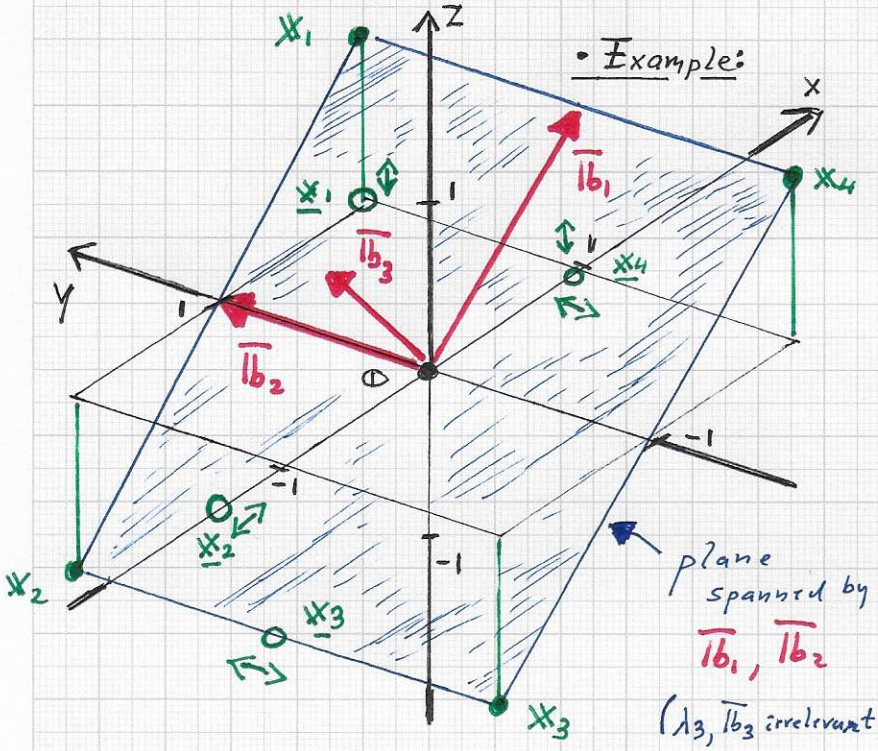
Stratovan

DIMENSION REDUCTION - Cont'd.

- **Missing data II:** A given sample data set $\{\underline{x}_i\}$ includes several data with missing coordinate values. An approximation of a PCA basis $\{\overline{b}_j\}$ can be constructed and improved iteratively.



- **Example:**

→ complete data set:

$$\{\underline{x}_i\} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \right\}$$

→ introduce missing values, '?':

$$\{\underline{x}_i\} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \right\}$$

→ PCA for complete data set:

$$C = \begin{pmatrix} 4 & 0 & 4 \\ 0 & 4 & 0 \\ 4 & 0 & 4 \end{pmatrix};$$

$\lambda_1 = 8,\ \lambda_2 = 4,\ \lambda_3 = 0;$

$$\overline{b}_1 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix},\ \overline{b}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},\ \overline{b}_3 = \begin{pmatrix} -\sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix}$$

plane spanned by $\overline{b}_1, \overline{b}_2$

($\lambda_3, \overline{b}_3$ irrelevant)

→ **step 1:** estimate a center $\mathbb{C}$ for the data set with missing values; i.e., compute averages of the known values, coordinate-wise:

$$\mathbb{C} \approx \begin{pmatrix} 1/3\ (1-1+1) \\ 1/2\ (1+1) \\ 1/2\ (-1-1) \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1 \\ -1 \end{pmatrix}$$

→ **step 2:** use the coordinate values of $\mathbb{C}$ to replace the '?' values; i.e., "repair" the $\underline{x}_i$ data with missing values using values of $\mathbb{C}$:

$$\underline{x}_1 \approx \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix},\ \underline{x}_2 \approx \begin{pmatrix} 1/3 \\ -1 \\ -1 \end{pmatrix},\ \underline{x}_3 \approx \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix},\ \underline{x}_4 \approx \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$$

■ DIMENSION REDUCTION - Cont'd.

