HOW TO REGRESS ON A SPHERE ELEGANTLY?

H.LUO

Abstract. There is no way of doing that. This method is highly data-structure dependent and should be avoided if we have no idea about the underlying structure about the data set.

(1) The regression on a sphere
Regression on a sphere is no more than the following diagram. It can either be regarded as a special case of regression problem or an estimation problem of the composition transformations $c_1 \circ f \circ c_2$. These two views leads to flattening approach OR direct estimation approach. [1]

$$\begin{array}{ccc} \mathbb{R}^d & & \mathbb{R}^d \\ \uparrow Chart & & \uparrow Chart \\ \mathbb{S}^{d-1} & \overset{Regression}{\rightarrow} & \mathbb{S}^{d-1} \end{array}$$

The key that makes a spherical regression to be special is that instead of density estimation, the regression function(if no action other than rigid actions are concerned) is lying in a finite dimensional space $SO(d)$, which makes the problem simpler. In following cases the spherical regression works nicely:
   (a) The data is obtained from a sphere. (Aden Gulf data, near-surface satellite scanning with perturbations)
   (b) The data seems to be related by the rigid motion groups. (Toy example of red-yellow points [1])
(2) Two approaches
   (a) Transformation Estimation
      (i) Pros
         (A) Finite dimensional space $SO(d)$.
         (B) Simple mimic of the $y = x + a + \epsilon$ model in linear case.
         (C) PCA-like techniques are applicable in this fashion, obviously more suitable[2].
      (ii) Cons
         (A) Limited transformations can be chosen as the correct transformation.
         (B) Not robust to data with many random perturbations.
         (C) Transformation group has not been extended to higher dimensional spheres in the context of a regression problem.
      (iii) Basic methods
         (A) Procrustes rotation estimation using SVD decomposition and watch the principal eigen vectors to determine the directions.

Date: 07-29-2016.

(B) Assuming the data have enough information to yield strong consistency $\frac{1}{n}\sum^n \boldsymbol{x'x} \to \boldsymbol{\Sigma}$ regular or rank $d-1$.
(C) [1] generalized the $SO(d)$ into the projective transformation group $PGL(d) \coloneqq GL(d)/scalaring\,group \cong SL(d)$ via $A \mapsto \frac{A}{[detA]^{\frac{1}{d}}}$. $SL(d)$ has tangent space the Lie group $sl(d) \coloneqq \left\{A \in \mathbb{R}^{d\times d} : traceA = 0\right\}$ at its identity element.

(b) Flattening
(i) Pros
(A) Reduction to a linear regression problem with calculation ease.
(B) Generally applicable to high dimensions.
(ii) Cons
(A) Global flattening of a sphere is a highly nonlinear map. [1]
(B) Global flattening can potentially distort relationships between predictor and response variables.

(3) Fisher von-Mises distribution
It is an element in the **dual space** of Sobolev space $L^{p,k}(\mathbb{S}^{d-1})$[3]. It is induced by the following density in $L^{p,k}(\mathbb{S}^{d-1})$, when $\kappa \to 0$ it is a Dirac funtional evaluated at $\boldsymbol{\mu}$. FVM happens to be a tempered distribution and we can even choose its Fourier transformation over the $\mathbb{S}^{d-1}$ which makes the calculation even more simpler and intrinsically related to the dynamic PDE over $\mathbb{S}^{d-1}$.
$f_\kappa = C_\kappa \cdot exp(\kappa \boldsymbol{\epsilon'\mu}), FVM = \langle f_\kappa, \bullet \rangle_{L^p}$ where C_κ is the normalizing constant.
Like the Gaussian process, the family of FVM distributions are closed under the action of $SO(d)$ as pointed out by [1], therefore it can also be regarded as a natural choice of Guassian process prior when we actually explore the infinite dimensional extension of this method. See Proposition 1 of [1].
(a) In the noiseless setting, there is a unique maximal MLE.
(b) In the noise setting, there is a strongly consistent MLE.

