# 4 Blind Source Separation: Independent Component Analysis ICA

## 4.1 Einleitung: Oja's ICA-Fixpunktalgorithmus

Wir betrachten das *blind source separation problem* (BSSP). Beim BSSP handelt es sich um ein recht interessantes Problem, bei dem man ohne weiteres nicht vermuten würde, daß es überhaupt eine Lösung besitzt. Die Problemstellung ist wie folgt, siehe Abbildung 4.1:

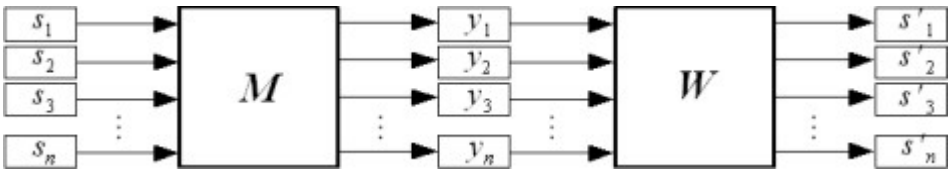


**Abbildung 4.1** Blind Source Separation

Es gibt eine Reihe von Quellen, die unabhängig voneinander n Signale $s_1, ..., s_n$ erzeugen. Diese Signale werden dann bei der Übertragung linear durch eine Matrix **M** (die sogenannte Mischmatrix) miteinander vermischt und es resultieren n Mischsignale $y_1, ..., y_n$ mit $[y_1, ..., y_n]^T = M \cdot [s_1, ..., s_n]^T$. Ziel ist es nun, *allein* aus der Beobachtung der $y_1, ..., y_n$ die ursprünglichen Signale $s_1, ..., s_n$ wiederherzustellen.

Dieses Problem tritt etwa auf, wenn mehrere Signale über eng benachbarte Leitungskabel übertragen werden und sich dadurch elektromagnetisch beeinflussen, bei mehreren Sprechern, deren Stimmen auch die Mikrophone anderer Sprecher erreichen, so daß die Mikrophone lediglich Mischsignale aufzeichnen, oder auch bei Radaraufnahmen und Bildern, die durch reflektierende Medien hindurch (Glas, Nebel, etc.) gemacht werden und deren horizontal und vertikal polarisierte Bildanteile dann aus unterschiedlichen Mischungen des ursprünglichen und des reflektierten Bildsignals bestehen.

Mit der Betonung auf *blind* source separation wird ausgedrückt, daß wir weder Zugriff auf einzelne Datenpunkte der Originalquellen $s_1, ..., s_n$, noch irgendwelche Informationen über die Mischmatrix **M** besitzen. Das Problem kann zwar im Prinzip durch Bestimmung der Inversen Matrix $W = M^{-1}$ gelöst werden, da $W \cdot [y_1, ..., y_n]^T = W \cdot M \cdot [s_1, ..., s_n]^T = M^{-1} \cdot M \cdot [s_1, ..., s_n]^T = [s_1, ..., s_n]^T$, aber da wir weder M, noch vereinzelte s kennen, kommt im Prinzip jede invertierbare Matrix **M** in Frage, die das beobachtete Ergebnis $y_1, ..., y_n$ hätte erzeugen können. Wählen wir beispielsweise **W** als eine *beliebige, invertierbare Matrix,* dann ist mit **M** = $W^{-1}$ für jedes **W** das Ergebnis konsistent mit unserer Beobachtung !

Es ist also gar nicht ohne weiteres klar, ob das Problem überhaupt eine Lösung besitzt. Interessanterweise stellt sich jedoch heraus, daß das Problem lösbar ist, sofern die Signale $s_1, ..., s_n$ stochastisch unabhängig voneinander sind und M invertierbar ist. Eine intuitive Begründung liegt darin, daß die vermischten Signale in diesem Fall Summe der Realisierungen unabhängiger Zufallsvariabler sind, was bedeutet, daß sie sich (aufgrund der Zentralen Grenzwertsätze aus der Stochastik) näher an einer Gaußverteilung befinden, als die Werte der ursprünglichen (unabhängigen) Signale. Der entscheidende Punkt ist nun der, daß **ein einzelner** Zeilenvektor $\underline{w_i^T}$ der Entmisch-Matrix W,

der die "Nicht-Gaußheit" der Komponente $s'_i = \underline{w_i}^T \cdot [y_1, ..., y_n]^T$ maximiert, tatsächlich eine Lösung des Problems darstellt und **bis auf die Skalierung** einem der $s_1, ..., s_n$ entspricht.

