From Dynamical Systems to Kernel Based Feature Spaces and Back

Peter Tino
School of Computer Science
University of Birmingham
UK
The Alan Turing Institute
Learning from time series data

We observe an “input time series” \( \{ u(t) \} \), \( u(t) \in \mathcal{U} \).

At certain times \( t_i \) we are asked to produce responses \( \{ y(t_i) \} \), \( y(t_i) \in \mathcal{Y} \), based on what we have observed till the response times \( t_i \), \( \{ u(t) \mid t \leq t_i \} \).

State space models - general model class capturing temporal structure through the notion of information processing state (IPS).

IPS at time \( t \), \( x(t) \in \mathcal{X} \), codes information from the entire history of time series items seen up to time \( t \) that is needed in order to perform a given task.
IPS at time $t$ codes for the entire history of time series items we have seen up to time $t$. This is equivalent to IPS at time $t - 1$ and input at time $t$.

**Recursive updates of IPSs**

$$f : \mathcal{X} \times \mathcal{U} \times \Theta_f \rightarrow \mathcal{X}$$

$$\mathbf{x}(t) = f(\mathbf{x}(t - 1), \mathbf{u}(t); \theta_f)$$

**Reading out the model response from IPSs**

$$h : \mathcal{X} \times \Theta_h \rightarrow \mathcal{Y}$$

$$\mathbf{y}(t) = h(\mathbf{x}(t); \theta_h)$$
Learning the dynamic couplings $\theta_f$ can be difficult

To latch a piece of information for a potentially unbounded number of time steps we need attractive sets.

Grammatical:
all strings containing odd number of 2’s

Saddle node bifurcation
Reservoir computation

- Fixed “contractive” input driven dynamics - no need to train!
  - “general purpose” IPS suitable for a variety of tasks

- Only simple static readout is trained - very efficient
  - we just need to do find out what to extract from the “rich” IPS to produce the output

- different flavours:
  - Echo State Network (Jaeger)
  - Liquid State Machine (Maass et al.)
  - Fractal Prediction Machine (Tino)
  - Neural Prediction Machine (Tino et al.)
Static case

\[ w^T u + b = 0 \]

\[ w^T u + b > 0 \]

\[ w^T u + b < 0 \]
Classification

\[ y = \text{sgn}(\mathbf{w}^T \mathbf{u} + b) \]

often

\[ \mathbf{w} = \sum_i \alpha_i \cdot \mathbf{u}_i \]

and so

\[ y = \text{sgn} \left( \sum_i \alpha_i \cdot \mathbf{u}_i^T \mathbf{u} + b \right) \]

\[ = \text{sgn} \left( \sum_i \alpha_i \cdot \langle \mathbf{u}_i, \mathbf{u} \rangle + b \right) \]
Classification - feature space, kernel trick

Embed the input space in a high-dimensional feature space:

\[ u \mapsto \phi(u) \in \mathcal{H} \]

\[
y = \text{sgn} \left( \sum_i \alpha_i \cdot \langle \phi(u_i), \phi(u) \rangle_{\mathcal{H}} + b \right)
\]

\[
= \text{sgn} \left( \sum_i \alpha_i \cdot K(u_i, u) + b \right)
\]

Kernel \( K(\cdot, \cdot) \) can allow us to implicitly work in a ”rich” Hilbert space \( \mathcal{H} \), even though it can itself be parametrized by only few free parameters!
Driven dynamical system with fading memory

Input stream: \( \ldots u(t - 2), u(t - 1), u(t); \quad u(i) \in \mathbb{R} \)

State space: \( S \subset \mathbb{R}^N \)

(Linear) Echo State Network:

\[
\begin{align*}
x(t) &= f(x(t - 1), u(t)) \\
&= W x(t - 1) + u(t) w
\end{align*}
\]

\( W \in \mathbb{R}^{N \times N} \) is a weight matrix providing the dynamical coupling, \( w \in \mathbb{R}^N \) is the input-to-state coupling vector.

