# ICA model, decorrelation, non-Gaussianity, and FastICA

Jarmo Hurri

# Purpose of this lecture

Provide the audience with an understanding of

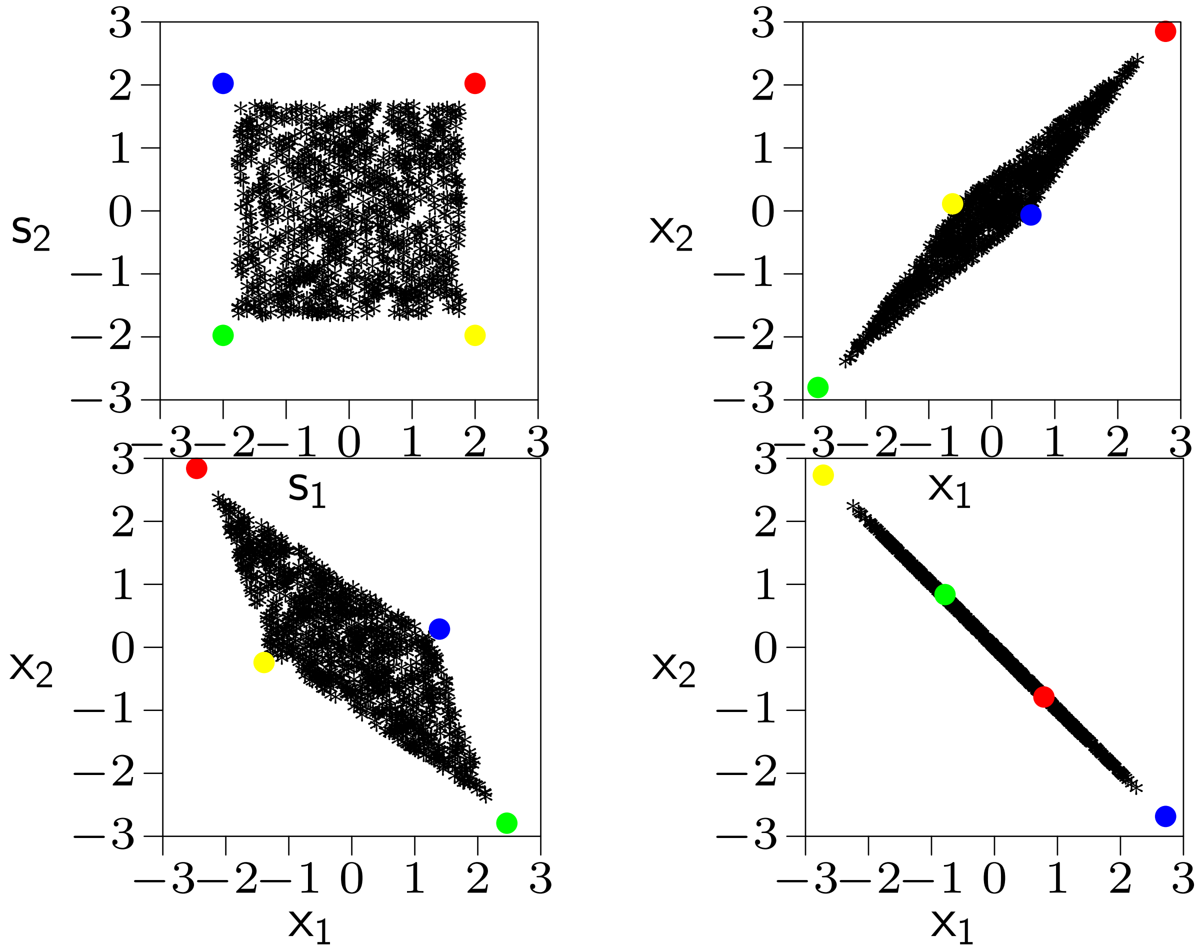
- the ICA data model
- why and how the model can be solved.

