Change Detection Using Directional Statistics

T. Ide ("Ide-san"), D. Phan, J. Kalagnanam
IBM T. J. Watson Research Center

The 25th International Joint Conference on Artificial Intelligence (IJCAI-16)
Problem setting: change detection from multi-variate noisy time-series data

- Change = difference between $p(x)$ and $p_t(x)$
  - $x$: $M$-dimensional i.i.d. observation
  - $p(x)$: p.d.f. estimated from training window
  - $p_t(x)$: p.d.f. estimated from the test window at time $t$

- Question 1: What kind of model should we use for the pdf?
- Question 2: How can we quantify the difference between the p.d.f.’s?
We use von Mises-Fisher distribution to model $p(x)$ and $p_t(x)$

- **vMF distribution**: “Gaussian for unit vectors”
  \[
  p(z \mid u, \kappa) = c_M(\kappa) \exp(\kappa u^\top z)
  \]
  - $z$: random unit vector of $||z|| = 1$
  - $u$: mean direction
  - $\kappa$: “concentration” (~ precision in Gaussian)
  - $M$: dimensionality

- We are concerned only with the direction of observation $x$:
  - $z = \frac{x}{||x||}$
    - Normalization is always made
    - Do not care about the norm
Why do we enforce normalization? Rationale from the real world

- Real mechanical systems often incur multiplicative noise
  - Example: two belt conveyors operated by the same motor
  - Multiplicative noise equally applied to correlated variables

- Normalization is simple but powerful method for noise reduction
Mean direction \( u \) is learned via weighted maximum likelihood to down-weight contaminated samples

- Weighted likelihood function

\[
L(u, \kappa) = \sum_{n=1}^{N} w^{(n)} b^{(n)} \left\{ \ln c_M(\kappa) + \kappa u^\top z^{(n)} \right\}
\]

- Regularization over sample weights

\[
R(w) = \frac{1}{2} \|w\|_2^2 + \nu \|w\|_1
\]

- p.d.f. is learned by solving

\[
(u^*, w^*) = \arg \max_{u, w} \left\{ L(u, \kappa) + \lambda R(w) \right\}
\]

The term related to \( \kappa \) is less important. \( \kappa \) is treated as a given constant.
Multiple patterns (directions) can be obtained by coupling maximum likelihood equations

- Find orthogonal sequence of the mean direction $u_1, u_2, \ldots, u_m$ by coupling the weighted regularized maximum likelihood

\[
\begin{align*}
(u_1^*, w_1^*) &= \arg \max_{u_1, w_1} \{ L(u_1, \kappa) + \lambda R(w_1) \} \\
(u_2^*, w_2^*) &= \arg \max_{u_2, w_2} \{ L(u_2, \kappa) + \lambda R(w_2) \} \\
& \vdots \\
(u_m^*, w_m^*) &= \arg \max_{u_m, w_m} \{ L(u_m, \kappa) + \lambda R(w_m) \}
\end{align*}
\]

Orthogonality condition

$u_i \top u_j = \delta_{i,j}$

Kronecker delta
Iterative sequential algorithm for the coupled maximum likelihood

- For each $i$, $w_i$ and $u_i$ are solved iteratively until convergence
- Analytic solution exists in each step
- Results in very simple fixed point equations

$$w_1 \circlearrowright u_1$$

$$w_2 \circlearrowright u_2 \perp u_1$$

$$\vdots$$

$$w_m \circlearrowright u_m \perp u_i$$

(i = 1, \ldots, m - 1)
(For reference) Derived fixed-point iteration algorithm

- Example: $i = 1$

Given $w_1$, solve
\[
\max_{u_1} \{ \kappa u_1^T X w_1 \} \text{ s.t. } u_1^T u_1 = 1
\]

Given $u_1$, solve
\[
\min_{w_1} \left\{ \frac{1}{2} \| w_1 - \frac{q}{\lambda} \|_2^2 + \nu \| w_1 \|_1 \right\}
\]

$q \equiv \ln c_M b + \kappa X^T u_1$

This Lasso problem is solved analytically

---

**Algorithm 1** RED algorithm.

**Input:** Initialized $w$. Regularization parameters $\lambda, \nu$. Concentration parameter $\kappa$. The number of major directional patterns $m$.

**Output:** $U = [u_1, \ldots, u_m]$ and $W = [w_1, \ldots, w_m]$.

for $j = 1, 2, \ldots, m$ do

while no convergence do

\[
u_j \leftarrow \kappa[l_M - U_{j-1}^T u_{j-1}] X w_j
\]

\[
u_j \leftarrow \text{sign}(u_j^T X w_j) \frac{u_j}{\| u_j \|_2}
\]

\[q_j \leftarrow \gamma b + \kappa X^T u_j
\]

\[w_j \leftarrow \text{sign}(q_j) \odot \max \left\{ \frac{|q_j|}{\lambda} - \nu 1, 0 \right\}
\]

end while

end for

Return $U$ and $W$. 

---
Theoretical property: The algorithm is reduced to the “trust-region subproblem” in $\nu \rightarrow 0$.

**Theorem 2.** When $\nu$ tends to 0, the nonconvex problem (5) is reduced to an optimization problem in the form of

$$\min_{u} \{ u^\top Qu + c^\top u \} \quad s.t. \quad u^\top u = 1, \quad (23)$$

which has a global solution obtained in polynomial time.

**Proof.** The non-convex optimization problem (23) is known as the trust region subproblem. For polynomial algorithms to the global solution, see [Sorensen, 1997; Tao and An, 1998; Hager, 2001; Toint et al., 2009]. Here we show how the algorithm is reduced to the trust region subproblem.

Useful to initialize the iterative algorithm.
With extracted directions, define the change score at time $t$ as

$$a^{(t)} = \min_{f,g} \int dx \ M(x|Uf, \kappa) \ln \frac{M(x|Uf, \kappa)}{M(x|U^{(t)}g, \kappa)}$$

$$f^\top f = 1, \ g^\top g = 1$$

Concisely represented by the top singular value of $U^\top U^{(t)}$

$$\mathbf{u}_1, \ldots, \mathbf{u}_m \quad \mathbf{v}_1, \ldots, \mathbf{v}_r$$

Training window (fixed or sliding)

Test window

$$U \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_m] \quad U^{(t)} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_r]$$
Experiment: Failure detection of ore belt conveyors

- vMF formulation successfully suppressed very noisy non-Gaussian noise of multiplicative nature
- ~40% of samples were automatically excluded from the model
- Better than alternatives
  - PCA, Hoteling $T^2$
  - Stationary subspace analysis [Blythe et al., 2012]

Learned sample weights

~ 40% samples have zero weight: automated sample size reduction

Training data

(simulation data)
Summary – Thank you for your attention!

- Proposed a new change detection algorithm featuring
  - (1) New feature extraction method based on weighted max. likelihood of the vMF distribution
  - (2) Change score based on parameterized KL divergence
- Showed the linkage with the trust region sub-problem