Introduction to Independent Component Analysis

Jarmo Hurri, Patrik Hoyer, Aapo Hyvärinen

Course timetable (this week)

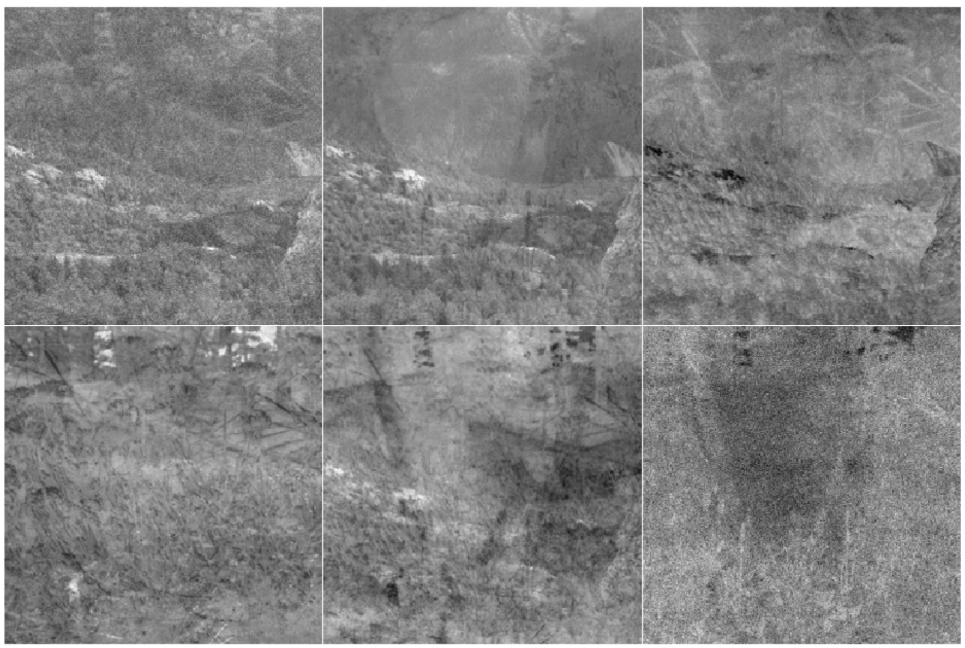
- today: ICA examples and mathematical background
- tomorrow: ICA model, decorrelation, non-Gaussianity, FastICA (me)
- Wednesday: practical considerations, ICA by maximum likelihood, image representations (Patrik)

Course timetable (next week)

- next Monday: nonnegative sparse coding, modeling residual dependencies (Patrik)
- next Tuesday: recent advances and open questions (Aapo)
- all classes at 14:15–16:00 in this room (A414)
- `http://www.cs.helsinki.fi/jarmo.hurri`