Die einzige Ausnahme bilden Gaußquellen selbst. Befinden sich nämlich mehrere Gaußquellen unter den unabhängigen Signalen, dann lassen sich diese leider *nicht* mehr voneinander trennen, da das Mischen zweier unabhängiger Gaußquellen selber wieder zwei unabhängige Gaußquellen ergibt. Ein wichtiger Punkt ist daher, daß höchstens eine Quelle unter den zu trennenden Signalen gaußisch sein darf.

Warum aber ***einem*** der $s_1, ..., s_n$ (und nicht etwa gerade genau $s_i$) und warum ***bis auf die Skalierung***? Betrachten wir uns dazu den Vorgang des Mischens durch die Mischmatrix **M**. Es gilt nun für eine beliebige Permutationsmatrix **P**, die die Reihenfolge der Komponenten eines Vektors ändert (eine Matrix aus Nullen und mit nur einer Eins auf jeder Zeile, aber nicht unbedingt auf der Hauptdiagonalen) und einem skalaren Faktor $a \neq 0$:

$$\begin{aligned}
[y_1, ..., y_n]^T &= M \cdot [s_1, ..., s_n]^T \\
&= M \cdot (P^{-1} \cdot P) \cdot [s_1, ..., s_n]^T \\
&= M \cdot (\tfrac{1}{a} \cdot P^{-1} \cdot P \cdot a) \cdot [s_1, ..., s_n]^T \\
&= (\tfrac{1}{a} \cdot M \cdot P^{-1}) \cdot a \cdot \left(P \cdot [s_1, ..., s_n]^T\right) \\
&= M' \cdot [a \cdot s_{\pi(1)}, ..., a \cdot s_{\pi(n)}]^T
\end{aligned}$$

Dabei bezeichnet $\pi(1), ..., \pi(n)$ die durch **P** vorgenommene Permutation. Wir sehen, dass die $y_1, ..., y_n$ auch durch beliebig skalierte und in ihrer Reihenfolge veränderte $s_1, ..., s_n$ erzeugt werden können, wobei die Mischmatrix **M'** eine andere ist. Daher ist es uns nur möglich, die einzelnen Signale bis auf die Reihenfolge und ihre Skalierung genau zu bestimmen.

nach oben

## 4.2 Signalvorverarbeitung

Üblicherweise führt man zum Trennen der einzelnen Quellen erst noch eine Vorverarbeitungsstufe durch, die die Lösung des Problems, das Finden einer geeigneten Entmischungsmatrix **W**, etwas vereinfacht. Siehe dazu Abbildung 4.2.

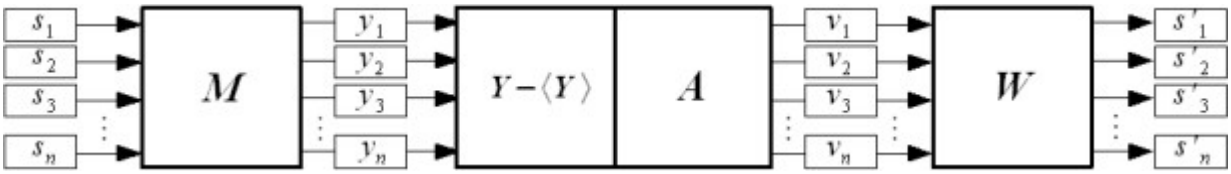


**Abbildung 4.2** Zentrieren und Weißen als Vorverarbeitungsstufe

**a)** Die beobachteten Signale $y_1, ..., y_n$ werden dabei als erstes mittels $y_i' = y_i - \langle y_i \rangle$ auf Erwartungswert 0 zentriert.

> *Dies hat den angenehmen Vorteil, daß wir durch Betrachten der $y_1', ..., y_n'$ anstelle von $y_1, ..., y_n$ nun annehmen dürfen, daß auch die unabhängigen Quellen zuvor auf Erwartungswert 0 zentriert worden sind. Es gilt nämlich, mit $\underline{y} = [y_1, ..., y_n]^T$, $\underline{y'} = [y_1', ..., y_n']^T$ und $\underline{s} = [s_1, ..., s_n]^T$:*

$$\begin{aligned} M^{-1}\underline{y'} &= M^{-1}\underline{y} - \langle M^{-1}\underline{y} \rangle \\ &= \underline{s} - \langle \underline{s} \rangle \end{aligned}$$

**b)** Nach dem Zentrieren werden die $y_1', ..., y_n'$ anschließend mittels einer Matrix **A** durch $[v_1, ..., v_n]^T = A \cdot [y_1', ..., y_n']^T$ dekorreliert und auf die Varianz 1 normiert; es wird ein sogenanntes *whitening* durchgeführt.

