Data analysis on Hilbert manifolds and shapes of planar contours

Leif Ellingson[†][1], Vic Patrangenaru [*][2] and Frits Ruymgaart[†]

[†] Department of Mathematics and Statistics, Texas Tech University
[*] Department of Statistics, Florida State University

1 Large sample behavior for means on Hilbert manifolds

DEFINITION 1.1. *Assume* $\mathbf{H}$ *is a separable, infinite dimensional Hilbert space over the reals. A chart on a separable metric space* $(\mathcal{M}, \rho)$ *is a one to one homeomorphism* $\varphi : U \to \varphi(U)$ *defined on an open subset* U *of* $\mathcal{M}$ *to a Hilbert space* $\mathbf{H}$. *A Hilbert manifold is a separable metric space* $\mathcal{M}$, *that admits an open covering by domain of charts, such that the transition maps* $\varphi_V \circ \varphi_U^{-1} : \varphi_U(U \cap V) \to \varphi_V(U \cap V)$ *are differentiable.*

EXAMPLE 1.1. *The projective space* $P(\mathbf{H})$ *of a Hilbert space* $\mathbf{H}$, *the space of all one dimensional linear subspaces of* $\mathbf{H}$, *has a natural structure of Hilbert manifold modelled over* $\mathbf{H}$. *Define the distance between two vector lines as their angle, and, given a line* $\mathbb{L} \subset \mathbf{H}$, *a neighborhood* $U_{\mathbb{L}}$ *of* $\mathbb{L}$ *can be mapped via a homeomorphism* $\varphi_{\mathbb{L}}$ *onto an open neighborhood of the orthocomplement* $\mathbb{L}^{\perp}$ *by using the decomposition* $\mathbf{H} = \mathbb{L} \oplus \mathbb{L}^{\perp}$. *Then if* $\mathbb{L}_1 \perp \mathbb{L}_2$, *the map is a* $\varphi_{\mathbb{L}_1} \circ \varphi_{\mathbb{L}_2}^{-1}$ *is differentiable map between open subsets in* $\mathbb{L}_1^{\perp}$, *respectively in* $\mathbb{L}_2^{\perp}$.

DEFINITION 1.2. *An embedding of a Hilbert manifold* $\mathcal{M}$ *in a Hilbert space* $\mathbb{H}$ *is a one-to-one differentiable function* $j : \mathcal{M} \to \mathbb{H}$, *such that for each* $x \in \mathcal{M}$, *the differential* $d_x j$ *is one to one, and the range* $j(\mathcal{M})$ *is a closed subset of* $\mathbb{H}$ *and the topology of* $\mathcal{M}$ *is induced via* j *by the topology of* $\mathbb{H}$.

EXAMPLE 1.2. *The Veronese-Whitney (VW) embedding* $j : P(\mathbf{H})$ *in* $\mathcal{L}_{HS} = \mathbf{H} \otimes \mathbf{H}$ *(Kent (1992)) is given by*

$$j([\gamma]) = \frac{1}{\|\gamma\|^2} \gamma \otimes \gamma. \tag{1}$$

DEFINITION 1.3. *If* $j{:}\mathcal{M} \to \mathbb{H}$ *is an embedding and given a random object* X *on* $\mathcal{M}$, *the associated Fréchet function is* $\mathcal{F}_j(x) = E(\|j(X) - j(x)\|^2)$. *The set of all minimizers of* $\mathcal{F}_j$ *is the extrinsic mean set of* X. *If the extrinsic mean set has one element only, that element is called the extrinsic mean and is labeled* $\mu_{E,j}$ *or simply* μ_E.

PROPOSITION 1.1. *Consider a random object* X *on* $\mathcal{M}$ *and assume* $j(X)$ *has the mean vector* μ. *Then the extrinsic mean set is the set of all points* $x \in \mathcal{M}$, *such that* $j(x)$ *is at minimum distance from* μ. *(iii) In particular,* μ_E *exists if there is a unique point on* $j(\mathbf{M})$ *at minimum distance from* μ, *the projection* $P_j(\mu)$ *of* μ *on* $j(\mathbf{M})$, *and in this case* $\mu_E = j^{-1}(P_j(\mu))$.

