Adaptive lineare Transformationen AS2-3

PCA-Transformation

PCA-Netze und Weissen

ICA-Transformation

Hebb'sches Lernen

$\Delta \mathbf{w} = w_i(t) - w_i(t-1) = \gamma_i(t)\, y_i x$ *Iterative Hebb'sche Lernregel*

$\Delta \mathbf{W} = \mathbf{W}(t) - \mathbf{W}(t-1) = \gamma(t)\, \mathbf{y}\mathbf{x}^T$

$\mathbf{W} = \mathbf{W(1)} + \mathbf{W(2)} + \mathbf{W(3)} + \ldots$

Problem: ex. kein „Vergessen", $\mathbf{w} \to \infty$

Unendliches Wachstum ??

Hebb'sches Lernen - Ergänzungen

UNIVERSITÄT FRANKFURT AM MAIN

Lösung 1: lin. Term, Abklingen der Synapsen

$\mathbf{w}(t) = \mathbf{w}(t\text{-}1) + \gamma(t)\, y\mathbf{x}$ *Iterative Hebb'sche Lernregel*

$\mathbf{w}(t) = \mathbf{w}(t\text{-}1) + \gamma(t)\,(y\mathbf{x} - \mathbf{w}(t\text{-}1))$ *Abklingen des Gewichts*

$\gamma^{-1}\frac{\partial}{\partial t}\mathbf{w} \approx \mathbf{w}(t)\text{-}\mathbf{w}(t\text{-}1) = y\mathbf{x} - \mathbf{w}(t\text{-}1)$ Diff.gleichung mit $\tau = 1/\gamma$

Erwartetes Ziel bei lin. System $y = \mathbf{w}^T\mathbf{x}$ $\quad \mathbf{C}_{xx} := \langle \mathbf{x}\mathbf{x}^T \rangle$

$$\frac{\partial}{\partial t}\mathbf{w} = \gamma_1\langle \mathbf{x}y \rangle - \gamma_2\,\mathbf{w} = \gamma_1\langle \mathbf{x}\mathbf{x}^T\mathbf{w} \rangle - \mathbf{w} = \gamma_1\langle \mathbf{x}\mathbf{x}^T \rangle\,\mathbf{w} - \gamma_2\mathbf{w} = \mathbf{0}$$

$\Leftrightarrow \mathbf{C}_{xx}\mathbf{w} = \lambda\mathbf{w}$ bei Fixpunkt $\mathbf{w} = \mathbf{w}^*$Eigenvektor von $\mathbf{C}_{xx}$

$\mathbf{w}^*$ stabil? 1-dim Beispiel: Nein!

$\frac{\partial}{\partial t}\mathbf{w} = \gamma_1 \mathbf{x}y - \gamma_2\,\mathbf{w}^n$ nicht-lin. Abklingterm $n>1$

Hebb'sches Lernen - Ergänzungen

UNIVERSITÄT FRANKFURT AM MAIN

Lösung : Normierung der Gewichte

$\mathbf{w}(t) = \mathbf{w}(t\text{-}1) + \gamma(t)\, y\mathbf{x}$ mit $|\mathbf{w}(t)| = 1$

Wie?

$$\hat{\mathbf{w}}(t) = \mathbf{w}(t\text{-}1) + \gamma(t)\, y\mathbf{x}$$

$$\mathbf{w}(t) = \frac{\hat{\mathbf{w}}}{|\hat{\mathbf{w}}|} = \frac{\mathbf{w}(t\text{-}1) + \gamma(t)\, y\mathbf{x}}{|\hat{\mathbf{w}}|}$$

Wohin konvergiert $\mathbf{w}(t)$?

Lernen: beschränkte Hebbsche Regel

UNIVERSITÄT FRANKFURT AM MAIN

Konvergenzziel?

