1) Different data distributions. The first experiment is to perform model criticism on *one* regression model, but for different data distributions. To generate this data, we select two points on $\mathcal{G}(1, 2)$ to define a geodesic curve, and then uniformly sample 51 points along this geodesic from one point at time 0 to the other point at time 1. Using the strategy described in Section 3, Gaussian noise with $\sigma = 0.05$ is then added to the sampled points along the geodesic, resulting in the 1st group (Group 1) of synthetic data. The 2nd group (Group 2) of synthetic data is simulated by concatenating two different geodesics. Again, 51 points are uniformly sampled along the curve and Gaussian noise is added. The left column in Fig. 1(a) shows the two synthetic data sets using their corresponding angles.



(a) Different data distributions (b) Different regression models

Fig. 1: Model criticism for synthetic data on the Grassmannian. (a) Different data distributions are fitted by one regression model (Std-GGR); (b) One data distribution is fitted by different regression models (*top*: Std-GGR, *bottom*: CS-GGR).



(a) Corpus callosum (b) Rat calvarium

Fig. 2: Model criticism for real data. From *top* to *bottom*: the regression model corresponds to Std-GGR, TW-GGR, and CS-GGR respectively.

Both groups are fitted using a standard geodesic regression model (Std-GGR); the qualitative and quantitative results of model criticism are reported in Fig. 1(a) and Table 1, respectively. Among 1000 trials, H_0 is rejected in only 8.1% of all cases for Group 1 (generated from one geodesic), but H_0 is always rejected for Group 2 (i.e., generated from two geodesics). As expected, Std-GGR is not an appropriate model to capture the distribution of the data belonging to multiple geodesics. Model criticism correctly identifies this, while the R^2 values are hard to interpret with respect to model suitability.

(2) Different regression models. The second experiment is to generate one data distribution, but to estimate different regression models. We first generate a section of a sine curve with x-axis as time and y-axis as the angle. Each angle θ corresponds to a point on the Grassmannian, i.e., $[\cos\theta; \sin\theta] \in \mathcal{G}(1,2)$. In this way, we can generate polynomial data on $\mathcal{G}(1,2)$ with associated time $t \in [0,1]$.

	Different data distributions		Different regression models	
	Group 1	Group 2	Std-GGR	CS-GGR
R^2	0.99	0.89	0.73	0.99
%(p-values < 0.05)	8.1%	100%	88.0%	4.5%

Table 1: Comparison of R^2 measure and model criticism for synthetic data.

	Corpus callosum			Rat calvarium		
	Std-GGR	TW-GGR	CS-GGR	Std-GGR	TW-GGR	CS-GGR
R^2	0.12	0.15	0.21	0.61	0.79	0.81
%(p-values < 0.05)	0.2%	0%	0%	100%	98.0%	1.3%

Table 2: Comparison of R^2 measure and model criticism for real data.

A visualization of the generated data with added Gaussian noise ($\sigma = 0.05$) is shown in the left column of Fig. 1(b). The data points are fitted by a standard geodesic regression model (Std-GGR) and its cubic spline variant (CS-GGR), respectively. The results of model criticism are shown in Fig. 1(b) and Table 1; in 88.0% of 1000 trials, H_0 is rejected for the Std-GGR model, while we only reject H_0 in 4.5% of all trials for CS-GGR. As designed, CS-GGR has better performance than Std-GGR and can appropriately capture the distribution of the generated data, as confirmed by model criticism.

Real data. Two real applications, shape changes of the corpus callosum during aging and landmark changes of the rat calvarium with age, are used to evaluate model criticism for three regression models on the Grassmannian.

(1) Corpus callosum shapes. The population of corpus callosum shapes is collected from 32 subjects with ages varying from 19 to 90 years. Each shape is represented by 64 2D boundary landmarks. We represent each shape matrix as a point on the Grassmannian using an affine-invariant shape representation [1]. As shown in Fig. 2(a) and Table 2, although the R^2 values of the three regression models are relatively low, our model criticism results with 1000 trials suggest that all three models may be appropriate for the observed data.

(2) Rat calvarium landmarks. We use 18 individuals with 8 time points from the Vilmann rat data[5]. The time points range from 7 to 150 days. Each shape is represented by a set of 8 landmarks. We project each landmark-based rat calvarium shape onto the Grassmannian, using the same strategy as for the corpus callosum shapes. From the model criticism results shown in Fig. 2(b) and Table 2 we can see that the equivalent linear (Std-GGR) and time-warped (TW-GGR) models cannot faithfully capture the distribution of the rat calvarium landmarks. However, the cubic-spline model is not rejected by model criticism and therefore appears to be the best among the three models. This result is

[5] Available online: http://life.bio.sunysb.edu/morph/data/datasets.html

consistent with the R^2 values, and also provides more information about the regression models. As we can see, the R^2 values of TW-GGR and CS-GGR are very close, but the model criticism suggests CS-GGR is the one that should be chosen for regression on this dataset.

5 Discussion

We proposed an approach to *criticize* regression models on the Grassmannian. This approach provides complementary information to the existing measure(s) for checking the goodness-of-fit for regression models, such as the customary R^2 statistic. While we developed the approach for the Grassmannian, the general principle can be extended to other smooth manifolds, as long as one knows how to add the modeled noise into samples and a positive definite kernel is available for the underlying manifold. Two important ingredients of model criticism are the estimation of the parameters of the noise model on the manifold (in our case Gaussian noise and therefore we estimate the standard deviation) and the selection of the kernel. Exploring, in detail, the influence of the estimator and the kernel and its parameters is left as future work.

Acknowledgments This work was supported by NSF grants EECS-1148870 and EECS-0925875.

References

1. Begelfor, E., Werman, W.: Affine invariance revisited. In: CVPR (2006) 7
2. Fletcher, P.: Geodesic regression and the theory of least squares on Riemannian manifolds. Int. J. Comput. Vision 105(2), 171–185 (2013) 1
3. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. 13, 723–773 (2012) 1, 2, 4
4. Hamm, J., Lee, D.D.: Grassmann discriminant analysis: a unifying view on subspace-based learning. In: ICML (2008) 4
5. Harandi, M.T., Salzmann, M., Jayasumana, S., Hartley, R., Li, H.: Expanding the family of Grassmannian kernels: An embedding perspective. In: ECCV (2014) 4
6. Hong, Y., Kwitt, R., Singh, N., Davis, B., Vasconcelos, N., Niethammer, M.: Geodesic regression on the Grassmannian. In: ECCV (2014) 1
7. Hong, Y., Singh, N., Kwitt, R., Vasconcelos, N., Niethammer, M.: Parametric regression on the Grassmannian, http://arxiv.org/abs/1505.03832 2, 5
8. Lloyd, J.R., Ghahramani, Z.: Statistical model criticism using kernel two sample tests, Preprint at http://mlg.eng.cam.ac.uk/Lloyd/papers/ 1, 2, 4
9. Niethammer, M., Huang, Y., Vialard, F.X.: Geodesic regression for image time-series. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI. vol. 6892, pp. 655–662. Springer, Heidelberg (2011) 1
10. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. J. Mach. Learn. Res. 4, 913–931 (2003) 4
11. Zhang, M., Fletcher, P.T.: Probabilistic principal geodesic analysis. In: NIPS. pp. 1178–1186 (2013) 4