The *VW mean* (extrinsic mean for a random object $X = [\Gamma]$ on $P(\mathbf{H})$ with respect to the VW embedding) exists if and only if $E(\frac{1}{\|\Gamma\|^2}\Gamma \otimes \Gamma)$ has a simple largest eigenvalue, in which case, the VW mean is $\mu_E = [\gamma]$, where γ is an eigenvector for this eigenvalue.

[1] Research partially supported by National Science Foundation Grant DMS-0805977

[2] Research supported by National Science Foundation Grants DMS-0805977, DMS-1106935

2 A one-sample test of the neighborhood hypothesis

Assume Σ_j is the extrinsic covariance operator of a random object X on the Hilbert manifold $\mathcal{M}$, with respect to the embedding $j : \mathcal{M} \to \mathbb{H}$. Let $\mathbf{M}_0$ be a compact submanifold of $\mathcal{M}$. Let $\varphi_0 : \mathcal{M} \to \mathbb{R}$ be the function

$$\varphi_0(p) = \min_{p_0 \in \mathbf{M}_0} \|j(p) - j(p_0)\|^2, \tag{2}$$

and let $\mathbf{M}_0^\delta, \mathbb{B}_0^\delta$ be given respectively by

$$\mathbb{M}_0^\delta = \{p \in \mathcal{M}, \varphi_0(p) \leq \delta^2\}, \mathbb{B}_0^\delta = \{p \in \mathcal{M}, \varphi_0(p) = \delta^2, \}. \tag{3}$$

Since φ_0 is Fréchet differentiable and all small enough $\delta > 0$ are regular values of φ_0, it follows that $\mathbf{B}_0^\delta$ is a Hilbert submanifold of codimension one in $\mathcal{M}$. Let ν_p be the normal space at a points $p \in \mathbf{B}_0^\delta$, orthocomplement of the tangent space to $\mathbb{B}_0^\delta$ at p. We define $\mathbb{B}_0^{\delta,X}$

$$\mathbb{B}_0^{\delta,X} = \{p \in \mathbf{B}_0, \Sigma_j|_{\nu_p}\text{is positive definite}\}. \tag{4}$$

DEFINITION 2.1. *The neighborhood hypothesis consists in the following two alternatives:*

$$H_0 : \mu_E \in M_0^\delta \cup B_0^{\delta,X} vs. H_1 : \mu_E \in (M_0^\delta)^c \cap (B_0^{\delta,X})^c. \tag{5}$$

Munk et al. (2008) show that, in general, the test statistic for these types of hypotheses has an asymptotically standard normal distribution for large sample sizes, in the case of random objects on Hilbert spaces. Here, we consider neighborhood hypothesis testing for the particular situation in which the submanifold $\mathbf{M}_0$ consists of a point m_0 on $\mathcal{M}$. We set $\varphi_0 = \varphi_{m_0}$, and since $T_{m_0}\{m_0\} = 0$ we will prove the following result.

THEOREM 2.1. *If $M_0 = \{m_0\}$, the test statistic for the hypotheses specified in (5) has an asymptotically standard normal distribution and is given by:*

$$T_n = \sqrt{n}\{\varphi_{m_0}(\hat{\mu}_E) - \delta^2\}/s_n, s_n^2 = 4\langle\hat{\nu}, S_{E,n}\hat{\nu}\rangle where \tag{6}$$

$$S_{E,n} = \frac{1}{n}\sum_{i=1}^{n}(\tan_{\hat{\mu}} d_{\overline{j(X)}_n} P_j(j(X_i) - \overline{j(X)}_n)) \otimes (\tan_{\hat{\mu}} d_{\overline{j(X)}_n} P_j(j(X_i) - \overline{j(X)}_n)) \tag{7}$$

is the extrinsic sample covariance operator for $\{X_i\}_{i=1}^n$, and

$$\hat{\nu} = (d_{\hat{\mu}_{E,n}} j)^{-1} \widehat{tan}_{j(\hat{\mu}_{E,n})}(j(m_0) - j(\hat{\mu}_{E,n})). \tag{8}$$

