Dynamics of Oja's rule:

$$\vec{w}_{t+1} = \vec{w}_t + \eta \left(y_t \vec{x}_t - y_t^2 \vec{w}_t \right), \quad \text{with } y_t = \vec{w}_t \cdot \vec{x}_t$$

An observation:

$$\vec{w}_{t+1}^{\,2} = \vec{w}_t^{\,2} + \underbrace{2\eta\, \vec{w}_t \cdot \left(y_t \vec{x}_t - y_t^2 \vec{w}_t \right)}_{2\eta\, y_t^2 \left(1 - \vec{w}_t^{\,2}\right)} + O(\eta^2)$$

Hence, if $\vec{w}_0 = \vec{0}$, we expect $|\vec{w}_t|$ to grow and reach one for large t.

Alternative rule with strict normalization enforcement:

- $\vec{w}_0$ ~~such~~ that $|\vec{w}_0| = 1$
- $$\vec{w}_{t+1} = \frac{\vec{w}_t + \eta\, y_t \cdot \vec{x}_t}{\left|\vec{w}_t + \eta\, y_t\, \vec{x}_t\right|}$$

$$= \frac{\vec{w}_t + \eta\, y_t\, \vec{x}_t}{|\vec{w}_t|} \left(1 + 2\eta \frac{y_t^2}{|\vec{w}_t|^2} + O(\eta^2) \right)^{-1/2}$$

$$\underset{|\vec{w}_t| = 1 = |\vec{w}_0|}{=} \vec{w}_t + \eta \left(y_t \vec{x}_t - y_t^2 \vec{w}_t \right) \quad : \text{Oja's rule}$$

Going to the diagonal basis of: $C_{ij} = \langle x_i x_j \rangle_{\vec{x}}$

$$C = \underbrace{O}_{\text{orthogonal matrix}} \Lambda O^T, \quad \text{with } \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N), \quad \lambda_1 \geq \dots \geq \lambda_N$$

Change of variable: $\begin{cases} \vec{w} \to O^T \vec{w} = \vec{w}' \\ \vec{x} \to O^T \vec{x} = \vec{x}' \end{cases}, \quad y \stackrel{!}{=} \vec{w}' \cdot \vec{x}' = y$

Then, we clearly have by linearity:

$$\vec{w}'_{t+1} = \vec{w}'_t + \eta\left(y'_t\,\vec{x}'_t - (y'_t)^2\,\vec{w}'_t\right)$$

with $\langle x'_i x'_j\rangle_{\vec{x}} = \lambda_i \delta_{ij}$

I'll drop the "′" in the following.

Evolution of $\vec{w}_t$:

I assume $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_t, \dots$ to be i.i.d from distribution with zero mean and ~~identity~~ covariance matrix Λ

Then $\vec{w}_t$ obeys a Markovian dynamics:

- $$\left\langle \vec{w}_{t+1} - \vec{w}_t \,\middle|\, \vec{w}_t = \vec{w}\right\rangle_{\vec{x}} = \eta\left\langle (\vec{w}\cdot\vec{x})\,\vec{x} - (\vec{w}\cdot\vec{x})^2\,\vec{w}\right\rangle_{\vec{x}}$$

$$\left\langle (\;-\;-\;)_i \,\middle|\, -\;-\right\rangle_{\vec{x}} = \eta\left(\lambda_i w_i - \left(\sum_k \lambda_k w_k^2\right) w_i\right)$$

- $$\left|\vec{w}_{t+1} - \vec{w}_t\right|^2 = \eta^2\left|y_t\,\vec{x}_t - y_t^2\,\vec{w}_t\right|^2$$

$$= \eta^2 y_t^2\left(\vec{x}_t^{\,2} - 2y_t \underbrace{\vec{x}_t\cdot\vec{w}_t}_{y_t} + y_t^2\,\vec{w}_t^{\,2}\right)$$

$$\underbrace{y_t^2\left(\vec{w}_t^{\,2} - 2\right)}_{\leq 0}$$

$$\leq \eta^2\left(\vec{x}_t^{\,2}\right)^2\underbrace{\left(\vec{w}_t\right)^2}_{\leq 1}$$

assumed to be bounded from above by X_4

$$\Rightarrow \mathrm{var}\left(\vec{w}_{t+1} - \vec{w}_t \,\middle|\, \vec{w}_t = \vec{w}\right) \leq O(\eta^2)$$