# Contents

# ICA-model

- observed data $\mathbf{x} = [x_1 \; x_2 \; \cdots \; x_m]^T$ (random vector)
- independent latent variables $\mathbf{s} = [s_1 \; s_2 \; \cdots \; s_n]^T$ (random vector), $f_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} f_{s_i}(s_i)$
- $\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{n} \mathbf{a_i} s_i$, $\mathbf{A} = [\mathbf{a}_1 \; \mathbf{a}_2 \; \cdots \; \mathbf{a}_n]$
- we observe only a sample from $\mathbf{x}$, we have to solve both $\mathbf{A}$ and $\mathbf{s}$ with as few assumptions as possible

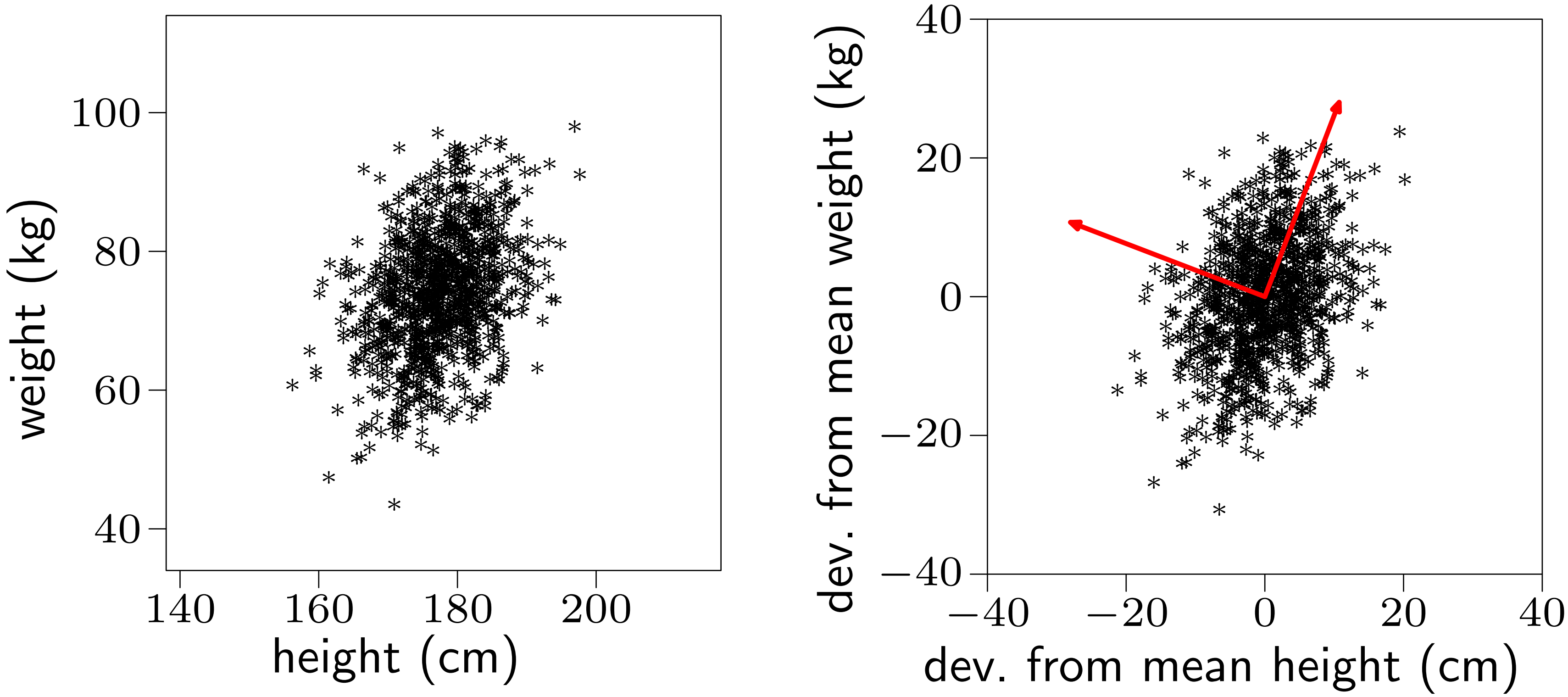
# ICA-mixture — examples



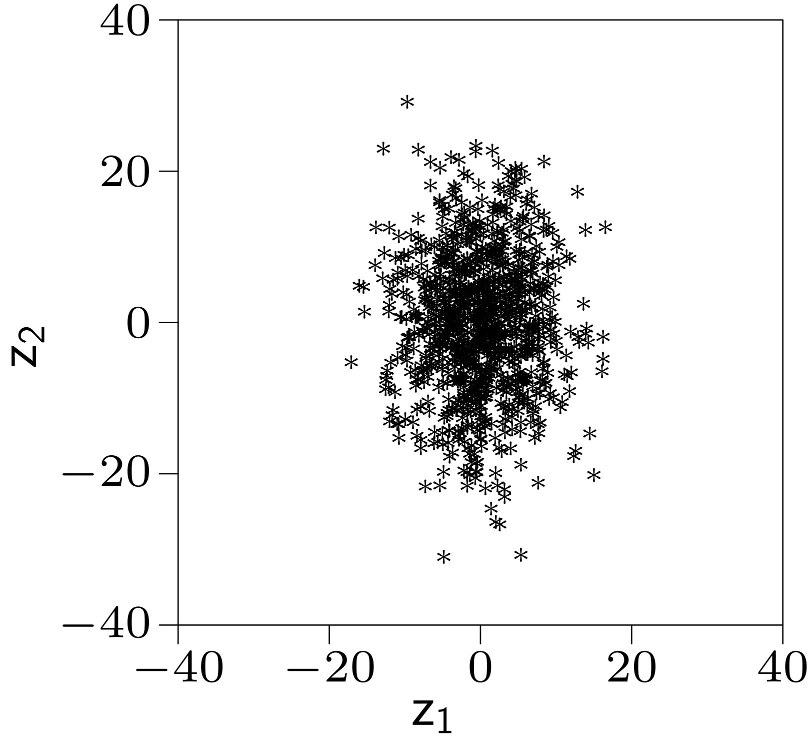
# Limitations of the ICA model

- in general, assume that at least as many observables and hidden components, $m \geq n$ (but, Patrik Wed.)
- assume that $\mathbf{A}$ invertible, $\mathbf{W} = \mathbf{A}^{-1}$
- component ordering & scale / sign indeterminacy:
  - $\mathbf{x} = \sum_i \left(\mathbf{a_i}\lambda_i\right)\left(s_i/\lambda_i\right)$
  - $\mathbf{P}$ permutation matrix, $\mathbf{\Lambda} = \mathbf{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$
  - $\mathbf{x} = \underbrace{\mathbf{A}\mathbf{P}^{-1}\mathbf{\Lambda}^{-1}}_{=\mathbf{A}_*}\underbrace{\mathbf{\Lambda}\mathbf{P}\mathbf{s}}_{=\mathbf{s}_*}$

# Linear correlations (1/3)



# Linear correlations (2/3)




- $\mathbf{C_x} = \mathbf{E}\mathbf{D}\mathbf{E}^T$
- $\mathbf{z} = \mathbf{E}^T\mathbf{x}$