*Warum der ganze Aufwand? Nun, das Zentrieren und das Weißen als Vorverarbeitungsstufe haben bestimmte Vorteile, die nachgeschalteten Algorithmen die Lösung des Quellentrennungsproblems etwas erleichtern. Da wir oben gesehen haben, daß wir die Skalierung der Signale nicht ermitteln können, legen wir einfach fest, daß die Entmischungsmatrix **W** so beschaffen sein soll, daß sie uns Ausgabesignale mit Varianz 1 liefert. Zusammen mit dem Optimierungsziel, daß die Signale $s'_1, ..., s'_n$ skalierte und permutierte Versionen der unabhängigen Quellen sind, folgt dann* $\left\langle \underline{s'}\,\underline{s'}^T \right\rangle = I$ *und es gilt mit dekorrelierten Signalen $v_1, ..., v_n$ der Varianz eins*

$$C_{vv} = \left\langle (\underline{v} - \langle \underline{v} \rangle)(\underline{v} - \langle \underline{v} \rangle)^T \right\rangle = \langle \underline{v}\underline{v}^T \rangle = \boldsymbol{I}$$

*und damit für die Quellen $s'_1, ..., s'_n$*

$$\begin{aligned} I &= \left\langle \underline{s'}\,\underline{s'}^T \right\rangle \\ &= \langle W \underline{v}\underline{v}^T W^T \rangle \\ &= W \langle \underline{v}\underline{v}^T \rangle W^T \\ &= W W^T \end{aligned}$$

*Dabei sehen wir, daß $W^T = W^{-1}$ sein muß. Durch den Vorgang des Weißens können nachfolgende Algorithmen sich also auf die Bestimmung einer **orthonormalen** Matrix W beschränken, was wesentlich schneller konvergiert. Dies ist ein entscheidender Vorteil gegenüber der Bestimmung einer beliebigen Matrix.*

Wie erhalten wir nun die dekorrelierten Signale der Varianz eins? Wie erhalten wir die dafür nötige Matrix **A**? Nun, wir erinnern uns, dass die PCA Signale dekorreliert. Wir müssen also nur die Eigenvektoren $\mathbf{e}_i$ der

Autokorrelationsmatrix bestimmen und diese geeignet in der Länge normieren, um zusätzlich die Varianz auf eins zu setzen. Wählen wir als neue Basisvektoren $\mathbf{b}_i$ (Zeilen von **A**) die geeignet skalierten Eigenvektoren

$$\mathbf{b}_i = \mathbf{e}_i / \lambda_i^{1/2}$$

so erreichen wir damit auch für die Signale $\underline{v} := [v_1, ..., v_n]^T$ die Varianz eins.

*Warum? Sei **E** eine Matrix aus den spaltenweise angeordneten normierten Eigenvektoren $\mathbf{e}_i$ der Autokorrelationsmatrix $C_{y'y'} = \left\langle \underline{y'}\,\underline{y'}^T \right\rangle$ und $\Lambda$ eine Diagonalmatrix, bestehend aus den zugehörigen Eigenwerten $l_i$, die ja die Varianzen der $y_i'$ angeben. **E** und $\Lambda$ können dabei einfach durch ein PCA-Netz, wie etwa das Sanger-Netz, ermittelt werden. Dann ist mit $A = E \cdot \Lambda^{-1/2} \cdot E^T$ , wobei $\Lambda^{-1/2}$ die (komponentenweise) Wurzel der inversen Matrix zu $\Lambda$ darstellt, für den Erwartungswert von $v_1, ..., v_n$*

$$\begin{aligned} \left\langle [v_1, ..., v_n]^T \right\rangle &= \left\langle E\Lambda^{-1/2}E^T \cdot [y_1', ..., y_n']^T \right\rangle \\ &= E\Lambda^{-1/2}E^T \cdot [\langle y_1' \rangle, ..., \langle y_n' \rangle]^T \\ &= 0 \end{aligned}$$

*Für die Varianzen von $\underline{v} := [v_1, ..., v_n]^T$ in der Diagonalen der Autokorrelationsmatrix $C_{vv}$ von $\underline{v}$ gilt*

$$
\begin{aligned}
C_{vv} &= \left\langle (\underline{v} - \langle \underline{v} \rangle)(\underline{v} - \langle \underline{v} \rangle)^T \right\rangle \\
&= \left\langle \underline{v}\underline{v}^T \right\rangle \\
&= \left\langle E\Lambda^{-1/2}E^T \underline{y}'\underline{y}'^T E\Lambda^{-1/2}E^T \right\rangle \\
&= E\Lambda^{-1/2}E^T \left\langle \underline{y}'\underline{y}'^T \right\rangle E\Lambda^{-1/2}E^T \\
&= E\Lambda^{-1/2} \cdot \left(E^T C_{\underline{y}'\underline{y}'} E\right) \cdot \Lambda^{-1/2}E^T \\
&= E\left(\Lambda^{-1/2}\Lambda\Lambda^{-1/2}\right)E^T \\
&= EE^T \\
&= I
\end{aligned}
$$