$w_i(t) = w_i(t\text{-}1) + \gamma_i(t)\,\langle y_i x \rangle$ Hebb'sche Lernregel mit NB $|\mathbf{w}| = \text{const.}$

$= w_i(t\text{-}1) + \gamma_i(t)\,\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}}$ Gradientenaufstieg

Also:

$\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} = \langle \mathbf{x}y \rangle = \langle \mathbf{x}\mathbf{x}^T \rangle\,\mathbf{w} = \mathbf{A}\mathbf{w}$ bei $y = \mathbf{x}^T\mathbf{w}$ *lin. Neuron*

Zielfunktion $R(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{A}\mathbf{w}$ mit $|\mathbf{w}| = \text{const} = 1$

Extremwert von $L(\mathbf{w},\mu) = R(\mathbf{w}) + \mu\cdot(|\mathbf{w}|^2 - 1)$ Lagrangefunktion

bei $\frac{\partial L}{\partial \mathbf{w}}(\mathbf{w},\mu) = \frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} + \mu 2\mathbf{w} = \mathbf{A}\mathbf{w} + \mu 2\mathbf{w} = 0$

$\mathbf{A}\mathbf{w} = \lambda\mathbf{w}$ Eigenwertgleichung mit $\lambda := -2\mu$

$\mathbf{w} \rightarrow EV(\mathbf{A})$ mit EW $\lambda = \max$

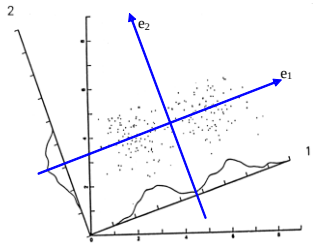
Principal Component Analysis PCA

UNIVERSITÄT FRANKFURT AM MAIN

Zerlegung in orthogonale Eigenvektoren = Basisvektoren

„Hauptkomponentenanalyse“, „principal component analysis“, „Karhunen-Loéve-Entwicklung“, „Hotelling-Transformation“,...

Eigenvektoren – Wozu?



Merkmals-transformation auf Hauptrichtungen

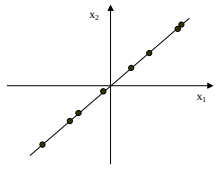
Principal Component Analysis PCA

UNIVERSITÄT FRANKFURT AM MAIN

Transformation auf Unkorreliertheit

$\langle (x_1-\langle x_1\rangle)\cdot(x_2-\langle x_2\rangle)\rangle = 0$ *Unkorrreliertheit* von x_1,x_2

Beispiel



Rauschfrei korrelierte Daten

$\mathbf{x} = (x_1,x_2)$ mit $x_2 = ax_1$

Rechnung: EV, EW = ?

Dekorrelation und Unabhängigkeit

UNIVERSITÄT FRANKFURT AM MAIN

- **DEF dekorreliert**: $\langle (y_i-\langle y_i\rangle)\cdot(y_j-\langle y_j\rangle)\rangle = \langle y_i-\langle y_i\rangle\rangle\langle y_j-\langle y_j\rangle\rangle = 0$ i≠j
- **Satz: PCA dekorreliert Daten**

DEF PCA: $y = \mathbf{w}^T\mathbf{x}$ mit $C_{xx}\mathbf{w} = \lambda\mathbf{w}$ Eigenvektoren

⇒ Mit PCA gilt $y_i'=y_i-\overline{y_i}$, $\mathbf{x}'=\mathbf{x}-\overline{\mathbf{x}}$

$$\langle y_i' y_j'\rangle = \langle \mathbf{w}_i^T\mathbf{x}'\,\mathbf{x}'^T\mathbf{w}_j\rangle = \mathbf{w}_i^T\langle \mathbf{x}'\,\mathbf{x}'^T\rangle\mathbf{w}_j = \mathbf{w}_i^T\mathbf{C}_{xx}\mathbf{w}_j = \mathbf{w}_i^T\lambda_j\mathbf{w}_j$$

$$= \begin{cases}\lambda_i \text{ bei } i=j\\ 0 \text{ bei } i\neq j\end{cases} \quad \text{da ja } \mathbf{w}_i^T\mathbf{w}_j = \begin{cases}1 \text{ bei } i=j\\ 0 \text{ bei } i\neq j\end{cases} \text{ gilt.}$$

- Daten sind **unabhängig** ⇒ Daten sind **dekorreliert**

DEF unabhängig: $P(x_i,x_j) = P(x_i)P(x_j)$ x_i,x_j Zufallsvariable

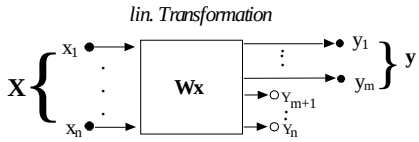
$$\Rightarrow \langle y_i y_j\rangle = \sum_{y_i,y_j} P(y_i,y_j)y_i y_j = \sum_{y_i}\sum_{y_j} P(y_i)P(y_j)y_i y_j = \sum_{y_i} P(y_i)y_i \sum_{y_j} P(y_j)y_j$$

$$= \langle y_i\rangle\langle y_j\rangle = 0 \text{ bei } \langle y_i\rangle = 0 \quad i\neq j$$

Aber: umgekehrt gilt nicht: dekorreliert ist **nicht** unabhängig!