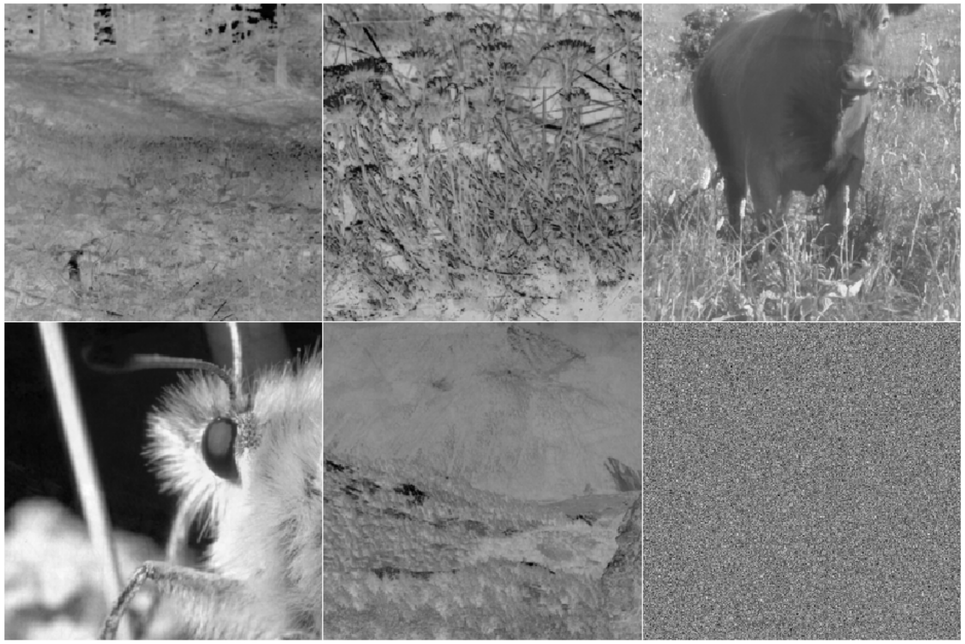
Contents

Example



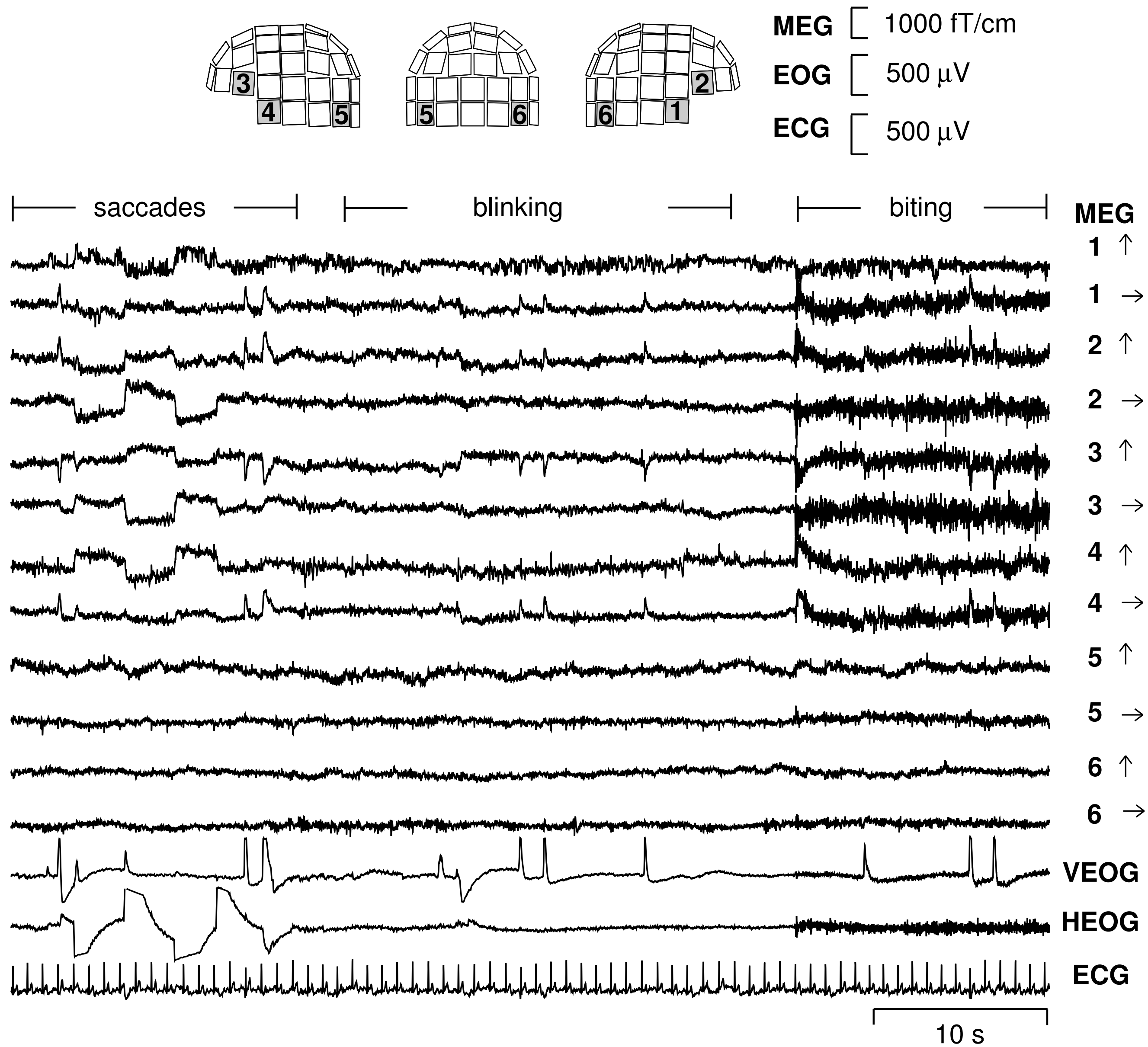
- 6 images
- linear mixtures of 6 originals
- determine originals

Independent components

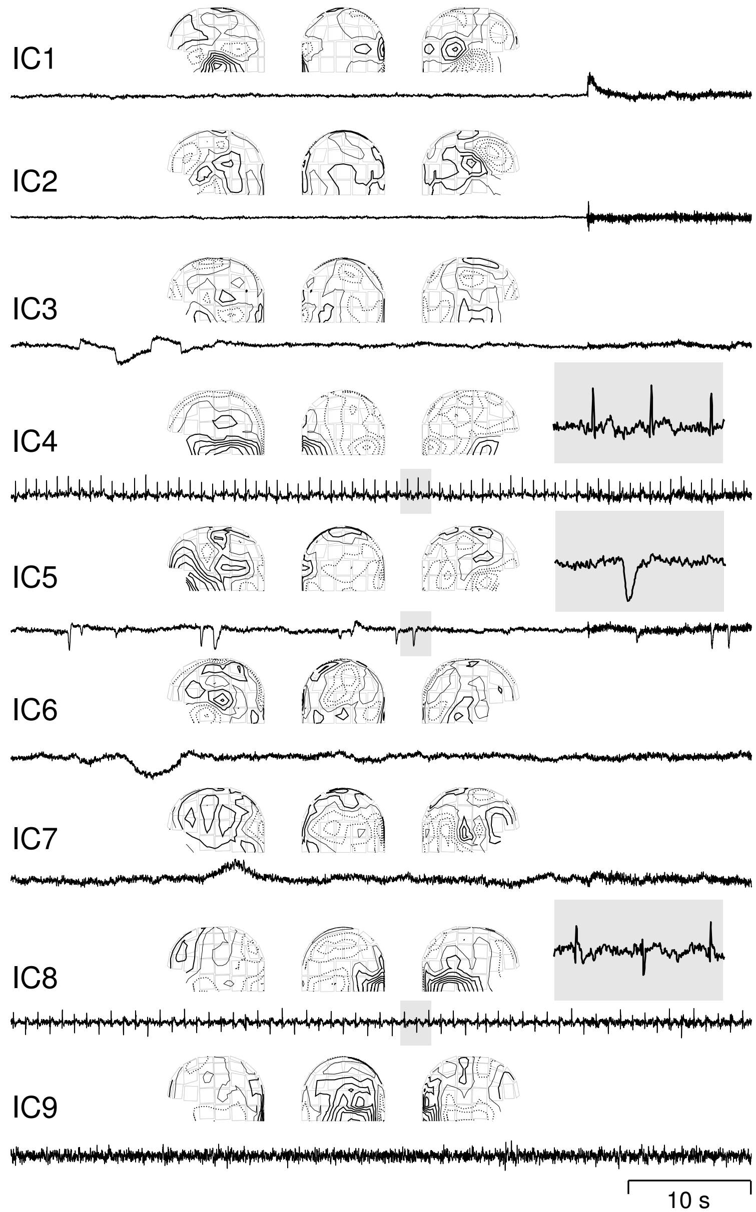


Independent components

- independent latent (hidden) variables
- linear phenomenon



3
4
5
5
6
6
1
2
MEG
1000 fT/cm
EOG
500 µV
ECG
500 µV
saccades
blinking
biting
MEG
1 ↑
1 →
2 ↑
2 →
3 ↑
3 →
4 ↑
4 →
5 ↑
5 →
6 ↑
6 →
VEOG
HEOG
ECG
10 s


IC1
IC2
IC3
IC4
IC5
IC6
IC7
IC8
IC9
10 s

Some ICA application areas

- biomedical signal analysis (EEG/MEG/MRI/fMRI)
- computational neuroscience
- multispectral image analysis
- bioinformatics (transcriptome analysis)
- gas sensor array analysis
- telecommunications

II.Multivariate optimization (1/2)