3 Neighborhood Hypothesis for Mean Shape of a Contour

We consider contours, boundaries of 2D topological disks in the plane. To keep the data analysis stable, and to assign a *unique* labeling, we make the *generic* assumption that there is a unique point p_0 on such a contour at the maximum distance to its center of mass so that the label of any other point p on the contour is the "counterclockwise" travel time at constant speed from p_0 to p. A *regular contour* $\tilde{\gamma}$ is regarded as the range of a piecewise differentiable *regular* arclength parameterized function $\gamma : [0, L] \to \mathbb{C}, \gamma(0) = \gamma(L)$, that is one-to-one on $[0, L)$. Two contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ *have the same direct similarity shape* if there is a direct similarity $S : \mathbb{C} \to \mathbb{C}$, such that $S(\tilde{\gamma}_1) = \tilde{\gamma}_2$. Two regular contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ have the same similarity shape if their centered counterparts satisfy to $\tilde{\gamma}_{2,0} = \lambda\tilde{\gamma}_{1,0}$, for some $\lambda \in \mathbb{C}\backslash 0$. Therefore Σ_2^{reg}, *set of all direct similarity shapes of regular contours,* is a dense and open subset of $P(\mathbf{H})$, the projective space corresponding to the Hilbert space $\mathbf{H}$ of all square integrable centered functions from S^1 to $\mathbb{C}$. Given any VW-nonfocal probability measure Q on $P(\mathbf{H})$, from Section 2 we see that if $\gamma_1, \dots, \gamma_n$ is a sample from Γ, then $\hat{\mu}_{E,n}$ is the projective point of the eigenvector corresponding to the largest eigenvalue of $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\gamma_i\|^2}\gamma_i\otimes\gamma_i$. Given n i.i.d.r. objects (i.i.d.r.o.'s) from a VW-nonfocal distribution on $P(\mathbf{H})$, the asymptotic distribution of $\overline{j(X)}_n$ is converges as follows

$$\sqrt{n}(\overline{j(X)}_n - \mu) \to_d \mathcal{G} \text{ as } n \to \infty, \tag{9}$$

where $\mathcal{G}$ has a Gaussian distribution $N_{\mathcal{L}_{HS}}(0, \Sigma)$ on $\mathcal{L}_{HS}$ a zero mean and covariance operator Σ. It follows that the projection $P_j : \mathcal{L}_{HS} \to j(P(\mathbf{H})) \subset \mathcal{L}_{HS}$ is given by

$$P_j(A) = \nu_A \otimes \nu_A, \tag{10}$$

where ν_A is the eigenvector of norm 1 corresponding to the largest eigenvalue δ_1^2 of A, $P_j(\mu) = j(\mu_E)$, and $P_j(\overline{j(X)}_n) = j(\hat{\mu}_{E,n})$.

REMARK 3.1. *Applying the delta method to (9) Ellingson et al.(2013) arrived at a C L T for the VW extrinsic sample mean $\hat{\mu}_{E,n}$. Because of the infinite dimensionality, in practice, a sample estimate for the covariance operator is always degenerate, so one can not studentize.*

We may reduce the dimensionality via the neighborhood hypothesis methodology. Suppose that $j : P(\mathbf{H}) \to \mathcal{L}_{HS}$ is the VW embedding in (1) and $\delta > 0$ is a given positive number. Using the notation in Section 2, we now can apply the result in Section 3 to random shapes of regular contours. Assume $x_r = [\gamma_r], \|\gamma_r\| = 1, r = 1, \dots, n$ is a random sample from a VW-nonfocal probability measure Q. Asymptotically the tangential component of the VW-sample mean around the VW-population mean has a complex multivariate normal distribution. In particular, if we extend the CLT for VW-extrinsic sample mean Kendall shapes in Bhattacharya and Patrangenaru (2005), to the infinite dimensional case, the j-extrinsic sample covariance operator $S_{E,n}$, when regarded as an infinite Hermitian complex matrix has the following entries

$$S_{E,n,ab} = n^{-1}(\hat{\delta}_1^2 - \hat{\delta}_a^2)^{-1}(\hat{\delta}_1^2 - \hat{\delta}_b^2)^{-1} \tag{11}$$

$$\sum_{r=1}^{n} < e_a, \gamma_r >< e_b, \gamma_r >^* | < e_1, \gamma_r > |^2, a, b = 2, 3, \dots$$