Transformation mit minimalem MSE

Allgemeine Situation



$$\min_{\mathbf{w}} R(\mathbf{W}) = \min \langle (\mathbf{x}-\hat{\mathbf{x}})^2 \rangle$$ *least mean squared error* (LMSE)

Wann minimal ?

Transformation mit minimalem MSE

Minimaler Rekonstruktionsfehler

$$\min_{\mathbf{w}} R(\mathbf{W}) = \min \langle (\mathbf{x}-\hat{\mathbf{x}})^2 \rangle$$ *least mean squared error* (LMSE)

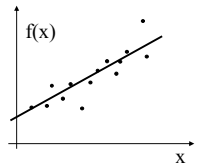
$$\mathbf{x} = \sum_{i=1}^{m} y_i \mathbf{w}_i + \sum_{i=m+1}^{n} y_i \mathbf{w}_i \qquad \hat{\mathbf{x}} = \sum_{i=1}^{m} y_i \mathbf{w}_i + \sum_{i=m+1}^{n} c_i \mathbf{w}_i \qquad y_i = \mathbf{x}^T \mathbf{w}_i$$

- **Was ist die beste Schätzung für die Konstanten c_i?**
  min $R(c_i)$ = ? **Rechnung!** Anhang B

- **Bei welchen Basisvektoren w_i ist der Fehler minimal ?**
  min $R(w_i)$ = ? **Rechnung!** Anhang B

Transformation mit minimalem MSE

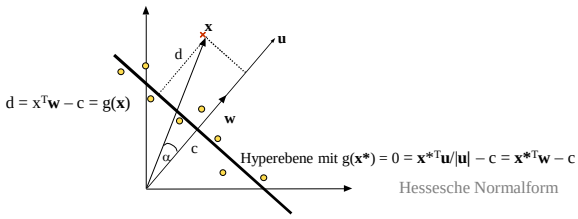
***m* Messungen**



- **Modellierung als Gerade**
  $\mathbf{y = f(x) = y_0 + ax}$
- **Beispiel: Ökonomie**
  Konsum y = f(Einkommen x)
  = Konsumsockel + a ·Einkommen
- **Problem:** Ergebnis hängt vom Koord.system ab

Transformation mit minimalem MSE

- **Minimaler mittlerer Abstand zur Gerade** (Hyperebene)



TLMSE = R(**w**,c) = $\langle d^2 \rangle$

Rechnung: Minimum des TLMSE (Kap.3.3.1)

PCA-Transformation

PCA-Netze und Weissen

ICA-Transformation

PCA Netze für geordnete Zerlegung

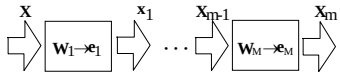
Sanger-Methode Sanger 1988

Vollständige Zerlegung von {x} in *n* Eigenvektoren (Gram-Schmidt)

$\{x\}_0 = \{x\}$, i=1

1. Suche die Richtung größter Varianz in $\{x\}_{i-1}$. Dies ist e_i.
2. Ziehe diese Richtung (Dimension) von $\{x\}_{i-1}$ ab. Wir erhalten $\{x\}_i$.
3. Wenn i<n, setze i := i+1, gehe zu 1.

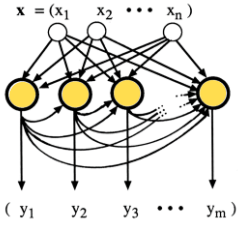
Diskret für stochastisches x, w: Sanger-Netz



PCA Netze durch laterale Inhibition

Asymmetrische Netze

Rubner, Tavan 1990



$$y_i = \mathbf{w}_i^T \mathbf{x} + \sum_{k<i} u_{ik} y_k \qquad \text{Aktivität}$$

$$= (\mathbf{w}_i + \sum_{k<i} u_{ik} \mathbf{w}_k)^T \mathbf{x} =: \tilde{\mathbf{w}}_i^T \mathbf{x}$$