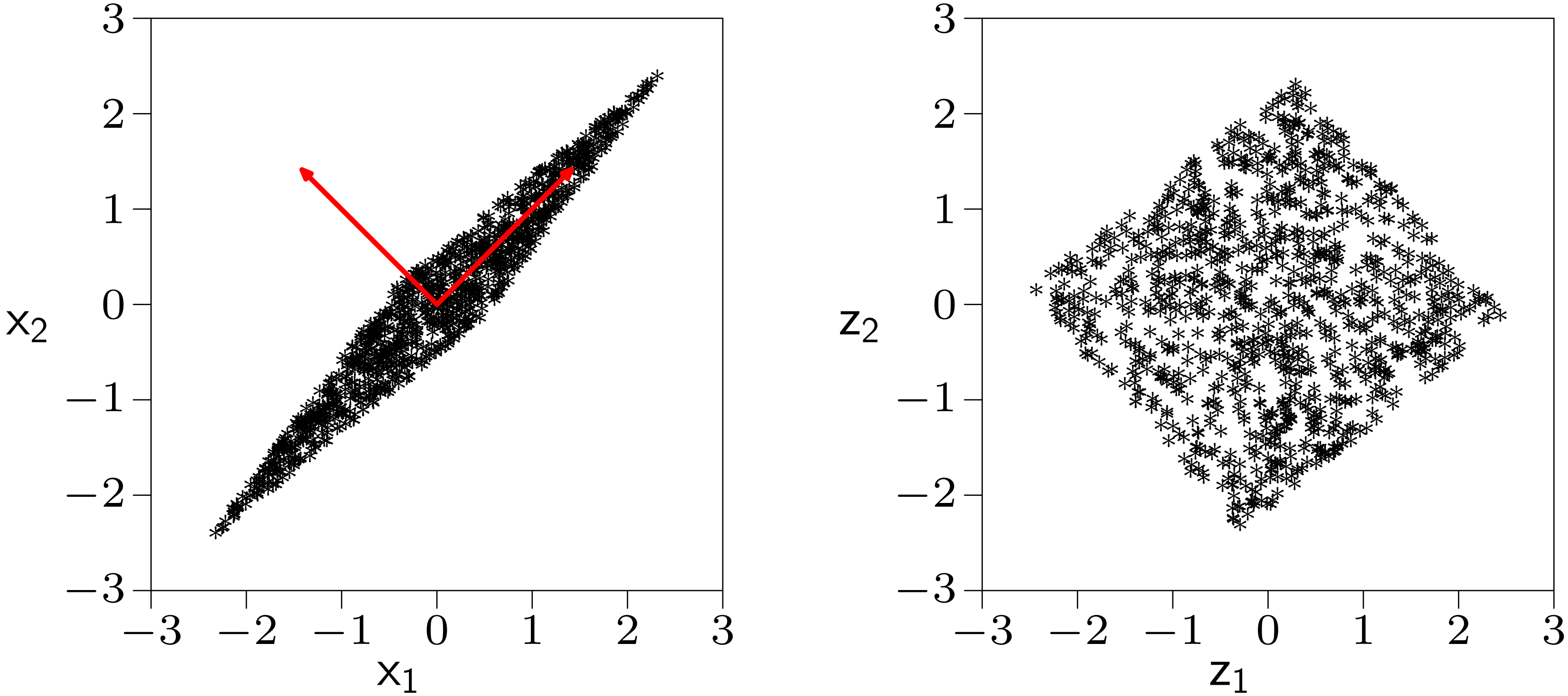
- $\mathbf{C_z} = \mathsf{E}\left\{\mathbf{z}\mathbf{z}^T\right\} = \mathsf{E}\left\{\mathbf{E}^T\mathbf{x}\mathbf{x}^T\mathbf{E}\right\} = \mathbf{E}^T\mathbf{E}\mathbf{D}\mathbf{E}^T\mathbf{E} = \mathbf{D}$