Contractive dynamics (ensure ESP): \( \nu = \sigma_{\text{max}}(W) < 1 \)
Interpret the state space model as a "dynamical kernel machine":

\[
\begin{align*}
\phi(...u(t-3), u(t-2), u(t-1)) &= x(t-1) \\
\phi(...u(t-3), u(t-2), u(t-1), u(t)) &= x(t) \\
&= f(x(t-1), u(t))
\end{align*}
\]

\[
\begin{align*}
y(t) &= h(x(t)) \\
&= \sum_i \alpha_i \cdot \langle x(t_i), x(t) \rangle + b \\
&= \sum_i \alpha_i \cdot \langle \phi(..., u(t_i - 1), u(t_i)), \phi(..., u(t - 1), u(t)) \rangle + b \\
&= \sum_i \alpha_i \cdot K(...u(t_i - 1)u(t_i), ...u(t - 1)u(t)) + b
\end{align*}
\]
"Support times" and "support time series"

\[
y(t) = \sum_{i} \alpha_i \cdot \langle \phi(..., u(t_i - 1), u(t_i)), \phi(..., u(t - 1), u(t)) \rangle + b
\]
... $u(-2), u(-1), u(0)$, $u(-j) \in \mathbb{R}$, $j \in \mathbb{N}_0$, 

Given a past horizon $\tau \gg 1$, we will represent 

$$u(-\tau + 1), u(-\tau + 2), ..., u(-1), u(0)$$

as a vector 

$$u(\tau) = (u(0), u(-1), ..., u(-\tau + 1))^\top$$

$$= (u_1, u_2, ..., u_\tau)^\top \in \mathbb{R}^\tau$$

$u_i = u(-i + 1)$, $i = 1, 2, ... \tau$. 
Setting the scene

The state $\mathbf{x}(0)$ reached from the initial condition $\mathbf{x}(-\tau)$ after seeing $\mathbf{u}(\tau)$ codes for information content in $\mathbf{u}(\tau)$ and will be considered the “feature space representation” of $\mathbf{u}(\tau)$:

$$
\phi(\mathbf{u}(\tau); \; \mathbf{x}(-\tau)) = \mathbf{x}(0)
$$

$$
= \mathbf{W}^T \mathbf{x}(-\tau) + \sum_{i=1}^{\tau} u_i \; \mathbf{W}^{i-1} \mathbf{w}.
$$
Temporal kernel

Given two time series at past horizon $\tau$,

$$u(\tau) = (u_1, u_2, \ldots, u_\tau)^\top, \quad v(\tau) = (v_1, v_2, \ldots, v_\tau)^\top$$

$$K(u(\tau), v(\tau); x(-\tau)) = \langle \phi(u(\tau); x(-\tau)), \phi(v(\tau); x(-\tau)) \rangle.$$

As expected, given the contractive nature of the dynamics, for sufficiently long past time horizons $\tau \gg 1$, kernel evaluation is insensitive to the initial condition $x(-\tau)$.

Simplify presentation by setting $x(-\tau)$ to the origin.
**Theorem**

**Time series over a bounded domain** \([-U, U]\), \(\nu = \sigma_{\text{max}}(W) < 1\), \(\|w\| \leq B\), \(\|x(-\tau)\| \leq A(\tau) = c \cdot \zeta^{-\tau}\), where \(\nu < \zeta < 1\) and \(c\) is a large enough positive const. depending on \(B, U, \nu, \zeta\). Then, for any \(u(\tau), v(\tau) \in [-U, U]^{\tau}\), it holds

\[
K(u(\tau), v(\tau); x(-\tau)) = K(u(\tau), v(\tau); 0) + \epsilon,
\]

where

\[
-\eta^T \left[ \frac{2c}{1-\nu} \cdot B \cdot U \right] \leq \epsilon \leq \eta^T \left[ c^2 \eta^T + \frac{2c}{1-\nu} \cdot B \cdot U \right],
\]

with \(\eta = \nu/\zeta < 1\).
The kernel and “metric tensor”

\[ K(u, v) = u^\top Q v = \langle u, v \rangle_Q, \]

where \( Q \) is a symmetric, positive semi-definite \( \tau \times \tau \) matrix of rank \( N_m = \text{rank}(Q) \leq N \) and elements

\[ Q_{i,j} = w^\top \left( W^\top \right)^{i-1} W^{j-1} w, \quad i, j = 1, 2, \ldots, \tau. \]

With increasing time indices \( i, j = 1, 2, \ldots, \tau \),

\[ |Q_{i,j}| < \nu^{i+j-2} \|w\|_2^2. \]