I assume $\begin{cases} \eta \to 0 \\ t \to \infty \end{cases}$ with $\boxed{\tau = t\eta}$ fixed

Consider evolution over time interval $[\tau, \tau+\varepsilon]$, i.e. from $t = \frac{\tau}{\eta}$ to $t' = \frac{\tau}{\eta} + \frac{\varepsilon}{\eta}$ (ε small, but first $\eta \to 0$)

Li, Wang, Zhang, NIPS2017

$$\vec{w}_{t'} - \vec{w}_t = \sum_{t_0=t}^{t'-1} \left(\vec{w}_{t_0+1} - \vec{w}_{t_0}\right)$$

$$= \sum_{t_0=t}^{t'-1} \left\{ \underbrace{\left\langle \vec{w}_{t_0+1} - \vec{w}_{t_0} \middle| \vec{w}_{t_0} \right\rangle_{\vec{x}_{t_0}}}_{\text{order } \eta} + \vec{Z}_{t_0} \right\}$$

$\frac{\varepsilon}{\eta}$ terms

average $= 0$

$\langle |\vec{Z}|^2 \rangle = O(\eta^2)$

$\boxed{1}$

$\mathrm{var}\left(\sum_{t_0} \vec{Z}_{t_0}\right) \sim O\left(\frac{\varepsilon}{\eta} \cdot \eta^2\right) \xrightarrow[\eta \to 0]{} 0$

hence, negligible: CLT

$\boxed{2}$

$\Downarrow$

$O(\varepsilon)$ small for small ε

Suggests that: $\vec{w}_t = \vec{W}(\tau = t\eta)$

i.e. $\vec{W}$ varies on slow time scale

hence

$\left\langle \vec{w}_{t_0+1} - \vec{w}_{t_0} \middle| \vec{w}_{t_0} \right\rangle$ does not vary by more than ε over $t_0 \in (t, t']$

$\tau_0 \in [\tau, \tau+\varepsilon]$

Finally: (after division by ε)

$$\frac{d W_i}{d\tau} = \lambda_i W_i - \left(\sum_k \lambda_k W_k^2\right) W_i$$

- Easy to check that, if $|\vec{w}(\tau=0)|=1$, then $|\vec{w}(\tau)|=1\ \forall \tau$.
- Solution:

$$w_i(\tau) = \frac{w_i(0)\, e^{\lambda_i \tau}}{\left(\sum_k e^{2\lambda_k \tau}\, w_k(0)^2\right)^{1/2}}$$

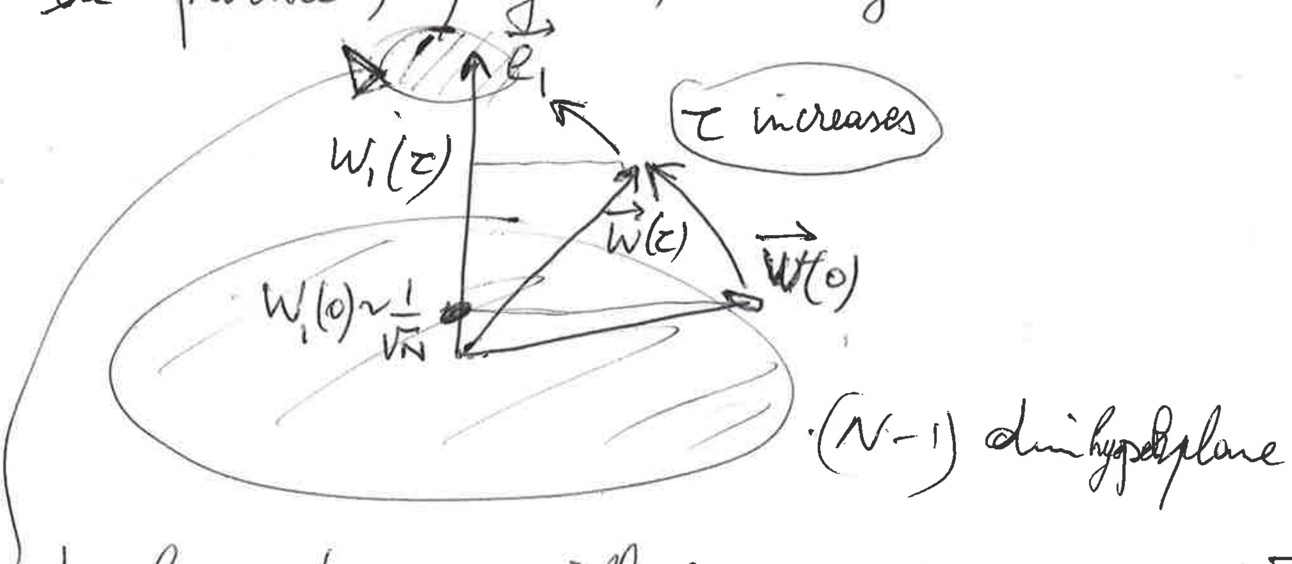
$$\xrightarrow[\tau\to\infty]{} \delta_{i1} \quad : \quad \text{PCA !}$$