• **Missing data II: Example continued...**

→ **step 3: compute a PCA-based orthonormal basis $\{\overline{lb}_j\}$;**

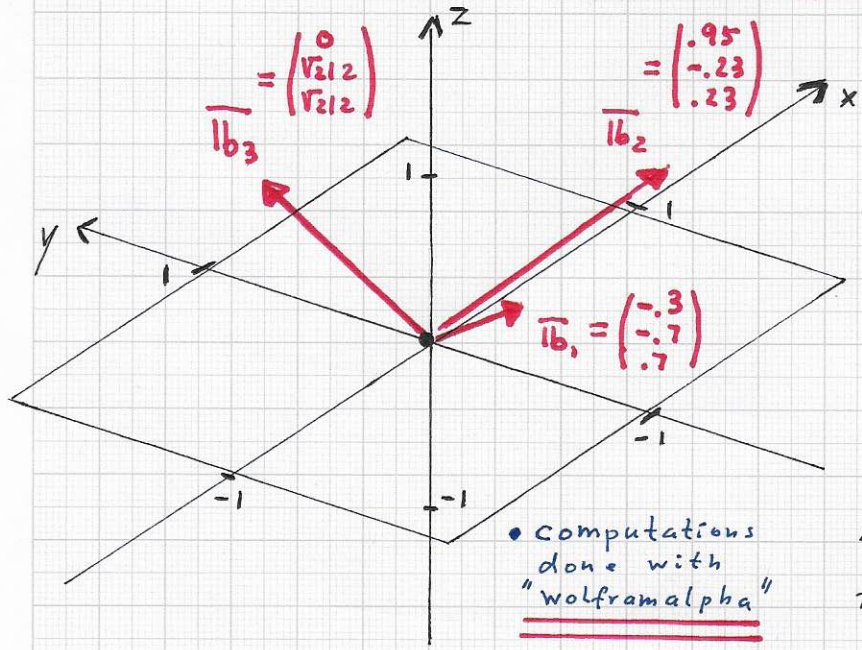
i.e., use the current estimates for $\{\underline{x}_i\}$ to perform PCA:

$$C = \begin{pmatrix} 3\frac{1}{9} & 1\frac{1}{3} & -1\frac{1}{3} \\ 1\frac{1}{3} & 4 & -4 \\ -1\frac{1}{3} & -4 & 4 \end{pmatrix}; \quad \lambda_1 = \frac{2}{9}\left(25+\sqrt{193}\right) \approx 8.64, \quad \lambda_2 = \frac{2}{9}\left(25-\sqrt{193}\right) \approx 2.47, \quad \lambda_3 = 0;$$

$\Rightarrow$ not normalized eigenvectors: $\mathcal{E}_1 = \begin{pmatrix} \frac{1}{6}\left(11-\sqrt{193}\right) \\ -1 \\ 1 \end{pmatrix} \approx \ldots$,

$\mathcal{E}_2 = \begin{pmatrix} \frac{1}{6}\left(11+\sqrt{193}\right) \\ -1 \\ 1 \end{pmatrix} \approx \ldots$, $\mathcal{E}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$,

$\Rightarrow$ orthonormal basis vectors: $\overline{lb}_1 \approx \begin{pmatrix} -.32 \\ -.67 \\ .67 \end{pmatrix}$, $\overline{lb}_2 = \begin{pmatrix} .95 \\ -.23 \\ .23 \end{pmatrix}$, $\overline{lb}_3 = \begin{pmatrix} 0 \\ \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$



• computations done with "wolframalpha"

→ **step 4: use the current $\overline{lb}_j$ estimates to perform best approximation of the given $\{\underline{x}_i\}$ to improve the estimates for all '?' coord. values;**

i.e., perform the method for "Missing data I" for all $\underline{x}_i$ data with '?' values:

$\Rightarrow$ best approx. of $\underline{x}_1$: $\left(\langle \overline{lb}_i, \overline{lb}_j \rangle_\bullet\right)\begin{pmatrix} \overline{x}_1 \\ \overline{y}_1 \\ \overline{z}_1 \end{pmatrix} = \left(\langle \underline{x}_1, \overline{lb}_i \rangle_\bullet\right)$, with '•' indicating that only the first two coords. are used.

Stratovan

DIMENSION REDUCTION - Cont'd.

- **Missing data II**: Example continued...