- differentiable objective $f(\mathbf{w})$
- derivative: local linear approximation of change Δf
- $f'(\mathbf{w}) = [\partial f / \partial w_1 \ \partial f / \partial w_2 \ \cdots \ \partial f / \partial w_n]$
- $f(\mathbf{w} + \Delta \mathbf{w}) - f(\mathbf{w}) \approx f'(\mathbf{w}) \Delta \mathbf{w}$
- gradient rule: $\mathbf{w}(k+1) = \mathbf{w}(k) \pm \alpha f'(\mathbf{w})^T$

Multivariate optimization (2/2)

- differentiable constraint $h(\mathbf{w}) = 0$
- necessary conditions:

$$f'(\mathbf{w}) + \lambda h'(\mathbf{w}) = \mathbf{0}^T$$

- different iterative methods (e.g., projected gradient)

III. Multivariate statistics

- notation
- basic statistics
- multivariate Gaussian density
- principal component analysis
- statistical independence

Notation

- random variables: $x, y, \ldots$
- density functions: $p_x(x), p_y(y), \ldots$
- random vectors: $\mathbf{x} = [x_1 \; x_2 \; \cdots \; x_n]^T$
  - each component a (continuous) random variable
  - $F(\mathbf{x}_0) = \mathsf{P}(\mathbf{x} \leq \mathbf{x}_0)$
  - $p_{\mathbf{x}}(\mathbf{x}_0) = \frac{\partial}{\partial x_1}\frac{\partial}{\partial x_2} \cdots \frac{\partial}{\partial x_n} F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_0}$

Basic statistics (1/3)

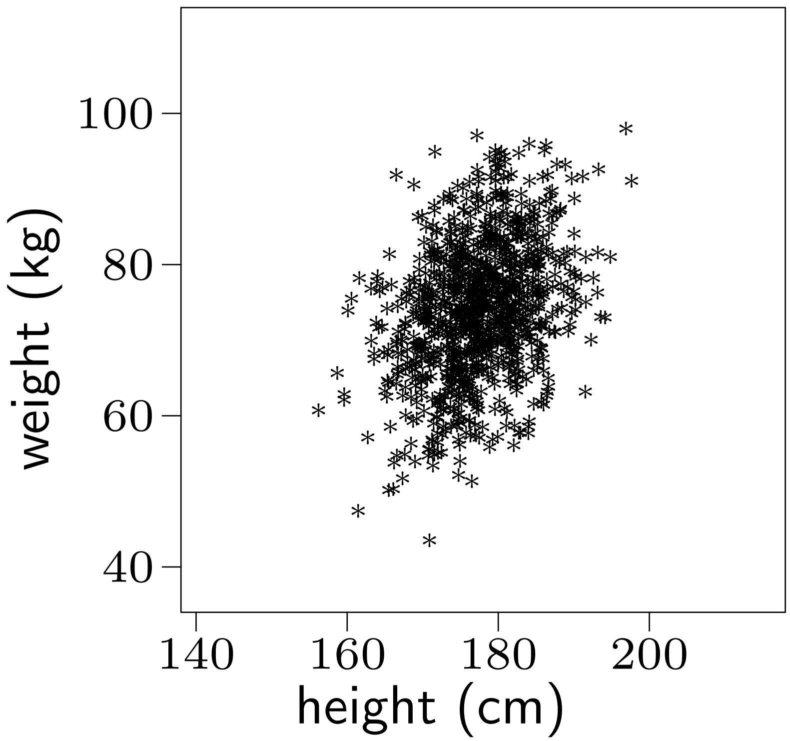
- expectation: $\mathsf{E}\left\{g(\mathbf{x})\right\} = \int g(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x})\, d\,\mathbf{x}$
- mean: $\mathsf{E}\left\{\mathbf{x}\right\} = \left[\mathsf{E}\left\{x_1\right\} \;\; \mathsf{E}\left\{x_2\right\} \;\; \cdots \;\; \mathsf{E}\left\{x_n\right\}\right]^T$
- correlation matrix:

$$\mathbf{R}_{\mathbf{x}} = \mathsf{E}\left\{\mathbf{x}\mathbf{x}^T\right\} = \begin{bmatrix} \mathsf{E}\left\{x_1^2\right\} & \mathsf{E}\left\{x_1 x_2\right\} & \cdots & \mathsf{E}\left\{x_1 x_n\right\} \\ \mathsf{E}\left\{x_2 x_1\right\} & \mathsf{E}\left\{x_2^2\right\} & \cdots & \mathsf{E}\left\{x_2 x_n\right\} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{E}\left\{x_n x_1\right\} & \mathsf{E}\left\{x_n x_2\right\} & \cdots & \mathsf{E}\left\{x_n^2\right\} \end{bmatrix}$$

Basic statistics (2/3)

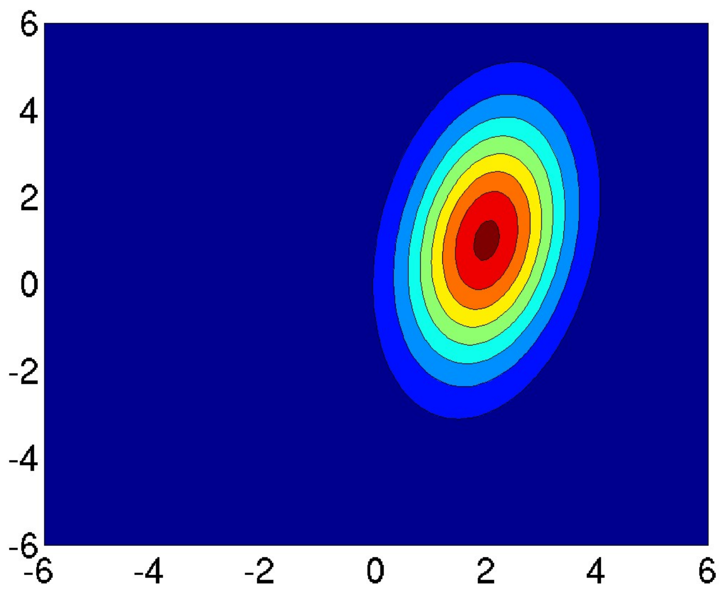
- covariance matrix $\mathbf{C}_{\mathbf{x}} = \mathsf{E}\left\{(\mathbf{x} - \mathsf{E}\{\mathbf{x}\})(\mathbf{x} - \mathsf{E}\{\mathbf{x}\})^T\right\}$
  - note: if $\mathsf{E}\{\mathbf{x}\} = \mathbf{0}$, then $\mathbf{C}_{\mathbf{x}} = \mathbf{R}_{\mathbf{x}}$