with respect to the complex orthobasis $e_2, e_3, e_4, \dots$ of unit eigenvectors in the tangent space $T_{\hat{\mu}_{E,n}}P(\mathbf{H})$. Recall that this orthobasis corresponds via the differential $d_{\hat{\mu}_{E,n}}$ with an orthobasis (over $\mathbb{C}$) in the tangent space $T_{j(\hat{\mu}_{E,n})}j(P(\mathbf{H}))$, therefore one can compute the components $\hat{\nu}^a$ of $\hat{\nu}$ from equation (8) with respect to $e_2, e_3, e_4, \dots$, and derive for s_n^2 in (6) the following expression

$$s_n^2 = 4\sum_{a,b=2}^{\infty} S_{E,n,ab}\hat{\nu}^a\overline{\hat{\nu}^b}, \tag{12}$$

where $S_{E,n,ab}$ given in equation (11) are regarded as entries of a Hermitian matrix.

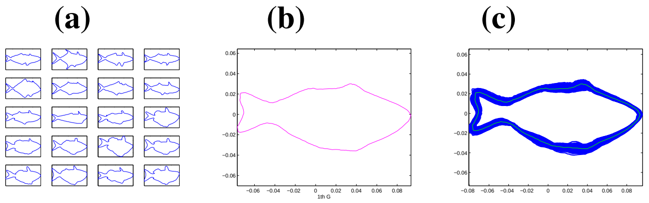

Figure 1: (a) Sample of 20 curves of red snapper fish and (b) the extrinsic mean shape of the sample of red snapper fish (c)Bootstrap 95% confidence region for the extrinsic mean shape of the red snapper fish

4 Bootstrap Confidence Regions for Means of Contours

Due to editorial space limitations, we present one result using selected planar closed curves kindly provided by Shantanu Joshi . A group of contours and its extrinsic mean shape is given below. Many others are given in Ellingson et. al. (2013). Similarly to the standard arithmetic mean, we see that the extrinsic mean provides a summary of the shapes by reducing the variability. This result is very noticeable with the red snapper above (Fig. 1). Another plus of the extrinsic mean is that the computation is fast (Bhattacharya et. al.(2012)).

One method for performing inference, is through nonparametric nonpivotal bootstrap (Efron (1979)). By repeatedly resampling from the available data and computing the distance between each resampled mean and the sample mean, we can obtain a confidence region for the extrinsic mean shape (for the sparse case, see Amaral et al (2010)). Above is an example of a 95% bootstrap confidence region for the set of contours in Fig. 1(a), based upon 400 resamples. The computations in Fig. 1(c) show that the confidence region behaves as expected, in that the width of the confidence region at various points along the contour reflects the amount of variability present in the data around the sample mean. For more examples see Ellingson et. al. (2013).

References

Amaral, G. J. A. ; Dryden, I. L.; Patrangenaru, V. and Wood, A.T.A. (2010). Bootstrap confidence regions for the planar mean shape. *JSPI.* **140**, 3026-3034.

Bhattacharya, R.N. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds- Part II, *Ann. Statist.*, **33**, No. 3, 1211- 1245.

Bhattacharya, R.N; Ellingson, L; Liu, X; Patrangenaru, V; and Crane, M. (2012) Extrinsic Analysis on Manifolds is Computationally Faster than Intrinsic Analysis, with Application to Quality Control by Machine Vision. *Applied Stochastic Models in Business and Industry.* **28**, 222-235.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.***7**, 1–26.

Ellingson, L., Patrangenaru, V. and Ruymgaart, F. (2013). Nonparametric Estimation of Means on Hilbert Manifolds and Extrinsic Analysis of Mean Shapes of Contours. *arXiv:1302.2126.*

Kent, J.T. (1992), New directions in shape analysis. *The Art of Statistical Science, A Tribute to G.S. Watson*, 115–127. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., Wiley, Chichester, 1992.

Klassen, E.; Srivastava, A. ; Mio, W. and Joshi, S. H. (2004). Analysis of Planar Shapes Using Geodesic Paths on Shape Spaces, IEEE Transactions on Pattern Analysis and Machine

Intelligence **26** 372 - 383.

Munk, A.; Paige, R.; Pang, J. ; Patrangenaru, V. and Ruymgaart, F. H.(2008). The One and Multisample Problem for Functional Data with Applications to Projective Shape Analysis. *J. of Multivariate Anal.* **. 99**, 815-833.