*Wir sehen, dass die Signale $v_1, ..., v_n$ durch die PCA nicht nur dekorreliert, sondern in der Tat auf Varianz = 1 normiert wurden.*

nach oben

## 4.3 Finden der ersten unabhängigen Quelle

Den Vorgang, ein Signal in unabhängige Quellen zu zerlegen (bzw. zu bestimmen, aus welcher Kombination unabhängiger Quellen ein bestimmtes Signal zusammengesetzt ist), bezeichnet man dabei auch als *independent component analysis* (ICA). In diesem Abschnitt wollen wir uns nun einen sehr effizienten ICA-Algorithmus ansehen mit dem wir das blind source separation Problem lösen können, der 1997 von Aapo Hyvärinen und Erkki Oja entdeckt wurde.

Es ist im Übrigen auch so, daß die Annahmen über die Varianz und den Erwartungswert der einzelnen Signale Voraussetzung für die Konvergenz einiger ICA-Algorithmen ist. Das gilt auch für den Algorithmus, den wir uns jetzt ansehen wollen, daher erfordert er die Vorverarbeitungsstufe (Zentrieren und Weißen), die wir zuvor besprochen haben.

Wie zuvor erwähnt, müssen wir zur Bestimmung einer unabhängigen Quelle die "Nicht-Gaußheit" des von uns rücktransformierten Signals $s_i' = \underline{w}_i^T \cdot [y_1, ..., y_n]^T$ maximieren, was die Frage aufwirft, wie man Gaußheit eigentlich genau messen kann. Ein mögliches Maß für die Gaußheit einer Quelle ist die sogenannte *Kurtosis*. Sie ist definiert als:

$$K(x) = \frac{\langle (x - \langle x \rangle)^4 \rangle}{\langle (x - \langle x \rangle)^2 \rangle^2} - 3$$

*Der Subtrahend von 3 mag etwas willkürlich erscheinen. Die tiefere Bedeutung liegt darin, daß auf diese Weise sichergestellt wird, daß die Kurtosis einer Gaußverteilung gerade Null ergibt. (Für Gaußverteilungen gilt gerade die Beziehung $\langle (x - \langle x \rangle)^4 \rangle = 3 \cdot \langle (x - \langle x \rangle)^2 \rangle^2$.) Bei spitzen Verteilungen nimmt sie positive Werte an. Man bezeichnet diese Verteilungen aufgrund ihres Aussehens dann auch als "supergaußisch". Negative Werte stellen flachere Verteilungen dar (wie etwa die Gleichverteilung), dementsprechend werden diese Verteilungen dann auch als "subgaußisch" bezeichnet.*

Da im Nenner gerade das Quadrat der Varianz auftaucht, läßt sich die Kurtosis bei Quellen mit Varianz 1 und Erwartungswert 0 auch schreiben als:

$$K(x) = \langle x^4 \rangle - 3$$

Damit läßt sich als Optimierungsziel das Erreichen einer maximalen oder minimalen Kurtosis eines Ausgabesignals $s_i'$ festlegen. Da die Reihenfolge der ursprünglchen Signale nicht ermittelt werden kann, können wir festlegen, daß $s_1'$ dieses Signal sein soll. Wir wollen dann den folgenden Ausdruck über $\underline{w}_1$ maximieren oder minimieren, unter der Nebenbedingung, daß W eine orthonormale Matrix ist, also insbesondere $\underline{w}_1^T \underline{w}_1 = 1$ gilt:

$$K(s_1') = K(\underline{w}_1^T \underline{v}) = \left\langle \left(\underline{w}_1^T \underline{v}\right)^4 \right\rangle - 3$$

Dieses Problem läßt sich mit dem sog. Lagrange-Verfahren lösen. Zusammen

mit der Nebenbedingung $\underline{w}_1^T \underline{w}_1 = 1$ folgt dann für die Zielfunktion

$$\max R(\underline{w}_1, p) \; oder \; \min R(\underline{w}_1, p)$$
$$R(\underline{w}_1, p) = \left\langle (\underline{w}_1^T \underline{v})^4 \right\rangle + p \cdot (\underline{w}_1^T \underline{w}_1 - 1)$$