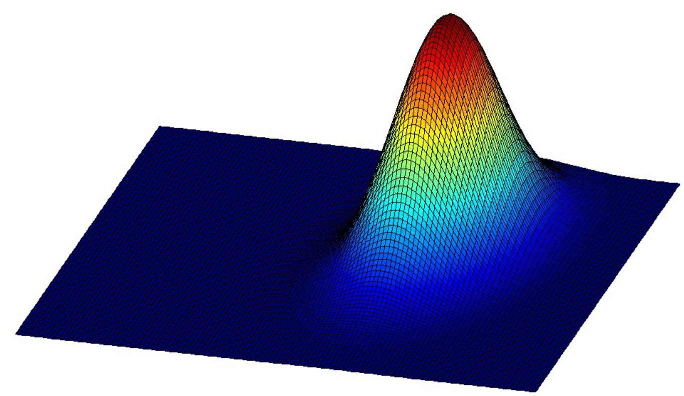
Basic statistics (3/3)



- x_1: height, x_2: weight
- $\mathbf{C_x} = \begin{bmatrix} 37.26 & 17.95 \\ 17.95 & 77.32 \end{bmatrix}$
- $\mathrm{cov}\left\{x_1, x_2\right\} > 0$
- $\mathsf{var}\left\{x_2\right\} > \mathsf{var}\left\{x_1\right\}$

Multivariate Gaussian density (1/2)

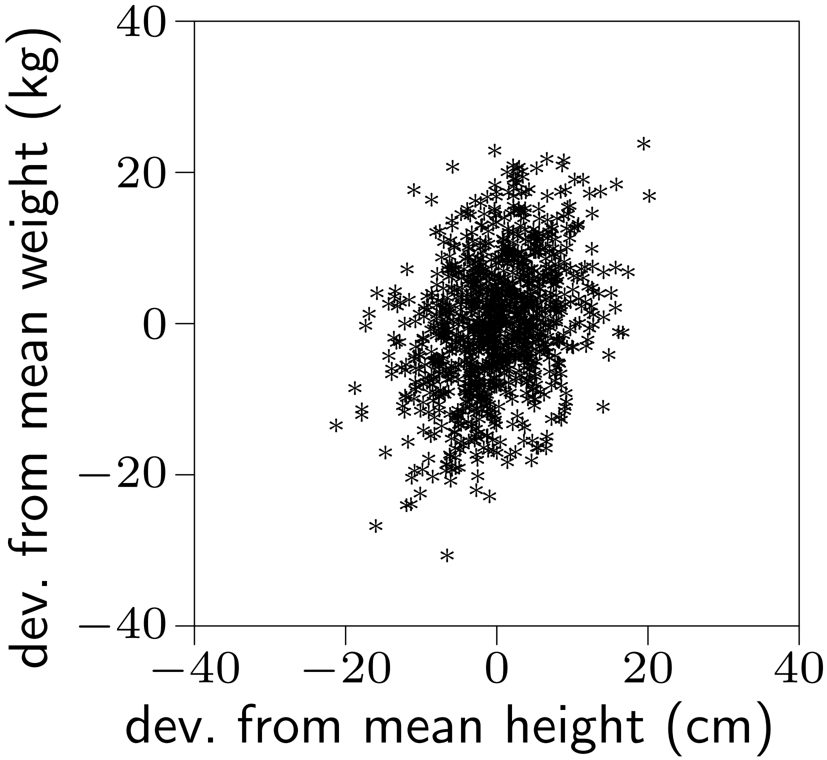




Multivariate Gaussian density (2/2)

- $p_{\mathbf{x}}(\mathbf{x}) = K \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m_x})^T \mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x}-\mathbf{m_x})\right)$
- $K = \left((2\pi)^{n/2} \det(\mathbf{C_x})^{1/2}\right)^{-1}$
- $\mathsf{E}\{\mathbf{x}\} = \mathbf{m_x}$
- covariance matrix $\mathbf{C_x}$
- completely specified by 1st and 2nd order statistics

Principal component analysis (1/4)



- describe data with <u>one</u> linear projection
- $\mathbf{w} = [w_1\ w_2\ \cdots\ w_n]^T$, $\|\mathbf{w}\|^2 = 1$
- new data $\mathbf{x}_* = \left(\mathbf{w}^T\mathbf{x}\right)\mathbf{w}$
- min $f(\mathbf{w}) = \mathsf{E}\left\{\|\mathbf{x} - \mathbf{x}_*\|^2\right\}$, s.t. $h(\mathbf{w}) = \|\mathbf{w}\|^2 - 1 = 0$

Principal component analysis (2/4)