$(\lambda_1 > \lambda_2 \geq \dots)$ (top component only so far)

Some remarks:

1. Here, $\eta \to 0$ & $t \to \infty$ with $t\eta = O(1)$

   In practice, η fixed, t large



at large t, we will have fluctuations around $\vec{e}_1$, unless $\eta \to 0$, e.g. such as $\frac{1}{t}$ $\left(\eta t \xrightarrow[t\to\infty]{} \infty\right)$

2. Dependence on N?
   - escape from $\perp$ hyperplane;
     $w_1(0) = \delta$ small requires $t \sim \dfrac{\log(\delta\sqrt{N})}{\eta(\lambda_1 - \lambda_2)}$ steps

     $\left(\simeq \frac{1}{\sqrt{N}}\, e^{2(\lambda_1-\lambda_2)\tau}\right)$

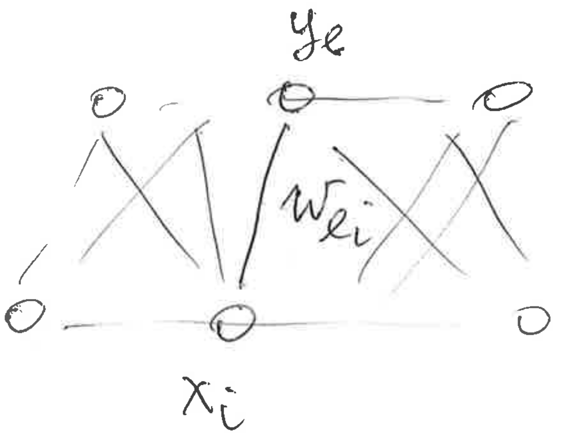
- in addition gap $\lambda_1 - \lambda_2$ may close as $N \to \infty$.

Theory here is in asymptotic setting: $\left\{ \begin{array}{l} N \text{ fixed} \\ \text{many data} \ldots \end{array} \right\}$

not high-dimensional setting, i.e. $N \sim t$

3. Extension to extraction of K top components:

(Sanger 1989)



$\ell = 1 \ldots K$

$i = 1 \ldots N$

$$\Delta w_{\ell i} = \eta \, y_\ell \left(x_i - \sum_{k=1}^{\ell} y_k w_{ki} \right)$$

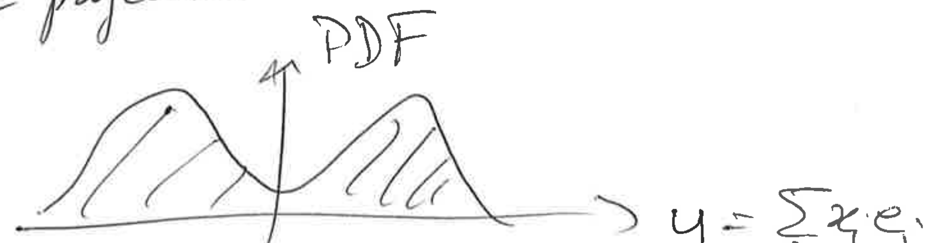
4. Other analysis, non quadratic:

$$\text{PCA} \iff \max \left\langle \left(\sum_i e_i x_i \right)^2 \right\rangle_{\vec{x}}, \text{ with } \sum_i e_i^2 = 1$$

Non quadratic extensions

$$V\left(\sum_i e_i x_i \right), \text{ with } V(y) \neq y^2$$

Example: $V(y) = y^4$: clustering, looks for large projections in absolute values



Modified Oja's rule:

$$\vec{w}_{t+1} - \vec{w}_t = \eta \, A(y_t) \left(\vec{x}_t - y_t \vec{w}_t \right)$$

with $A(y) = \dfrac{dV}{dy}$.