$\Delta \mathbf{w}_i = \gamma\, \mathbf{x} y_i$ Hebb-Lernen

$\Delta u_{ik} = -\mu\, y_i y_k$ Anti-Hebb-Lernen

Anti-Hebb auch aus Prinzip „kleinste gemeins. Information“ $H(y_k, y_l)$

Whitening Filter

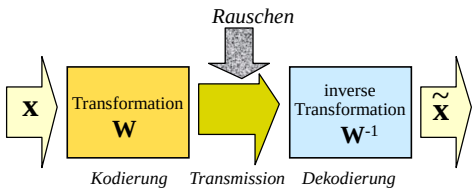
Shannon: Whitening für alle Frequenzen, d.h. alle diskreten Signalbänder

Übertragung auf parallele Signale y_i : gleiche Varianz aller durch Transformation $\mathbf{W}$.

Anhebung zu geringer Amplituden: Wähle $\mathbf{W}$ so,

dass $\langle y_i y_j \rangle = 1$ bei i = j, sonst = 0; also $\langle \mathbf{y}\mathbf{y}^T \rangle = \mathbf{I}$

Absenkung der Amplituden: durch inverse Matrix $\mathbf{W}^{-1}$



Whitening Filter

Anhebung bei parallelen Signalen

Wenn für die Transformation $\mathbf{W}$ eine orthonormale Basis $\mathbf{M}^{-1} = \mathbf{M}^T$ gewählt wird, ist das Ziel des Lernens

mit $\mathbf{I} = \langle \mathbf{y}\mathbf{y}^T \rangle = \langle \mathbf{W}\mathbf{x}\mathbf{x}^T\mathbf{W}^T \rangle = \mathbf{W}\langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{W}^T = \mathbf{W}\mathbf{A}\mathbf{W}^T$

auch $\mathbf{W}^T\mathbf{I} = \mathbf{W}^T(\mathbf{W}\mathbf{A}\mathbf{W}^T) = \mathbf{A}\mathbf{W}^T$

bzw. $\mathbf{w}_k = \mathbf{A}\mathbf{w}_k$ **Eigenvektoren** $\mathbf{w}_k$ von $\mathbf{A}$ mit $\lambda = 1$

Also:

1. Signal zentrieren und PCA durchführen. Wir erhalten orthonormale Eigenvektoren $\mathbf{e}_i$ mit Eigenwerten λ_i.
2. $\mathbf{e}_i$ normieren: $\mathbf{w}_i = \mathbf{e}_i / \lambda_i^{1/2}$ so dass $|\mathbf{w}_i|^2 = 1/\lambda_i$.
   Es ergibt sich orthogonale Transformationsmatrix $\mathbf{W}$ mit Zeilenvektoren $\mathbf{w}_i$

Sie erfüllt $\langle y_i y_i \rangle = \langle \mathbf{w}_i^T \mathbf{x}\mathbf{x}^T \mathbf{w}_i \rangle = \mathbf{w}_i^T \mathbf{A} \mathbf{w}_i = \mathbf{w}_i^T \mathbf{w}_i \lambda_i = 1$

Whitening Filter

UNIVERSITÄT FRANKFURT AM MAIN

Absenkung (Rücktransformation) $W^{-1} = ?$

Wenn $\mathbf{B} = (\lambda_1\mathbf{w}_1,..., \lambda_n\mathbf{w}_n)$ gewählt wird, ist mit $|\mathbf{w}_i|^2 = \lambda_i^{-1}$

mit $\mathbf{W}\cdot\mathbf{B} = \begin{pmatrix} \mathbf{w}_1^T \\ ... \\ \mathbf{w}_n^T \end{pmatrix} \begin{pmatrix} \lambda_1\mathbf{w}_1 & \cdots & \lambda_n\mathbf{w}_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & 1 \end{pmatrix} = \mathbf{I},$

Also ist $\mathbf{W}^{-1} = \mathbf{B}$ mit den Spalten aus den Zeilen von $\mathbf{W}$.