# Linear correlations (3/3)

- uncorrelated multivariate Gaussian is independent

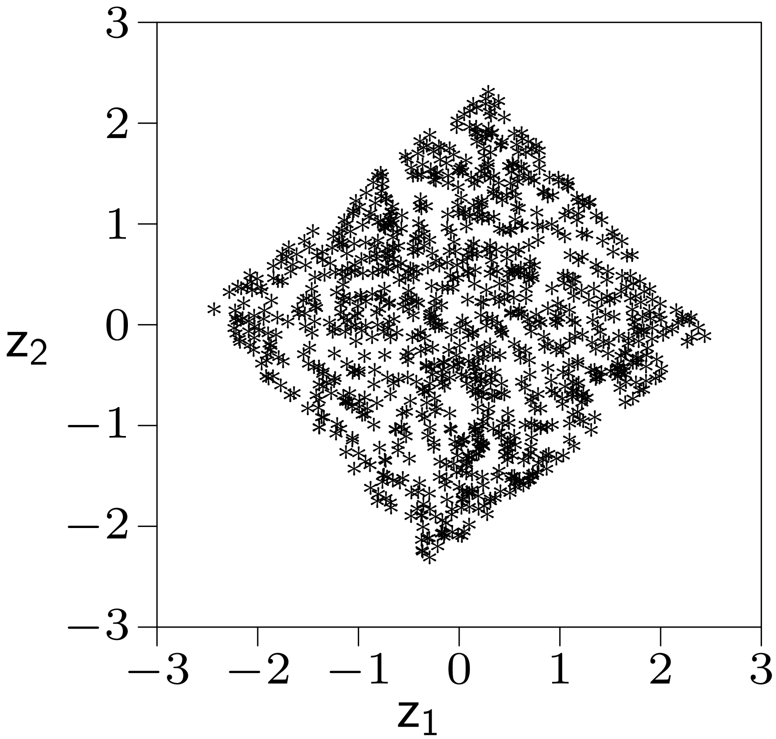
$$
\begin{aligned}
f_{\mathbf{z}}(\mathbf{z}) &= K \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{C}_{\mathbf{z}}^{-1} \mathbf{z}\right) \\
&= K \exp\left(-\sum_{i=1}^{n}\left(\frac{z_i}{\sqrt{2}\sigma_i}\right)^2\right) \\
&= \prod_{i=1}^{n} \sqrt[n]{K} e^{-\left(\frac{z_i}{\sqrt{2}\sigma_i}\right)^2} = \prod_{i=1}^{n} f_{z_i}(z_i)
\end{aligned}
$$

# Whitening (1/4)




- $\mathbf{z} = \underbrace{\mathbf{D}^{-1/2}\mathbf{E}^T}_{=\mathbf{V}}\mathbf{x}$, $\mathsf{E}\left\{\mathbf{z}\mathbf{z}^T\right\} = \mathbf{I}$

# Whitening (2/4)




- removes linear dependencies
- normalizes variance of projections
- new problem: search for orthonormal basis

- "whitening": frequency contents in a decorrelated signal

# Whitening (3/4)

- let $\mathbf{w} = [w_1 \ w_2 \ \cdots \ w_n]^T$, $\|\mathbf{w}\| = 1$
- let $y = \mathbf{w}^T\mathbf{z}$, $\mathsf{E}\{\mathbf{z}\} = \mathbf{0}$

