Uniformity in directional statistics: data-driven tests and free-lunch learning

Peter Jupp

University of St. Andrews, U.K.

http://www.mcs.st-andrews.ac.uk/~pej/

Directional statistics

Sample space:

circle: $S^1 = \{(\cos\theta, \sin\theta)\} = \left\{\mathbf{x} \in \mathbb{R}^2 : \mathbf{x}^T\mathbf{x} = 1\right\}$

sphere: $S^2 = \left\{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x}^T\mathbf{x} = 1\right\}$

rotation group: $SO(3) = \left\{\mathbf{X} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_3\right\}$

compact Riemannian manifold M

Uniform distributions

circle: $\theta \sim \pm\theta + c$

sphere: $\mathbf{x} \sim \mathbf{U}\mathbf{x}$ ($\mathbf{U}$ orthogonal)

rotation group: $\mathbf{X} \sim \mathbf{U}\mathbf{X}\mathbf{V}$ ($\mathbf{U}$, $\mathbf{V}$ orthogonal)

compact Riemannian manifold:
invariant under isometries

Testing uniformity on the circle

Rayleigh (1919)

For observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on S^1,

$$T_n \;=\; n \left\| \bar{\mathbf{x}} \right\|^2 \;=\; n \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \right\|^2$$

Reject uniformity for large T_n.

Sobolev tests of uniformity

$$\mathbf{t} : M \to \mathbb{E} \quad (\text{vector space})$$

For observations $x_1, \ldots, x_n$ on M,

$$T_n \;=\; n \left\| \bar{\mathbf{t}} \right\|^2 \;=\; \frac{1}{n} \left\| \sum_{i=1}^{n} \mathbf{t}(x_i) \right\|^2$$

Reject uniformity for large T_n.

T_n invariant under isometries.

Sobolev tests: construction of $\mathbf{t}$

Giné (1975)

$$E_k = k\text{th eigenspace of Laplacian} \qquad (k \geq 1)$$

$$\mathbf{t}_k : M \to E_k \subset L^2(M)$$

Define

$$\mathbf{t} : M \to L^2(M)$$

by

$$x \mapsto \mathbf{t}(x) \;=\; \sum_{k=1}^{\infty} a_k \mathbf{t}_k(x)$$

where

$$\sum_{k=1}^{\infty} {a_k}^2 \dim\, E_k < \infty$$

Problem

How to choose $a_1, a_2, \ldots$?

$$\begin{aligned} \text{few } a_k \neq 0 &\quad \Rightarrow \quad \text{(often) simple to calculate} \\ \text{all } a_k \neq 0 &\quad \Leftrightarrow \quad \text{consistent against all alternatives} \end{aligned}$$

Embarrassment of choice!

Data-driven tests of uniformity

Ledwina (1994, . . .): data-driven tests of fit on $\mathbb{R}$

Directional version:

(i) $(a_1, a_2, \ldots) = \left(\underbrace{1, 1, \ldots, 1}_{k}, 0, 0, \ldots \right)$

$$S_k \;=\; \sum_{r=1}^{k} \frac{1}{n} \left\| \sum_{i=1}^{n} \mathbf{t}_r(x_i) \right\|^2$$

score test

(ii) choose k using BIC (Schwarz, 1978)

Bayes Information Criterion:

Choose k to maximise

$$B_S(k) \;=\; S_k - \left(\sum_{r=1}^{k} \text{dim } E_r\right) \log n$$

Penalised score test

Penalises higher-dimensional models

Nice properties

Under uniformity, as $n \to \infty$,

$$\begin{aligned} \hat{k} &\xrightarrow{P} 1 \\ S_{\hat{k}} &\xrightarrow{d} \chi^2_{\nu_1} \end{aligned}$$

Consistent against all alternatives

Example: sphere, S^2

$$S_k \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{r=1}^{k}(2r+1)P_r(\mathbf{x}_i^T\mathbf{x}_j)$$

$$B_S(k) \;=\; S_k - k(k+2)\log n.$$

Example: sphere, S^2

	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 25$
1	8407	9635	9826	9900	9951
2	1060	338	167	97	49
3	357	25	6	3	0
4	92	2	1	0	0
5–10	84	0	0	0	0

Empirical distribution of $\hat{k}$
(10,000 simulations)

Example: sphere, S^2

α	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
0.10	0.132	0.114	0.108	0.110	**0.102**
	0.114	**0.105**	**0.101**	**0.102**	**0.097**
0.05	0.078	0.062	0.058	0.058	**0.051**
	0.056	**0.050**	**0.048**	**0.048**	0.043
0.01	0.043	0.023	0.017	0.014	**0.011**
	0.000	0.018	0.015	**0.010**	**0.008**

$P(S_{\hat{k}} \geq \chi^2_{3;\alpha})$ and $P(S^*_{\hat{k}} \geq \chi^2_{3;\alpha})$

$$S^*_{\hat{k}} = \left\{1 + \left(1.37 - 0.31 S_{\hat{k}}\right)/n\right\} S_{\hat{k}}$$

Red: values in $\alpha \pm 2\sqrt{\alpha(1-\alpha)/10,000}$.