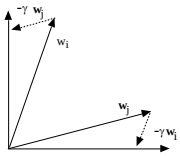
Also: *Rücktransformation*

1. Aus der PCA haben wir $\mathbf{e}_i$, λ_i mit $|\mathbf{e}_i|^2 = 1$ und so die Matrix $\mathbf{W}$ mit $|\mathbf{w}_i|^2 = \lambda_i^{-1}$
2. Basis $\mathbf{b}_i$ bilden aus $\mathbf{W}$: $\mathbf{b}_i = (\lambda_1 w_{1i}, \lambda_2 w_{2i}, ..., \lambda_n w_{ni})$

Orthonormalisierende Netze

UNIVERSITÄT FRANKFURT AM MAIN

Heuristische Methode — Silva, Almeida 1991



Ziel: Projektion $a_{ij} := \mathbf{w}_i^T\mathbf{w}_j$ eines Basisvektors $\mathbf{w}_i$ auf einen anderen $\mathbf{w}_j$ vermindern

$$\mathbf{w}_i(t) = \mathbf{w}_i(t\text{-}1) - \gamma(t)\, a_{ij}\mathbf{w}_j(t\text{-}1)$$
$$= \mathbf{w}_i(t\text{-}1) - \gamma(t)\, (\mathbf{w}_i^T(t\text{-}1)\mathbf{w}_j(t\text{-}1))\, \mathbf{w}_j(t\text{-}1)$$
$$+ \gamma(1-\langle y_i^2\rangle)\, \mathbf{w}_i(t) \quad \textit{Normierung}$$

Ziel: orthogonal im Datenraum: $a_{ij} := \langle y_i y_j\rangle = w$

Alle Einflüsse

$$\mathbf{w}_i(t) = \mathbf{w}_i(t\text{-}1) - \gamma(t) \sum_{j\neq i} \langle y_i y_j\rangle\, \mathbf{w}_j(t\text{-}1) + \gamma(1-\langle y_i^2\rangle)\, \mathbf{w}_i(t)$$

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \gamma \mathbf{C}_{yy}(t)\, \mathbf{W}(t) + \gamma \mathbf{W}(t) \quad \text{Matrixversion}$$

Orthonormalisierende Netze

UNIVERSITÄT FRANKFURT AM MAIN

Konvergenz der Heuristischen Methode — Silva, Almeida 1991

Beh: $C_{yy} \to I$ mit $C_{yy} = \langle \mathbf{y}\mathbf{y}^T\rangle = W\langle \mathbf{x}\mathbf{x}^T\rangle W^T = WC_{xx}W^T$

Bew: Einsetzen der Matrixversion $W(t+1) = W(t) - \gamma C_{yy}(t)\, W(t) + \gamma W(t)$

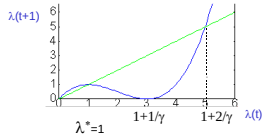
$\Rightarrow \quad C_{yy}(t+1) = \gamma^2 C_{yy}^3 - 2\gamma(1+\gamma)C_{yy}^2 + (1+\gamma)^2 C_{yy}$

Darstellung im Eigenvektorraum

$C_{yy}\, e_i = e_i\lambda_i \Rightarrow C_{yy}E = E\Lambda \Rightarrow C_{yy} = E\Lambda E^T$

ändert die Eigenvektoren nicht, sondern nur die Eigenwerte λ zu

$\lambda_i(t+1) = \gamma^2\lambda^3 - 2\gamma(1+\gamma)\lambda^2 + (1+\gamma)^2\lambda$ *Fixpunktgleichung*

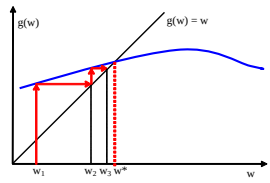


Konvergenz $\lambda \to 1 \Rightarrow C_{yy} \to I$ q.e.d.

Ex. keine Zielfunktion !

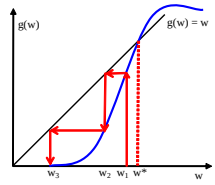
Konvergenz von Fixpunkten

Stabiler Fixpunkt



$|g'(w)| < 1$ bzw. $\left|\det\left(\frac{\partial g_i}{\partial w_j}\right)\right| < 1$

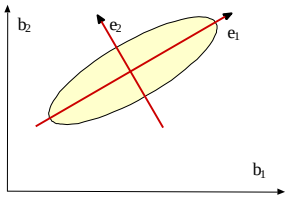
Labiler Fixpunkt



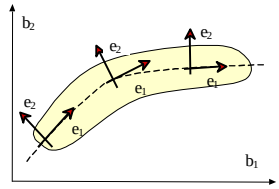
$|g'(w)| > 1$

Ausblick: lineare und nichtlineare PCA

Lineare Hauptachsentransformation



Dynamische Hauptachsentransformation

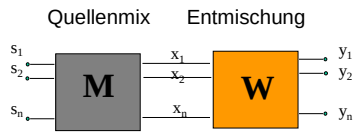


PCA-Transformation

PCA-Netze und Weissen

ICA-Transformation

Lineares ICA-Modell



Ziel: $W \rightarrow M^{-1}$
$y \rightarrow s$

mit $p(\mathbf{y}) = p(y_1,..,y_n) = p(y_1)..p(y_n)$ *unabhängige Kanäle*

Unabhängigkeit notwendig zur Quellentrennung.