Note: \( Q = C^\top C \), where \( C \) is the controllability matrix
\( C = [w, Ww, W^2w, \ldots, W^{\tau-1}w] \).
Temporal kernel

Understanding the kernel

Eigen-analysis of the “metric tensor” \( Q \in \mathbb{R}^{\tau \times \tau} \) (SVD of \( C \))

Rotate axes - orthogonal \( M \in \mathbb{R}^{\tau \times \tau} \)

\[
K(u, v) = \langle u, v \rangle_Q \\
= u^\top M\Lambda M^\top v \\
= \langle \Lambda^{\frac{1}{2}} M^\top u, \Lambda^{\frac{1}{2}} M^\top v \rangle \\
= \langle \tilde{u}, \tilde{v} \rangle
\]
Semi-inner product

What if $Q$ is not full rank, i.e. positive semi-definite?

$\langle \cdot, \cdot \rangle_Q$ is a semi-inner product on $\mathbb{R}^d$.

$u \in \ker(Q)$ from kernel of the linear operator $Q$ have zero length.

Strictly speaking, $\langle \cdot, \cdot \rangle_Q$ is inner product in the quotient of $\mathbb{R}^d$ by $\ker(Q)$, $\mathbb{R}^d/\ker(Q)$ (image of $Q$).

Since this distinction is not crucial for our argumentation, slightly abusing mathematical terminology, we will still refer to $K(\cdot, \cdot)$ as a kernel and to $Q$ as the associated metric tensor.
Kernel motifs and motif weights

Eigenvectors $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_\tau \in \mathbb{R}^\tau$ of $Q$ (columns of $M$), the corresponding real non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_\tau$ arranged on the diagonal of the diag matrix $\Lambda$.

Motifs - $N_m \leq N \leq \tau$ eigenvectors $\mathbf{m}_i$ with positive eigenvalue $\lambda_i > 0$.

$K(\cdot, \cdot)$ acts as semi-inner product on $\mathbb{R}^\tau$ and as inner product on $\text{span}\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{N_m}\}$.
Random dynamic coupling $\tilde{W}$, zero-mean i.i.d. entries

$\tilde{W}_{i,j}, i, j = 1, 2, ..., N$, generated i.i.d. from a zero-mean distribution with variance $\sigma_0^2 > 0$ and finite fourth moment. Re-scale to the desired $\nu \in (0, 1)$:

$$W = \frac{\nu}{\sigma_{\text{max}}(\tilde{W})} \tilde{W},$$

For $N \gg 1$, the largest eigenvalue of $N^{-1} \tilde{W}^\top \tilde{W}$ converges to $4\sigma_0^2$ a.s.

Rescaling

$$W = \frac{\nu}{2\sqrt{N}\sigma_0} \tilde{W}$$

can be thought of as generating $W_{i,j}$ i.i.d. from a zero-mean distribution with standard deviation

$$\sigma = \frac{\nu}{2\sqrt{N}}.$$
Random dynamic coupling $\mathbf{W}$, zero-mean i.i.d. entries

$$Q \approx \|\mathbf{w}\|_2^2 \ \text{diag} \left( 1, \left(\frac{\nu}{2}\right)^2, \left(\frac{\nu}{2}\right)^4, \ldots, \left(\frac{\nu}{2}\right)^{2(\tau-1)} \right).$$

Eigenvectors form the standard basis $\{\mathbf{e}_i\}$ with eigenvalues

$$\hat{\lambda}_i = \|\mathbf{w}\|_2^2 \left(\frac{\nu}{2}\right)^{2(i-1)}.$$
$K$ implements weighted correlation

The temporal kernel has a **rigid Markovian flavor with shallow memory**:

$$K(u, v) = \sum_{i=1}^{N} \lambda_i \langle m_i, u \rangle \langle m_i, v \rangle \approx \sum_{i=1}^{N} \hat{\lambda}_i \langle e_i, u \rangle \langle e_i, v \rangle$$

$$= \|w\|^2 \sum_{i=1}^{N} \left( \frac{\nu}{2} \right)^{2(i-1)} u_i v_i,$$

compares the corresponding recent entries of the time series and weights down comparisons of past elements with rapidly decaying weights.
Illustrative example

100-dimensional state space ($N = 100$)

100 realizations of $\tilde{W}$, $W_{i,j}$ randomly distributed according to $N(0, 1)$.