... $\Rightarrow$ best approx. of $\underline{x}_1$:

$$\begin{pmatrix} \langle \overline{b}_1, \overline{b}_1 \rangle_\bullet & \dots & \langle \overline{b}_1, \overline{b}_3 \rangle_\bullet \\ \vdots & & \\ \langle \overline{b}_3, \overline{b}_1 \rangle_\bullet & \dots & \langle \overline{b}_3, \overline{b}_3 \rangle_\bullet \end{pmatrix} \begin{pmatrix} \overline{x}_1 \\ \overline{y}_1 \\ \overline{z}_1 \end{pmatrix} = \begin{pmatrix} \langle \underline{x}_1, \overline{b}_1 \rangle_\bullet \\ \vdots \\ \langle \underline{x}_1, \overline{b}_3 \rangle_\bullet \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} .5513 & -.1499 & -.4738 \\ -.1499 & .9554 & -.1626 \\ -.4738 & -.1626 & .5 \end{pmatrix} \begin{pmatrix} \overline{x}_1 \\ \overline{y}_1 \\ \overline{z}_1 \end{pmatrix} = \begin{pmatrix} -.99 \\ .72 \\ .7071 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \overline{x}_1 \\ \overline{y}_1 \\ \overline{z}_1 \end{pmatrix} = \begin{pmatrix} -.0261 \\ 1.0438 \\ 1.7289 \end{pmatrix}$$

$$\Rightarrow \underline{x}_1 = -.0261 \begin{pmatrix} -.32 \\ -.67 \\ .67 \end{pmatrix} + 1.0438 \begin{pmatrix} .95 \\ -.23 \\ .23 \end{pmatrix} + 1.7289 \begin{pmatrix} 0 \\ \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1.445 \end{pmatrix}$$

(only computation of third coord. value, 1.445, needed!)

$\Rightarrow$ best approx. of $\underline{x}_2, \underline{x}_3, \underline{x}_4$: perform same computations!

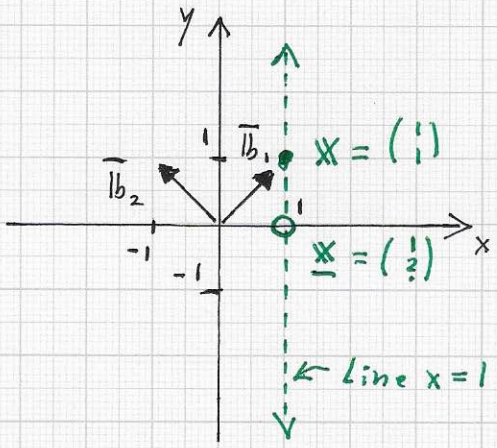
➔ **Iteration**:
- perform step 3, using $\{\underline{x}_i\}$ to improve $\{\overline{b}_j\}$;
- perform step 4, using $\{\overline{b}_j\}$ to best-approximate $\{\underline{x}_i\}$;
- stop when orthonormal basis $\{\overline{b}_j\}$ has "converged."

- **Note**: It is unclear whether one should estimate a new data set center $\mathbb{C}$ and perform mean-subtraction after every computation of updated $\underline{x}_i$-values. Intuitively, it seems to make sense and be desirable to do this, since the mean indeed changes after every iteration. ∴ BH

- **Reference**: Everson and Sirovich, The Karhunen-Loève transform for incomplete data, J. Opt. Soc. Am., A, pp. 1657-1664, 1995.

■ DIMENSION REDUCTION - Cont'd.

• Missing data I and missing data II: simple examples



"Computing only missing ?-values when basis $\{\overline{b}_i\}$ is given, known"

• Ex.: missing data I

→ known orthonormal basis: $\overline{b}_1 = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$, $\overline{b}_2 = \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$

→ datum $\underline{x} = \begin{pmatrix} 1 \\ ? \end{pmatrix}$ misses y-coord. value