$$\mathbf{u}\colon\ \mathbf{u}^T\mathbf{w} = 0,\ \|\mathbf{u}\|^2 = 1$$

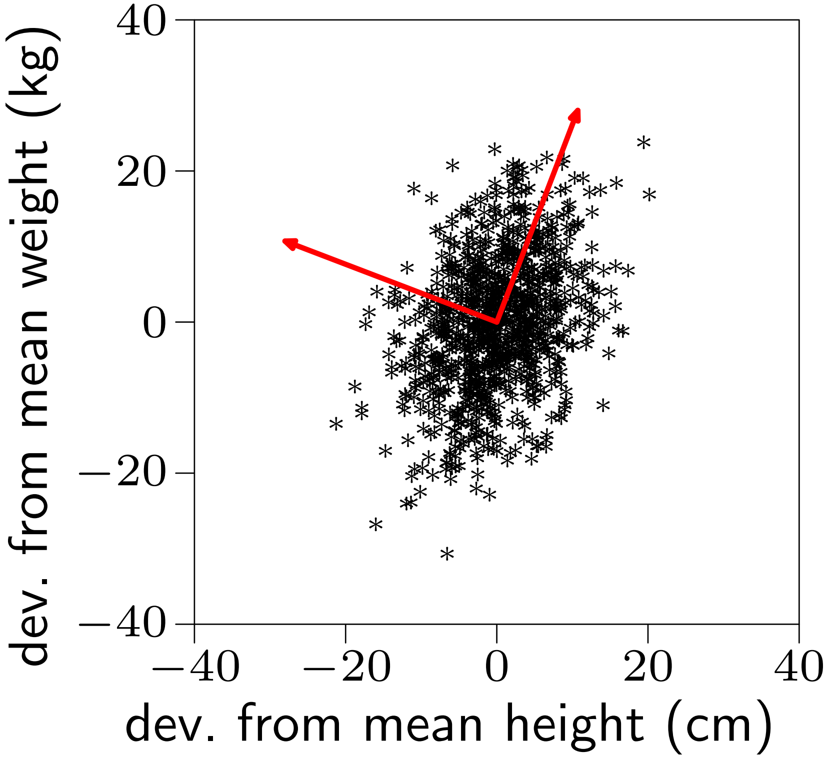
$$\begin{aligned}
f(\mathbf{w}) &= \mathsf{E}\{\|\underbrace{(\mathbf{w}^T\mathbf{x})\mathbf{w} + (\mathbf{u}^T\mathbf{x})\mathbf{u}}_{=\mathbf{x}} - \underbrace{(\mathbf{w}^T\mathbf{x})\mathbf{w}}_{=\mathbf{x}_*}\|^2\} \\
&= \mathsf{E}\left\{\|(\mathbf{u}^T\mathbf{x})\mathbf{u}\|^2\right\} \text{ expected norm of proj.} \\
&= \mathsf{E}\left\{\mathbf{u}^T(\mathbf{u}^T\mathbf{x})(\mathbf{u}^T\mathbf{x})\mathbf{u}\right\} \\
&= \mathbf{u}^T\mathsf{E}\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{u} = \mathbf{u}^T\mathbf{R}_\mathbf{x}\mathbf{u}
\end{aligned}$$

Principal component analysis (3/4)

$$\begin{aligned} f(\mathbf{u}) &= \mathbf{u}^T \mathbf{R}_{\mathbf{x}} \mathbf{u} \\ h(\mathbf{u}) &= \mathbf{u}^T \mathbf{u} - 1 = 0 \end{aligned}$$

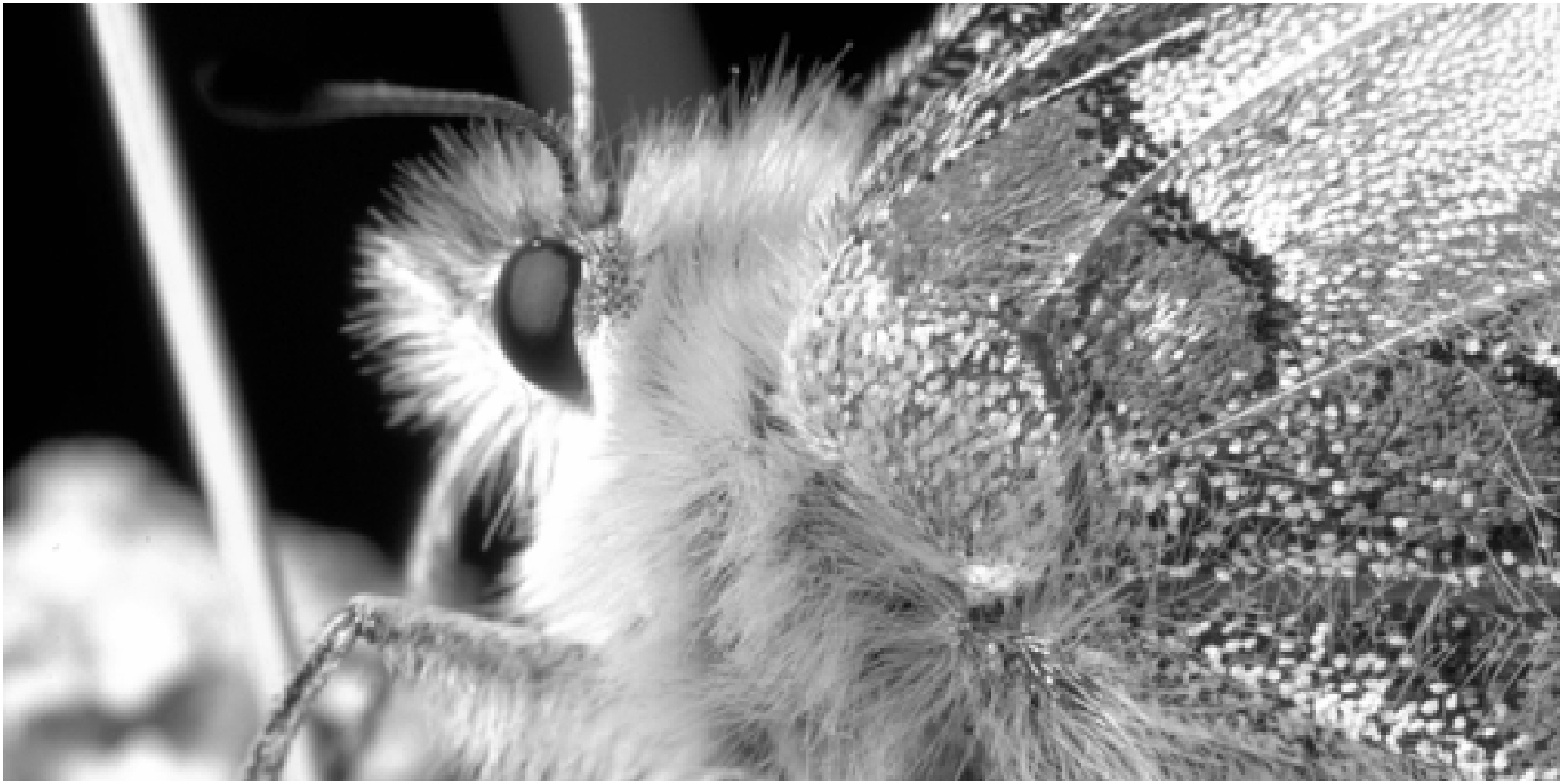
$$\begin{aligned} f'(\mathbf{u}_{\mathsf{opt}}) - \lambda h'(\mathbf{u}_{\mathsf{opt}}) &= \mathbf{0}^T \\ 2\mathbf{u}_{\mathsf{opt}}^T \mathbf{R}_{\mathbf{x}} - \lambda 2 \mathbf{u}_{\mathsf{opt}}^T &= \mathbf{0}^T \\ \mathbf{R}_{\mathbf{x}} \mathbf{u}_{\mathsf{opt}} &= \lambda \mathbf{u}_{\mathsf{opt}} \\ f(\mathbf{u}_{\mathsf{opt}}) &= \lambda \end{aligned}$$