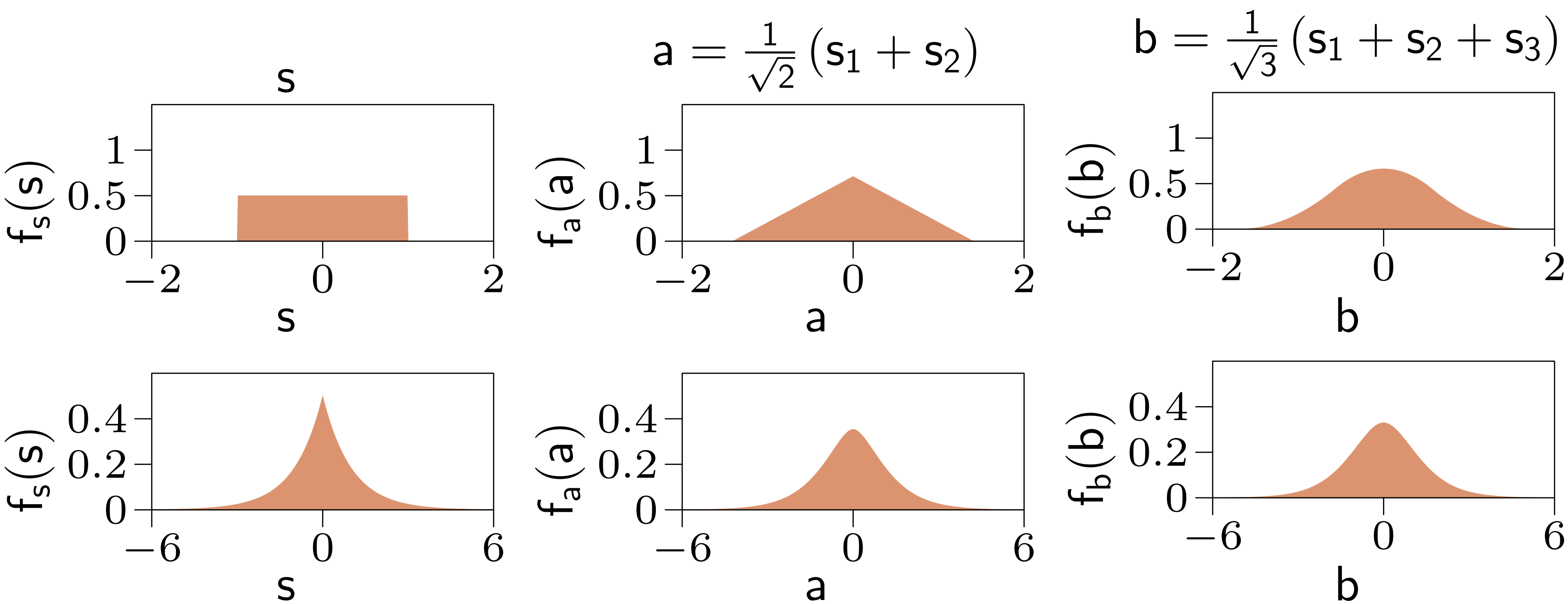
$$
\begin{aligned}
\text{var}\{y\} &= \mathsf{E}\{y^2\} \\
&= \mathsf{E}\{\mathbf{w}^T\mathbf{z}\mathbf{z}^T\mathbf{w}\} \\
&= \mathbf{w}^T\mathsf{E}\{\mathbf{z}\mathbf{z}^T\}\mathbf{w} = \|\mathbf{w}\|^2 = 1
\end{aligned}
$$

# Whitening (4/4)

- let $\mathbf{A}^{-1} = \mathbf{W}$
- $\mathbf{z} = \mathbf{V}\mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{V}^{-1}\mathbf{z}$
- $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x} = \mathbf{W}\mathbf{x} = \underbrace{\mathbf{W}\mathbf{V}^{-1}}_{=\mathbf{W}_*}\mathbf{z} = \mathbf{W}_*\mathbf{z}$
- $\mathbf{I} = \mathsf{E}\left\{\mathbf{s}\mathbf{s}^T\right\} = \mathsf{E}\left\{\mathbf{W}_*\mathbf{z}\mathbf{z}^T\mathbf{W}_*^T\right\} = \mathbf{W}_*\underbrace{\mathsf{E}\left\{\mathbf{z}\mathbf{z}^T\right\}}_{=\mathbf{I}}\mathbf{W}_*^T = \mathbf{W}_*\mathbf{W}_*^T$

# Non-Gaussianity (1/3)

- central limit theorem: the sum of independent r.v's approaches a Gaussian distribution when $n \longrightarrow \infty$