Yelling,Weinstein (1994): auch hinreichend!

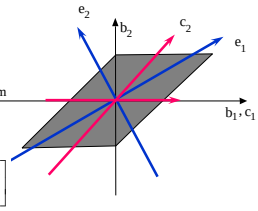
Lineare Koordinatentransformationen

Beispiel:

- PCA-Hauptkomponentenanalyse *Richtung stärkster Varianz*
- ICA- Unabhängigkeitsanalyse *statistische Unabhängigkeit*

Beispiel:

$x = x_1+x_2$,
$y = x_2$
mit x_1,x_2
zufällig uniform
aus [-1,+1]



$c_1 := x-y$, $c_2 := y$
also $c_1 = x_1$ unabh.
von $c_2 = x_2$

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{M}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

ICA-Einschränkungen

- **Quellenzahl = Mischzahl**
  M **muß regulär sein** $\Leftrightarrow$ **nur dann ex.** M^{-1}
- **Ausgabereihenfolge unbestimmt**
  Reihenfolge in $p(y_1)..p(y_n)$ **ist unwichtig** $\Rightarrow M^{-1}$ **bis auf Permutation P bestimmbar:** $M^{-1} \rightarrow M^{-1} P$
- **Unbekannte Skalierung**
  $\Rightarrow \sigma_i := 1$ **Beweis:** Rechnung Kap.3.4
- **2 Gaußsche Quellen lassen sich nicht trennen**
  $\Rightarrow$ **max 1 Gaußsche Quelle** **Beweis:** Rechnung Kap.3.4

DEF Information

$I \sim n = ld(2^n) = ld$ (Zahl der möglichen Daten)

$I \sim ld(1/P)$ [Bit]

DEF $I(X) := \ln(1/P(x^k)) = -\ln(P(x^k))$ *Information*

DEF $H(X) := \sum_k P(x^k) I(x^k) = \langle I(x^k) \rangle_k$ *Entropie*

$$H(X) := \int_{-\infty}^{+\infty} p(x) \ln p(x)^{-1} dx$$ *differenzielle Entropie*

Frage: Wieviel Information hat eine 32-bit floating-point Zahl?

DEF $I(X;Y) = H(X) + H(Y) - H(X,Y)$ *Transinformation mutual information*

ICA - Algorithmen 1a

Ziel: ***minimale Transinformation zwischen den Ausgaben y_i***

x = InputKanäle, stoch. Variable

y = Output

Transinformation $I(y_1;y_2) = H(y_1) + H(y_2) - H(y_1,y_2)$

minimal bei $I(y_1;y_2) = 0$ bzw. maximaler Entropie $H(y_1,y_2) = H(y_1) + H(y_2)$
bzw. $p(y_1,y_2) = p(y_1) \cdot p(y_2)$ stochastische Unabhängigkeit der Variablen y_i

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \frac{\partial}{\partial \mathbf{w}} \gamma I(y_1;y_2;..;y_n)$$ *Gradientenabstieg*

(Amari, Young, Cichocki 1996)

Entwicklung von $p(y_1,y_2,..,y_n)$ in $I(y_1;y_2;..;y_n)$ nach höheren Momenten

$\mathbf{W}(t+1) = \mathbf{W}(t) - (\mathbf{1} - \mathbf{f}(\mathbf{y})\mathbf{y}^T)\mathbf{W}(t)$ mit $f_i(y_i) = \frac{3}{4} y^{11} + \frac{25}{4} y^9 - \frac{14}{3} y^7 - \frac{47}{4} y^5 + \frac{29}{4} y^3$

ICA - Algorithmen 1b

Ziel: ***maximale Transinformation*** (Bell, Sejnowski 1995)
zwischen Eingabe und Ausgabe

Transinformation $I(X;Y) = H(Y) - H(Y|X)$ aus Def.
$H(Y|X)$ ist konstant im determinist. System