(4) The contribution in [1]
(a) Model
(i) Sufficient information to guarantee consistency.
With probability one, any i.i.d. random sample of sample size $n(> d)$must have a subsample of size $d+1$ in which any d elements are linearly independent.
(ii) Regression function
$\boldsymbol{y} \mid \boldsymbol{x} = FVM(\boldsymbol{\mu}_\kappa = \frac{A\boldsymbol{x}_i}{\|A\boldsymbol{x}_i\|}, \kappa)$ where $A \in PGL(d) \cong SL(d)$ with supremum norm.
(b) Solution
(i) MLE (profile MLE)
$\hat{A} = argmax_{A\in SL(d)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i' \cdot \frac{A\boldsymbol{x}_i}{\|A\boldsymbol{x}_i\|}$;
$\hat{\kappa} = argmax_{\kappa>0}\,[C_\kappa]^n \cdot exp\left[\kappa \cdot \sum_{i=1}^n \boldsymbol{y}_i' \cdot \frac{A\boldsymbol{x}_i}{\|A\boldsymbol{x}_i\|}\right]$
The strong consistency comes from the MLE when the Fisher information matrix is regular, which is ensured by our sufficient information assumption described above. The asymptotic normality is totally a consequence of the MLE. T**heir proof can**

be simplified by using the Laplace method of expansion. Using the asymptotic normality we can derive an approximate confident region as the author showed in the paper.

(ii) Newton-Raphson Procedure on $SL(d)$
Let the function of concern to be $f_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i' \cdot \frac{A\boldsymbol{x}_i}{\|A\boldsymbol{x}_i\|}$, the authors are using classical NR method to solve the equation $\nabla_A f_n = 0$, thus they derive $\nabla_A f_n$ to make linear interpolation. $\delta_A f_n = \nabla_A f_n + \left\langle \nabla_A f_n, \frac{A^{-1}}{\|A^{-1}\|} \right\rangle \cdot \frac{A^{-1}}{\|A^{-1}\|}$ [1]
This is the tangent shooting vector of f_n in $T_A(SL(d)) \cong sl(d)$. Meanwhile, the calculation of derivatives and Hessian of f_n is actually equivalent to finding the first and second fundamental form of the integral curves of ∇f_n over $\mathbb{S}^{d-1}$. As I mentioned before, such calculation can also be proceeded using generalized function and Fourier coefficients elegantly [3], yet in this scenario I agree that the direct computation will reveal the geometric nature better.
Prof.Kurtek gave some inspiring comments. To minimize the L^2norm of residuals $\|\boldsymbol{y} - A\boldsymbol{x}\|^2$ is equivalent to maximize their correlationship $\langle \boldsymbol{y}, A\boldsymbol{x}\rangle$ due to $\|\boldsymbol{y} - A\boldsymbol{x}\|^2 = \|\boldsymbol{y}\|^2 + \|A\boldsymbol{x}\|^2 + 2\langle \boldsymbol{y}, A\boldsymbol{x}\rangle$ and A will preserve the norm.

(A) Initialization. Let $\sum_{i=1}^n \boldsymbol{y}_i\boldsymbol{x}_i'$ and choose its SVD vector matrix as initial value. **This choice makes no special sense to me, and I do not think it will improve the performance of the whole algorithm.**
(B) Construct an orthonormal basis on $T_A(SL(d))$ and $\delta_A f_n$. Hessian matrix is also computed to aid the choice of update direction[2].
(C) Project $\delta_A f_n$ to every direction of the orthonormal basis and choose the rapidest descending direction as update direction.

(5) Examples and comments
(a) Fixed-mean model
(b) Rigid rotation model in the classical spherical regression setting
(c) Log-linear model
(d) Classification using the $PGL(d)$, the dis-similiarity measure is chosen to be the maximized likelihood function.[3] [4]

[1] I think it is not appropriate to denote delta functional in such a way, yet they are practical statisticans, so...

[2] "In other words, if we perturb A in the direction of V , the Hessian measures the resulting changes in the gradient vector" [1]

[3] In short this paper discuss only one model, that is the transformation model with FVM noise using the MLE solved numerically by the NR method using Hessian to choose the rapidest direction.

[4] One comment I have to make is about the cloud formation example. I feel it very inappropriate to approximate dynamics using $PGL(d)$ even for approximated results since if we proceed like that we implicitly assume that all perturbations in this dynamics are transient, which is super strong an assumption I would like to discard.

References

[1] Rosenthal, Michael, et al. "Spherical regression models using projective linear transformations." Journal of the American Statistical Association 109.508 (2014): 1615-1624.

[2] Fletcher, P. Thomas, et al. "Principal geodesic analysis for the study of nonlinear statistics of shape." IEEE transactions on medical imaging 23.8 (2004): 995-1005.

[3] Zemanian, Armen H. Distribution theory and transform analysis: an introduction to generalized functions, with applications. Courier Corporation, 1965.