Principal component analysis (4/4)

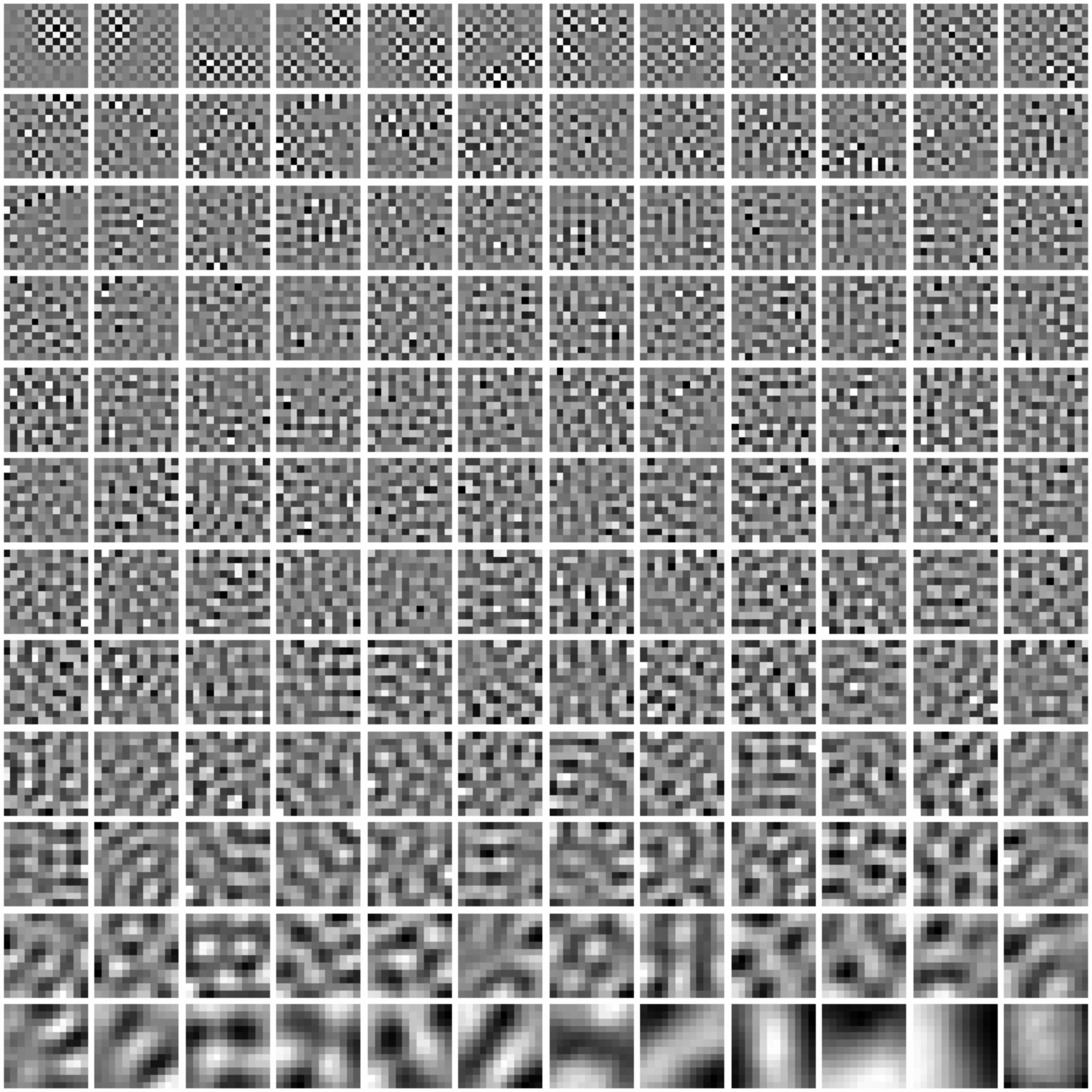


- orthogonal PCA basis
- deflationary minima / maxima of variance
- dimensionality reduction
- noise attenuation