Each $\tilde{W}$ renormalized to $W$ with $\nu = 0.995$.

Input coupling $w$ - random vector with elements generated i.i.d. according to $N(0, 1)$ and then renormalized to unit vector.

Past horizon $\tau = 200$

Calculated the metric tensor $Q$, as well as its approximation $\hat{Q}$
Illustrative example

Show true motifs \( \mathbf{m}_i \) (eigenvectors of \( \mathbf{Q} \)) for the first four dominant motifs as the mean and std dev across 100 realizations.

For clarity, only show the first 10 dimensions.

Also present the corresponding eigenvalues
- solid bars - means of the actual eigenvalues \( \lambda_i \)
- std dev

Theoretically predicted values - red line.
Random dynamic coupling $W$

Eigenvectors - motifs

motif 1

motif 2

motif 3

motif 4
Random dynamic coupling $W$

Eigenvalues - (squared) motif weights

![Graph showing eigenvalues of Q]
Symmetric coupling $\mathbf{W}$ of rank $N_k \leq N$. Let $s_1, s_2, \ldots, s_{N_k}$ be the eigenvectors of $\mathbf{W}$ corresponding to non-zero eigenvalues $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{N_k}$. Denote by $\tilde{\mathbf{w}}_a = \mathbf{s}_a^\top \mathbf{w}$ the projection of the input coupling $\mathbf{w}$ onto the eigenvector $\mathbf{s}_a$. Then,

$$K(\cdot, \cdot) = \sum_{a=1}^{N_k} \tilde{\mathbf{w}}_a^2 K^{(a)}(\cdot, \cdot),$$

each kernel $K^{(a)}$ with a single motif

$$\mathbf{m}^{(a)} = (1, \sigma_a, \sigma_a^2, \ldots, \sigma_a^{\tau-1})^\top \in \mathbb{R}^\tau.$$
Ring topology of $W$ - SCR

\[ u(t) \xrightarrow{\rho} x(t) \xrightarrow{\rho} y(t) \]
Ring topology of $\mathbf{W}$ - SCR

Cyclic permutation matrix $\mathbf{P}_{N \times N}$

\[
\{1 \rightarrow 2, 2 \rightarrow 3, \ldots, N - 1 \rightarrow N, N \rightarrow 1\}, \text{ represented by } P_{i+1,i} = 1, i = 1, 2, \ldots, N - 1 \text{ and } P_{1,N} = 1, \text{ all the other elements of } \mathbf{P} \text{ are zero.}
\]

Rescale to the desired $\sigma_{\text{max}}$: $\mathbf{W} = \nu \cdot \mathbf{P}$

Time horizon $\tau = \ell N$, for some positive integer $\ell > 1$.

Motifs have an intricate block structure.
Theorem

\( \mathbf{W} = \nu \cdot \mathbf{P}_{N \times N}, \quad \nu \in (0, 1) \). Let \( \tilde{m}_i \in \mathbb{R}^N, \ i = 1, 2, \ldots, N \), be motifs of the temporal kernel under past time horizon \( N \), with motif weights \( \tilde{\omega}_i \). Then, given a different past time horizon \( \tau = \ell \cdot N \), for some positive integer \( \ell > 1 \), the motifs \( \mathbf{m}_i \in \mathbb{R}^\tau \) have the following block form:

\[
\mathbf{m}_i = \left( \tilde{m}_i^\top, \nu^N \tilde{m}_i^\top, \nu^{2N} \tilde{m}_i^\top, \ldots, \nu^{(\ell-1)N} \tilde{m}_i^\top \right)^\top, \quad i = 1, 2, \ldots N.
\]

The corresponding motif weights are equal to

\[
\omega_i = \tilde{\omega}_i \left( \frac{1 - \nu^{2\tau}}{1 - \nu^{2N}} \right)^{\frac{1}{2}}.
\]
Empirically, when $W$ is a scaled permutation matrix, a very simple setting of input coupling $w$ is sufficient:

All elements of $w$ can have the same absolute value, but the sign pattern should be aperiodic.