# Non-Gaussianity (2/3)

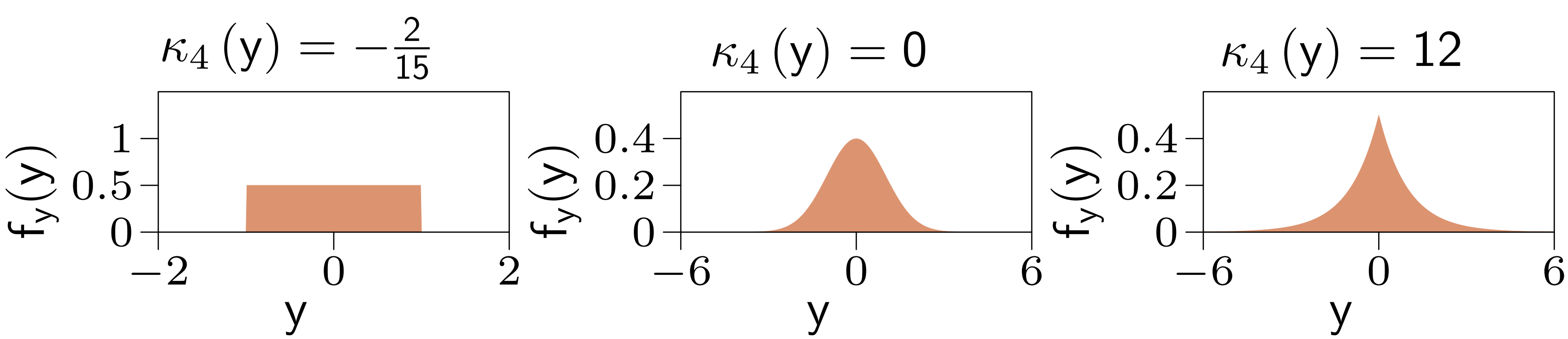
uniform distributions

# Non-Gaussianity (3/3)

- assume that at most one component of $\mathbf{s}$ has a normal distribution
- when components mixed, mixture "closer" to a Gaussian than the originals
- $\Rightarrow$ components can be found by searching for maximally non-Gaussian linear combinations of the observed data $\mathbf{x}$

# Kurtosis (1/2)

- a measure of non-Gaussianity
- measures the peaknedness of a (unimodal) distribution
- $\kappa_4(y) = \mathrm{E}\{y^4\} - \underbrace{3\left(\mathrm{E}\{y^2\}\right)^2}_{=3 \text{ if } \mathrm{E}\{y\}=0 \text{ and whitened}}$



# Kurtosis (2/2)

- if $y_1$ and $y_2$ are statistically independent,
  $\kappa_4 (\alpha_1 y_1 + \alpha_2 y_2) = \alpha_1^4 \kappa_4 (y_1) + \alpha_2^4 \kappa_4 (y_2)$
- solves the ICA problem when the model holds
- can be optimized with a number of different algorithms

# FastICA (1/6)

- whitened data $\mathbf{z}$
- linear transformation $y = \mathbf{w}^T\mathbf{z}$, var $\{y\} = 1 \Leftrightarrow \|\mathbf{w}\| = 1$
- maximize kurtosis $f(\mathbf{w}) = \kappa_4(y) = \mathsf{E}\left\{y^4\right\}$ with constraint $h(\mathbf{w}) = \|\mathbf{w}\|^2 - 1 = 0$
- at optimum $f'(\mathbf{w}) + \lambda h'(\mathbf{w}) = \mathbf{0}^T \Rightarrow 4\mathsf{E}\left\{\left(\mathbf{w}^T\mathbf{z}\right)^3\mathbf{z}\right\} + 2\lambda\mathbf{w} = \mathbf{0}$