PCA example — original

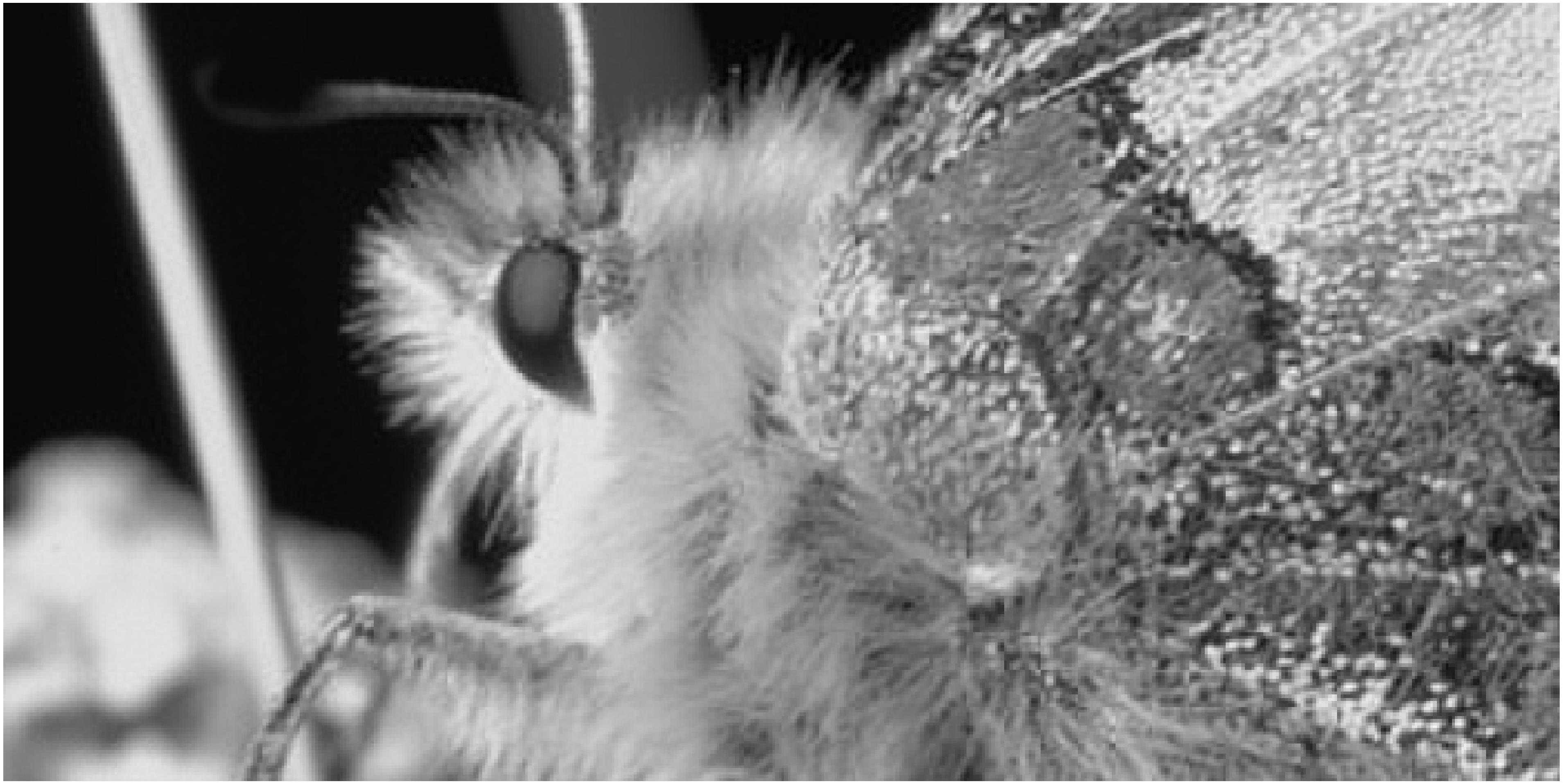


PCA example — PCA basis

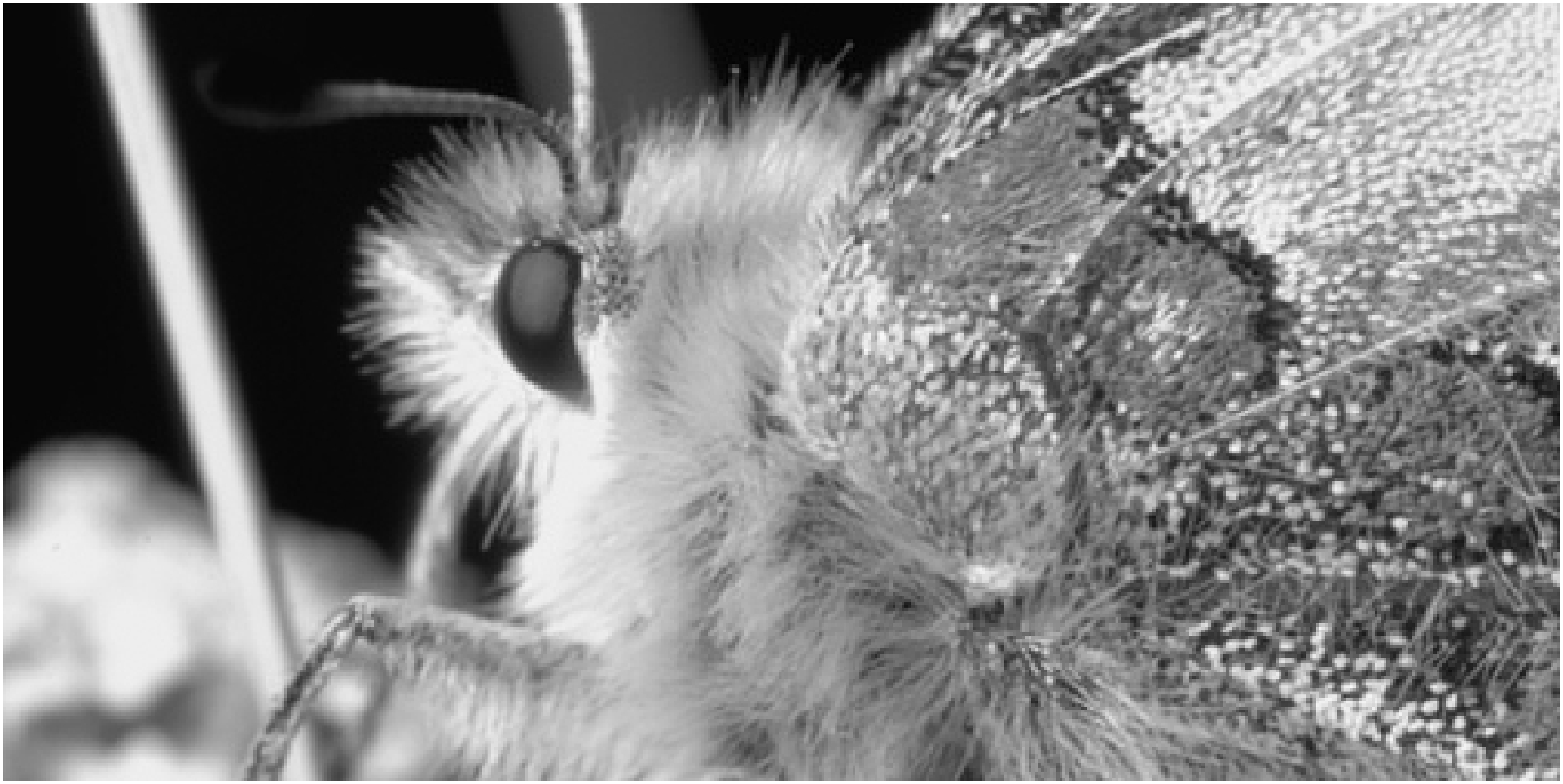


PCA example — 90% compressed

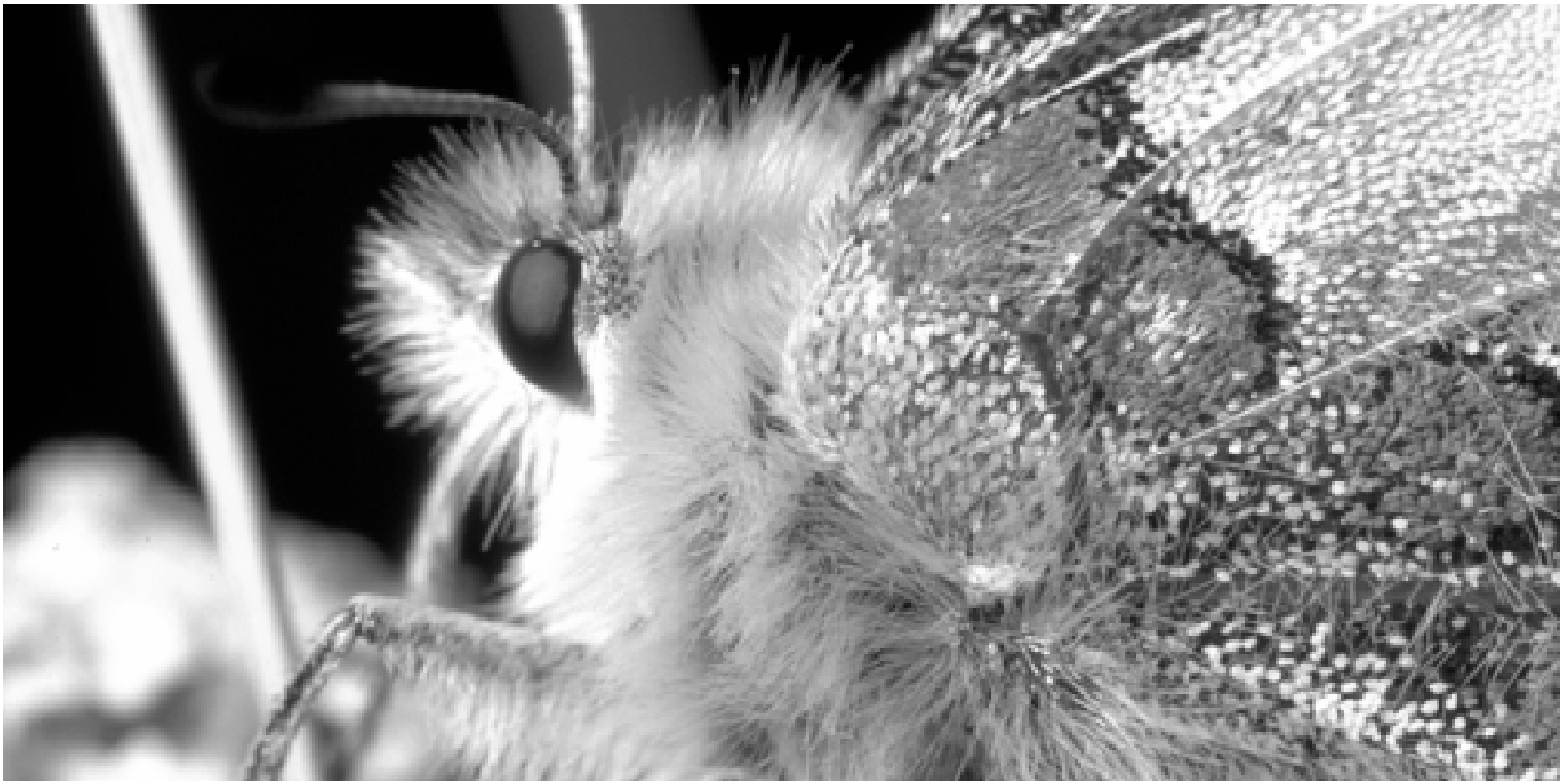
PCA example — 75% compressed



PCA example — 50% compressed



PCA example — original



Statistical independence

- random variables x and y
- statistical independence: knowing the value of x does not provide information about the distribution of y
- $p_y\left(y|x\right) = \frac{p_{x,y}(x,y)}{p_x(x)} = p_y\left(y\right)$
  $\Leftrightarrow p_{x,y}\left(x,y\right) = p_x\left(x\right)p_y\left(y\right)$

IV. Estimation theory

- basic concepts
- maximum-likelihood estimation

Basic concepts (1/3)

- estimation: finding an approximate value or distribution of some parameter (e.g., mean, standard deviation) from random samples of the population

- estimator: a function computing the approximate value or distribution from the samples (e.g., $\hat{\mu} = 1/T\sum_{i=1}^{T} x(i)$)

- estimate: numerical value of the estimator for a given set of samples

Basic concepts (2/3)

- some good properties of estimators:
  - unbiasedness: $\mathsf{E}\left\{\hat{\theta}\right\} = \theta$
  - consistency:

$$\forall \epsilon > 0 \left[\lim_{T \to \infty} \mathsf{P}\left(\left\| \theta - \hat{\theta} \right\| < \epsilon \right) = 1 \right]$$