Intuitively, it is clear that for such $W$ a periodic input coupling $w$ will induce symmetry in the dynamic processing and such a symmetry should be broken.

Exactly what representational capabilities are lost by imposing a periodicity in $w$?

Start by considering a periodic $w \in \mathbb{R}^N$ formed by $k > 1$ copies of a periodic block $s \in \mathbb{R}^p$, $w = (s^\top, s^\top, \ldots, s^\top)^\top$. 
Minimal setting of input coupling \( \mathbf{w} \)

Periodic \( \mathbf{w} \in \mathbb{R}^N \) formed by \( k > 1 \) copies of a periodic block \( \mathbf{s} \in \mathbb{R}^p \),
\[
\mathbf{w} = (\mathbf{s}^T, \mathbf{s}^T, \ldots, \mathbf{s}^T)^T.
\]

\( \mathbf{P} \in \mathbb{R}^{p \times p} \) is the right shift permutation matrix operating on vectors from \( \mathbb{R}^p \).

Matrix \( \mathbf{T} \in \mathbb{R}^{p \times p} \) with elements
\[
T_{i,j} = \rho^{i+j-2} \langle \mathbf{s}, \mathbf{P}^{j-i} \mathbf{s} \rangle, \quad i, j = 1, 2, \ldots, p. \tag{1}
\]
**Theorem**

\[ W = \nu \cdot P_{N \times N}, \; \nu \in (0, 1). \]

Let the input coupling \( w \in \mathbb{R}^N \) consist of \( k > 1 \) copies of a periodic block \( s \in \mathbb{R}^p \). Denote by \( m_i \in \mathbb{R}^p \), \( i = 1, 2, \ldots, p \), eigenvectors of \( T \) with the corresponding eigenvalues \( \lambda_i \).

Then, given a past time horizon \( \tau = \ell \cdot N \), there are at most \( p \) temporal kernel motifs \( m_i \in \mathbb{R}^\tau \) of non-zero motif weight. Furthermore, the kernel motifs have the following block form,

\[ m_i = (\overline{m}^\top_i, \nu^p \overline{m}^\top_i, \nu^{2p} \overline{m}^\top_i, \ldots, \nu^{\tau-p} \overline{m}^\top_i)^\top, \quad i = 1, 2, \ldots p, \]

with the corresponding motif weights

\[ \omega_i = \left( \frac{\lambda_i}{1 - \nu^{2\tau}} \right)^{\frac{1}{2}}. \]
Illustrative examples

Illustrative examples – setting

Influence of the dynamic and input coupling, \( W \) and \( w \), respectively, on the strength and richness of motifs of the temporal kernel.

State space dimensionality \( N = 100 \).

Re-normalize \( W \in \mathbb{R}^{100 \times 100} \) to spectral radius \( \rho = 0.995 \).

Input coupling \( w \) is renormalized to unit length.

Show motifs with motif weights up to \( 10^{-2} \) of the highest motif weight.
Illustrative examples

Random $\mathbf{W}$. $W_{i,j}, w_j \sim \mathcal{N}(0, 1)$, i.i.d.

Almost identical results for other distributions for $\mathbf{W}, \mathbf{w}$ and settings of $\mathbf{w}$. 

P. Tino
From Dynamical Systems to Kernel Based Fea:
Random symmetric Wigner $W$, random $w$

The number and nature of the motifs stayed unchanged across a variety of generative mechanisms for $W$ and $w$. 

P. Tino
From Dynamical Systems to Kernel Based Feature Spaces and Back
Illustrative examples

SCR, random \( w \)

**motifs**

<table>
<thead>
<tr>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>210</td>
<td>220</td>
<td>230</td>
<td>240</td>
</tr>
</tbody>
</table>

**motif weights**

<table>
<thead>
<tr>
<th>0</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

P. Tino
From Dynamical Systems to Kernel Based Fe:
Illustrative examples

$\text{SCR, } \mathbf{w} \in \{-1, +1\}^N$, signs follow binary expansion of $\pi$

motifs

motif weights
SCR, $w \in \{-1, +1\}^N$, signs follow binary expansion of $e$
Illustrative examples

$\text{SCR, } w \in \{-1, +1\}^N$, periodic, $p = 10$
Illustrative examples

Thank you! Interested?