# FastICA (2/6)

- $\lambda_* = -\frac{\lambda}{2} \Rightarrow \lambda_* \mathbf{w} = \mathsf{E}\left\{ \left( \mathbf{w}^T \mathbf{z} \right)^3 \mathbf{z} \right\}$
- $\Rightarrow$ <u>direction</u> of $\mathbf{w}$ fixed under iteration

$$\mathbf{w}(k+1) = \mathsf{E}\left\{ \left( \mathbf{w}(k)^T \mathbf{z} \right)^3 \mathbf{z} \right\}$$

- additional twist needed for fast convergence:

$$\mathbf{w}(k+1) = \mathsf{E}\left\{ \left( \mathbf{w}(k)^T \mathbf{z} \right)^3 \mathbf{z} \right\} - 3\mathbf{w}(k)$$

# FastICA (3/6)

- summary of FastICA:

$$\mathbf{w}_{\mathsf{I}}(k+1) = \mathsf{E}\left\{\left(\mathbf{w}(k)^T\mathbf{z}\right)^3\mathbf{z}\right\} - 3\mathbf{w}(k)$$

$$\mathbf{w}(k+1) = \frac{\mathbf{w}_{\mathsf{I}}(k+1)}{\|\mathbf{w}_{\mathsf{I}}(k+1)\|}$$

# FastICA (4/6)

- kurtosis $\mathsf{E}\left\{y^4\right\}$ sensitive to outliers
- FastICA for a general nonlinearity $g(y) = G'(y)$, $G$ non-quadratic:

$$\mathbf{w}_\mathsf{l}(k+1) = \mathsf{E}\left\{g\left(\mathbf{w}(k)^T\mathbf{z}\right)\mathbf{z}\right\} - \mathsf{E}\left\{g'\left(\mathbf{w}^T\mathbf{z}\right)\right\}\mathbf{w}$$

- for example, $g(y) = \tanh(\alpha y)$

# FastICA (5/6)

- multiple components: deflation or symmetric algorithm
- deflation: intermediate Gram-Schmidt orthogonalization ($\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_{\ell-1}]^T$)

$$
\begin{aligned}
\mathbf{w}_{\ell,\mathsf{I}}(k+1) &= \mathsf{E}\left\{\left(\mathbf{w}_\ell(k)^T\mathbf{z}\right)^3\mathbf{z}\right\} - 3\mathbf{w}_\ell(k) \\
\mathbf{w}_{\ell,\mathsf{II}}(k+1) &= \mathbf{w}_{\ell,\mathsf{I}} - \mathbf{W}^T\mathbf{W}\mathbf{w}_{\ell,\mathsf{I}} \\
\mathbf{w}_\ell(k+1) &= \frac{\mathbf{w}_{\ell,\mathsf{II}}(k+1)}{\|\mathbf{w}_{\ell,\mathsf{II}}(k+1)\|}
\end{aligned}
$$

# FastICA (6/6)

- symmetric algorithm: simultaneous updates / orthogonalization

$$
\begin{aligned}
\mathbf{w}_{\ell,\mathsf{I}}(k+1) &= \mathsf{E}\left\{\left(\mathbf{w}_\ell(k)^T\mathbf{z}\right)^3\mathbf{z}\right\} - 3\mathbf{w}_\ell(k), \\
&\quad \ell = 1, \ldots, n \\
\mathbf{W}(k+1) &= \mathbf{W}_{\mathsf{I}}(k+1)^T\left(\mathbf{W}_{\mathsf{I}}(k+1)\mathbf{W}_{\mathsf{I}}(k+1)^T\right)^{-1/2}
\end{aligned}
$$

# Summary

- linear model $\mathbf{x} = \mathbf{A}\mathbf{s}$, components of $\mathbf{s}$ statistically independent
- observe $\mathbf{x}$, solve $\mathbf{A}$ and $\mathbf{s}$ (except multiplier, order)
- whitening decorrelates and unifies variance
- after whitening solve for orthogonal basis by maximizing non-Gaussianity
- a family of fixed-point algorithms (FastICA)

# What else...

- today: exercises and handouts
- tomorrow: Patrik