Die Ableitung ergibt die optimalen Werte für $\mathbf{w}_1$

$$\nabla_{\underline{w}_1} R(\underline{w}_1, p) = 4 \left\langle (\underline{w}_1^T \underline{v})^3 \underline{v} \right\rangle + 2p\underline{w}_1$$

Alle lokalen Optima $\underline{w}_1^*$ sind gegeben durch:

$$4 \left\langle (\underline{w}_1^{*T} \underline{v})^3 \right\rangle + 2p\underline{w}_1^* = 0 \iff \left\langle (\underline{w}_1^{*T} \underline{v})^3 \underline{v} \right\rangle + a\underline{w}_1^* = \underline{w}_1^*$$

für ein geeignetes *a*. Wir können erkennen, daß $\underline{w}_1^*$ ein Fixpunkt der folgenden Iteration ist:

$$\begin{aligned} \underline{w}_1 &\leftarrow \left\langle (\underline{w}_1{}^T \underline{v})^3 \underline{v} \right\rangle + a\underline{w}_1 \\ \underline{w}_1 &\leftarrow \underline{w}_1 \cdot \frac{1}{||\underline{w}_1||_2} \end{aligned}$$

Man kann diese Iteration auch wieder als Gradientenabstieg/aufstieg interpretieren (dazu einfach ein $\underline{w}$ von $a\underline{w}$ abspalten).

nach oben

## 4.4 Trennen der restlichen Quellen

Wir haben also gesehen, daß die Iterationsvorschrift (startend mit einem zufälligen $\underline{w}_1$) sicherstellt, daß $\underline{w}_1$ gegen (irgend)einen Vektor der Entmischungsmatrix B konvergiert und uns damit durch $\underline{s}_1 = \underline{w}_1^T \underline{v}$ ein zugehöriges entmischtes und unabhängiges Quellensignal liefert. Dabei wird diejenige Quelle zuerst gefunden, für die der Ausdruck $\sqrt{|K(s_m)|} \cdot |z_m(0)|$

maximal ist unter allen m. Um auch die anderen Quellen zu finden, müssen wir nur noch sicherstellen, daß dieser Ausdruck bezüglich des nächsten Vektors $\underline{w}_2$, den wir iterieren, irgendwie auf Null gesetzt wird. Anschließend können wir die Fixpunktiteration von neuem starten und werden einen neuen Vektor $\underline{w}_2$ der Entmischungsmatrix erhalten, usw.

Glücklicherweise reicht es dazu aus, die Vektoren $\underline{w}_1, ..., \underline{w}_n$, die wir ermitteln, zu orthonormalisieren. Die Orthonormalisierung von $\underline{w}_i$ läßt sich dabei ganz einfach nach dem Gram-Schmidt-Verfahren bewerkstelligen, bei dem wir alle Komponenten um ihre Anteile in Richtung der anderen $\mathbf{w}_i$ reduzieren

$$\begin{aligned} \underline{w_i} &\leftarrow \underline{w_i} - \sum_{j=1}^{i-1} \underline{w}_j^T \underline{w_i} \cdot \underline{w}_j \\ \underline{w_i} &\leftarrow \underline{w_i} \cdot \frac{1}{||\underline{w}_i||_2} \end{aligned}$$

Wir erhalten dann abschließend das folgende Iterationsschema:

$i \leftarrow 1$

1. Initialisierung

   $\underline{w}_i \leftarrow$ zufällig aus $\mathbb{R}^n$

2. Iteriere, bis konvergiert :

   $\underline{w}_i \leftarrow \langle (\underline{w_i}^T \underline{v})^3 \underline{v} \rangle - 3\underline{w}_i$

   $\underline{w}_i \leftarrow \underline{w}_i - \sum_{j=1}^{i-1} \underline{w}_j^T \underline{w}_i \cdot \underline{w}_j$

   $\underline{w}_i \leftarrow \underline{w}_i \cdot \frac{1}{||\underline{w}_i||_2}$

3. Nächster Schritt :

   $i \leftarrow i + 1$

   wenn noch nicht alle Vektoren gefunden, gehe zu 1

Das Vorliegen von Konvergenz in Schritt 2 kann dabei durch Berechnung der Richtungsänderung von $\underline{w}_i$ nach Ausführen eines Iterationsschrittes ermittelt werden. Konvergenz liegt vor, wenn sich die Richtung von $\underline{w}_i$ entweder nicht

oder um genau 180 Grad verändert. In diesem Fall ist $\underline{w}_i$ gerade gegen ein $\pm\underline{b}$ konvergiert.

nach oben