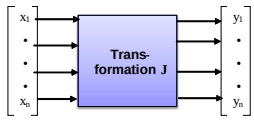
Maximimierung von I(X;Y) durch Maximierung von H(Y)

$$R(\mathbf{w}) := H(Y) = H(X) + \int_{-\infty}^{+\infty} p(\mathbf{x}) \ln |\det \mathbf{J}| d\mathbf{x} \quad \text{mit } \mathbf{J} = \left(\frac{\partial y_i}{\partial x_j} \right)$$

Nicht-lin. Ausgabe y, z.B. $y = \tanh(z)$

Informationstransformation

$H(\mathbf{Y}) = ?$ $\quad H(\mathbf{Y}) = H(\mathbf{Y}(\mathbf{X}))$



Transformation **kontinuierlicher** Zufallsvariabler

$d\mathbf{y} = |\det \mathbf{J}|\, d\mathbf{x}$
$\Rightarrow p(\mathbf{y}(\mathbf{x})) = p(\mathbf{x})\, |\det \mathbf{J}|^{-1}$

$$H(\mathbf{Y}) = -\int_{-\infty}^{+\infty} p(\mathbf{y}) \ln p(\mathbf{y})\, d\mathbf{y}$$

$$= -\int_{-\infty}^{+\infty} (p(\mathbf{x})\, |\det \mathbf{J}|^{-1}) \ln (p(\mathbf{x}) \cdot |\det \mathbf{J}|^{-1})\ |\det \mathbf{J}|\, d\mathbf{x}$$

$$= -\int_{-\infty}^{+\infty} p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} + \int_{-\infty}^{+\infty} p(\mathbf{x}) \ln |\det \mathbf{J}|\, d\mathbf{x}$$

$\mathbf{y}(\mathbf{x}) = \mathbf{W}\mathbf{x}$
$\Rightarrow \mathbf{J} = \mathbf{W}$

$$H(\mathbf{Y}) = H(\mathbf{X}) + \ln |\det(\mathbf{W})|$$

ICA - Algorithmen 1b

Ziel: ***maximale Transinformation*** (Bell, Sejnowski 1995)
zwischen Eingabe und Ausgabe

Gradientenregel

$$\mathbf{W}(t+1) = \mathbf{W}(t) + \gamma\left(\left[\mathbf{W}^T\right]^{-1} - 2\mathbf{y}\mathbf{x}^T\right)$$ Rechnung: 1-dim Fall (Kap.3.4.1)

„Natürl."Gradient Amari 1985

$$\frac{d\mathbf{W}}{dt} = \Delta\mathbf{W} = \gamma(t)\, \frac{\partial R(\mathbf{W})}{\partial \mathbf{W}} \mathbf{W}^T\mathbf{W} = \gamma\left(\mathbf{I} - 2\mathbf{y}\mathbf{z}^T\right)\mathbf{W}$$

Statist. Momente und Kurtosis

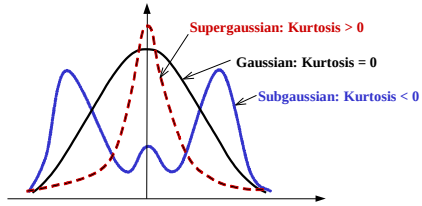
Momente einer Zufallsvariablen $\mathbf{x}$:

$\alpha_i = \langle x \rangle_i$, z.B. $\alpha_1 = \langle x \rangle$ Mittelwert

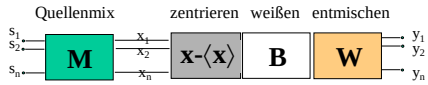
Zentrale Momente einer Zufallsvariablen $\mathbf{x}$:

$m_k = \langle (x-\alpha_1)^k \rangle$, z.B. $m_2 = \langle (x-\alpha_1)^2 \rangle$ Varianz

Wölbungsmaß Kurtosis: $\mathrm{kurt}(x) = [\langle (x-\alpha_1))^4 - 3m_2^2]/m_2^2$



ICA-Algorithmen: Vorverarbeitungsfolge



Zentrieren $\langle \mathbf{x} \rangle = 0 \qquad (\mathbf{x} - \langle \mathbf{x} \rangle)^2 = 1$

- Mittelwertbildung, z.B. iterativ durch $w_0(t+1) = w_0(t) - \gamma\,(w_0 - x)$, $\gamma = 1/t$