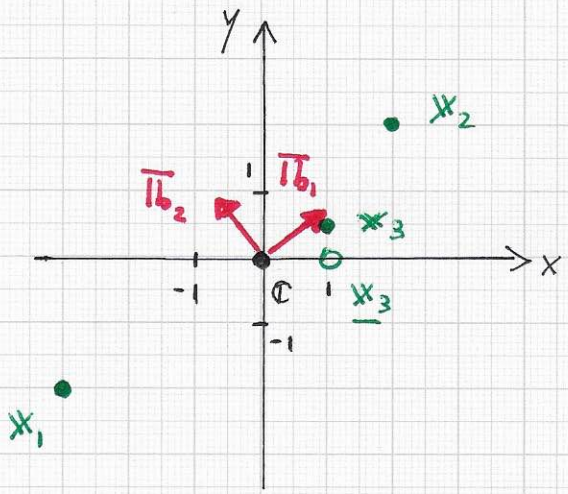
→ perform best approximation, using only the x-values for inner products:

$$\begin{pmatrix} \langle \overline{b}_1, \overline{b}_1 \rangle_\bullet & \langle \overline{b}_1, \overline{b}_2 \rangle_\bullet \\ \langle \overline{b}_2, \overline{b}_1 \rangle_\bullet & \langle \overline{b}_2, \overline{b}_2 \rangle_\bullet \end{pmatrix} \begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix} = \begin{pmatrix} \langle \underline{x}, \overline{b}_1 \rangle_\bullet \\ \langle \underline{x}, \overline{b}_2 \rangle_\bullet \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}$$

$\Rightarrow$ solution family: $\overline{y} = \sqrt{2} - \overline{x}$

$$\Rightarrow \underline{x} = \overline{x} \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} + (\sqrt{2} - \overline{x}) \begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} = \begin{pmatrix} \overline{x}\sqrt{2} - 1 \\ 1 \end{pmatrix} \overset{x=1}{=} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$



"Iterative computation of unknown basis vectors and $\underline{x}_i$ data with ?-values"

• Ex.: missing data II

→ orthonormal basis NOT known

→ given data: $\underline{x}_1 = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$, $\underline{x}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\underline{x}_3 = \begin{pmatrix} 1 \\ ? \end{pmatrix}$

→ Center (mean): $C = \begin{pmatrix} 1/3 \cdot 0 \\ 1/2 \cdot 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

($\Rightarrow$ subsequent mean-subtraction step would not change data!)

→ replace ?-value by mean-value:

$\underline{x}_1 = \begin{pmatrix} -3 \\ -3 \end{pmatrix}$, $\underline{x}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\underline{x}_3 = \begin{pmatrix} 1 \\ 6 \end{pmatrix}$

■ DIMENSION REDUCTION – Cont'd.

• Ex.: missing data II ...

→ first approximation of orthonormal basis:

$$C = \begin{pmatrix} -3 & 2 & 1 \\ -3 & 2 & 0 \end{pmatrix} \begin{pmatrix} -3 & -3 \\ 2 & 2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 14 & 10 \\ 10 & 8 \end{pmatrix} \Rightarrow \lambda_1 = 11 + \sqrt{109} = 21.44$$

$$\lambda_2 = 11 - \sqrt{109} = 0.56$$

$$\Rightarrow \overline{b_1} \approx \begin{pmatrix} .8 \\ .6 \end{pmatrix}, \overline{b_2} \approx \begin{pmatrix} -.6 \\ .8 \end{pmatrix}$$

→ HERE: $\lambda_1 \gg \lambda_2 \Rightarrow$ $\overline{b_1}$ "important," $\overline{b_2}$ "irrelevant"!

→ compute next approximation for ?-value of $\underline{x}_3$:
perform inner product computations only for x-coordinate, considering basis vector $\overline{b_1}$ only:

$$\begin{pmatrix} \langle \overline{b_1}, \overline{b_1} \rangle_\bullet \end{pmatrix} \begin{pmatrix} \bar{x} \end{pmatrix} = \begin{pmatrix} \langle \underline{x}_3, \overline{b_1} \rangle_\bullet \end{pmatrix}$$

$$\Leftrightarrow \quad .64\,\bar{x} = 1 \cdot .8$$

$$\Leftrightarrow \quad \bar{x} = .8$$

$$\Rightarrow \quad \underline{x}_3 = \bar{x}\,\overline{b_1} = .8 \begin{pmatrix} .8 \\ .6 \end{pmatrix} = \begin{pmatrix} .64 \\ .48 \end{pmatrix}$$

this x-value is NOT considered; it is given as x=1

$$\overset{x=1}{=} \begin{pmatrix} 1 \\ .48 \end{pmatrix}$$

→ perform iterative improvement of basis vectors and originally unknown values for '?'.

∴ BH