Free-Lunch Learning

"There's no such thing as a free lunch"

In learning theory there is!

(i) Learn a foreign language; vocabulary $A_1 \cup A_2$
A_1, A_2 involve n_1, n_2 independent associations

(ii) Forget (partially) $A_1 \cup A_2$

(iii) Relearn *only* subset A_2

Then A_1 comes flooding back! 'Free lunch'!

Artificial Neural Network

ANN with weight vector $\mathbf{w} \in \mathbb{R}^n$ sends input $\mathbf{x}$ to output $\mathbf{w}^T\mathbf{x}$:

$$\mathbf{x} \;\mapsto\; \mathbf{w}^T\mathbf{x}$$

Teaching ANN to associate inputs $\mathbf{x}_1, \ldots, \mathbf{x}_c$ with outputs $d_1, \ldots, d_c$ puts weight vector into

$$\{\mathbf{w} : \mathbf{X}\mathbf{w} = \mathbf{d}\}$$

where

$$\begin{aligned} \mathbf{X} &= (\mathbf{x}_1, \ldots, \mathbf{x}_c)^T \\ \mathbf{d} &= (d_1, \ldots, d_c)^T \end{aligned}$$

Learning, forgetting, relearning

(i) Learn $A_1 \cup A_2$: weight vector is $\mathbf{w}_0$

$$\mathbf{X}_1\mathbf{w}_0 = \mathbf{d}_1 \qquad \mathbf{X}_2\mathbf{w}_0 = \mathbf{d}_2$$

(ii) Forget (partially) $A_1 \cup A_2$: weight vector is $\mathbf{w}_1$

squared error on A_1 is $\|\mathbf{X}_1\mathbf{w}_1 - \mathbf{d}_1\|^2$

(iii) Relearn *only* subset A_2: weight vector is $\mathbf{w}_2$
(orthogonal projection of $\mathbf{w}_1$ onto A_2)

squared error on A_1 is $\|\mathbf{X}_1\mathbf{w}_2 - \mathbf{d}_1\|^2$

After forgetting $A_1 \cup A_2$,

squared error on A_1 is $\|\mathbf{X}_1\mathbf{w}_1 - \mathbf{d}_1\|^2$

After relearning A_2,

squared error on A_1 is $\|\mathbf{X}_1\mathbf{w}_2 - \mathbf{d}_1\|^2$

Amount of FLL is

$$\delta = \|\mathbf{X}_1\mathbf{w}_1 - \mathbf{d}_1\|^2 - \|\mathbf{X}_1\mathbf{w}_2 - \mathbf{d}_1\|^2$$

$\delta > 0 \Leftrightarrow$ relearning A_2 brings $\mathbf{w}$ closer to A_1

(a) Synaptic drift

Forgetting $A_1 \cup A_2$ moves weight vector from $\mathbf{w}_0$ to $\mathbf{w}_1$.

$$\mathbf{w}_1 \;=\; \mathbf{w}_0 + \mathbf{v}$$

with

$\mathbf{v}$ isotropic ($\|\mathbf{v}\|^{-1}\mathbf{v}$ uniform)

'Free Lunch' results

If $n_1 + n_2 \leq n$, rows of $\mathbf{X}_1$ are i.i.d. and $\mathbf{d}_1$ is isotropic then

$$\begin{aligned} \text{median}\left(\delta\right) &> 0 \\ \mathrm{E}\left[\delta\right] &= \frac{n_1 n_2}{n^2} \mathrm{E}\left[\|\mathbf{x}\|^2\right] \mathrm{E}\left[\|\mathbf{v}\|^2\right] \\ P\left(\delta > 0\right) &\to 1 \qquad n \to \infty \end{aligned}$$

Under drift, free lunch is very probable!

(b) Synaptic fading

Forgetting $A_1 \cup A_2$ moves weight vector from $\mathbf{w}_0$ to $\mathbf{w}_1$.

$$\mathbf{w}_1 = r\mathbf{w}_0$$

If $n_1 + n_2 \leq n$, rows of $\mathbf{X}_1$ are i.i.d. and $\mathbf{d}_1$ is isotropic then

$$\begin{aligned} \mathrm{E}[\delta] &\leq 0 \\ P(\delta > 0) &\to 0, \qquad n \to \infty \end{aligned}$$

Under fading, free lunch is very unlikely