Basic concepts (3/3)

- data model: mathematical representation of data
  - parametric / nonparametric
  - static / dynamic
  - probabilistic / deterministic
  - example: $p_x(x|\mu,\sigma) = Ke^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Maximum-likelihood estimation (1/3)

- select data model parameter values that maximize the probability of the observed data
- $\hat{\theta}_{\mathsf{ML}} = \arg\max_{\theta} \mathsf{P}\left(x(1), x(2), ..., x(T) \,|\, \theta\right)$
- continuous data: solve the likelihood equation

$$\frac{\partial}{\partial \theta} \ln p_{x(1),...,x(T)}\left(x(1), ..., x(T)|\theta\right)\Big|_{\theta=\hat{\theta}_{\mathsf{ML}}} = 0$$

Maximum-likelihood estimation (2/3)

- when samples are independent

$$p_{x(1),...,x(T)}(x(1),...,x(T)|\theta) = \prod_{i=1}^{T} p_{x(i)}(x(i)|\theta)$$

- example: mean of a Gaussian

- $p_{x(1),...,x(T)}(x(1),...,x(T)|\mu) = K^T \prod_{i=1}^{T} e^{-\frac{(x(i)-\mu)^2}{2\sigma^2}}$

Maximum-likelihood estimation (3/3)

$$
\begin{aligned}
\frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{T}\left(x(i)-\mu\right)^2\right]\Bigg|_{\mu=\mu_{\mathsf{ML}}} &= 0 \\
\sum_{i=1}^{T}\left(x(i)-\mu_{\mathsf{ML}}\right) &= 0 \\
\mu_{\mathsf{ML}} &= \frac{1}{T}\sum_{i=1}^{T} x(i)
\end{aligned}
$$

Tomorrow

ICA in images and video (and equations...)