Weißen

- PCA durchführen: $\mathbf{w}_i$ Eigenvektoren von $\mathbf{C} = \langle \mathbf{x}\mathbf{x}^T \rangle$ mit $|\mathbf{w}_i| = 1$ und Eigenwerten λ_i
- Gewichtsvektoren $\mathbf{w}_i$ normieren zu $\mathbf{w}_i / \lambda_i^{1/2}$. Dies führt zu $\langle y^2 \rangle = \mathbf{w}_i^T \langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{w}_i = \mathbf{w}_i^T \lambda_i \mathbf{w}_i = 1$

Entmischen

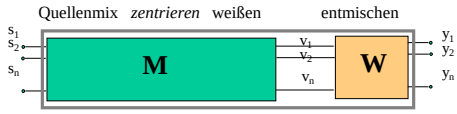
- ICA Algorithmen, z.B. minimale Transinformation, maximale Kurtosis etc.
- *Speziell*: dekorrelierte $\mathbf{x}$ benötigen nur eine orthogonale Matrix $\mathbf{W}$ (Vereinfachung)

ICA – Algorithmen 2

Ziel: extremale Kurtosis (Delfosse, Loubaton 1995)

Extrema bei s_j = **unabh**. Komp, und $z_j = +/-1$

$$\text{kurt}(y) = \text{kurt}(\mathbf{w}^T\mathbf{v}) = \text{kurt}(\mathbf{w}^T\mathbf{M}\mathbf{s}) = \text{kurt}(\mathbf{z}^T\mathbf{s}) = \sum_{j=1}^{n} z_j^4 \,\text{kurt}(s_j)$$



Matrix Z

$(\mathbf{v}_i - \langle \mathbf{v}_i \rangle)^2 = 1 \quad \langle \mathbf{v}_i \rangle = 0$

ICA – Algorithmen 2

Ziel: extremale Kurtosis bei $y = \mathbf{w}^T\mathbf{v}$

$$R(\mathbf{w}) = \langle (\mathbf{w}^T\mathbf{v})^4 \rangle - 3 \langle (\mathbf{w}^T\mathbf{v})^2 \rangle^2 = \min_{\mathbf{w}}$$

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \gamma \,\text{grad}\, R(\mathbf{w})$$

$$= \mathbf{w}(t) + \gamma\, 4 \left(\langle (\mathbf{w}^T\mathbf{v})^3 \mathbf{v} \rangle - 3|\mathbf{w}|^2 \mathbf{w} \right)$$

Bei $|\mathbf{w}| = 1$ ist die Richtung gegeben durch

$$\mathbf{w}(t+1) = \alpha \left(\langle (\mathbf{w}^T\mathbf{v})^3 \mathbf{v} \rangle - \beta \mathbf{w} \right)$$

- **Lernalgorithmus** für einzelnes Neuron (Hyvarinen, Oja 1996)

$$\mathbf{w}(t+1) = \langle (\mathbf{w}^T\mathbf{v})^3 \mathbf{v} \rangle - 3\,\mathbf{w} \qquad \text{Fixpunktalgorithmus}$$

mit $|\mathbf{w}| = 1$

ICA – Algorithmen 2

- **Sequentielle Extraktion aller Komponenten**

  Gegeben: Trainingsmenge $\{\mathbf{v}(0)\}$

  $$\mathbf{w}_1(t+1) = \langle(\mathbf{w}_1{}^T\mathbf{v})^3\mathbf{v}\rangle - 3\,\mathbf{w}_1 \qquad \text{mit } |\mathbf{w}_1| = 1$$

  Konvergenz zum 1. ICA-Vektor.

  Dann neue Trainingsmenge durch $\mathbf{v}(1) = \mathbf{v}(0) - \mathbf{w}_1 y_1$

  $$\mathbf{w}_2(t+1) = \langle(\mathbf{w}_2{}^T\mathbf{v})^3\mathbf{v}\rangle - 3\,\mathbf{w}_2 \qquad \text{mit } |\mathbf{w}_2| = 1$$

  Konvergenz zum 2. ICA-Vektor, usw.

- **Schnellere Konvergenz: Orthogonalisierung**

  $$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) - \sum_{j=1}^{i-1}(\mathbf{w}_i\,\mathbf{w}_j)\,\mathbf{w}_j \quad j < i$$

ICA-Anwendung: Audioanalyse

Mischung | **entmischte Quellen**