

Ganganalyse

Modellering og estimation

Klaus Kähler Holst

5. Januar 2006

- 1 Introduktion
- 2 Model for ledvinkelsrotation
- 3 PCA
- 4 Perspektivering

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Motivation:

- Medicinske anvendelser.

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Motivation:

- Medicinske anvendelser.
 - Hjælp til diagnostisering af visse sygdomme.

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Motivation:

- Medicinske anvendelser.
 - Hjælp til diagnostisering af visse sygdomme.
 - Har en behandling en effekt?

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Motivation:

- Medicinske anvendelser.
 - Hjælp til diagnostisering af visse sygdomme.
 - Har en behandling en effekt?
 - Forbedring af bevægelsesmønstre.

Ganganalyse er studiet af hvordan et menneske bevæger sig.

Motivation:

- Medicinske anvendelser.
 - Hjælp til diagnostisering af visse sygdomme.
 - Har en behandling en effekt?
 - Forbedring af bevægelsesmønstre.
- Biometri: Identifikation og verifikation baseret på gang.
 - En stor fordel i forhold til eksisterende biometriske metoder er, at identifikationen kan ske på lang afstand.

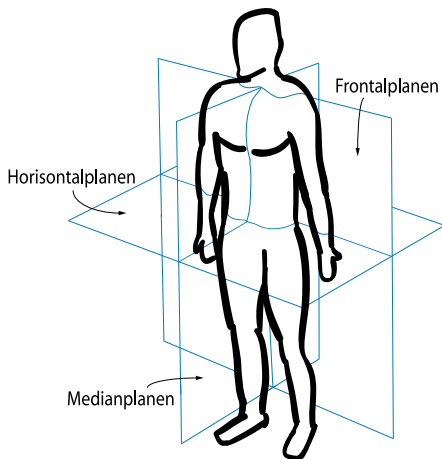
Hvad er menneskelig gang?

Menneskelig gang kan ses som translationer og rotationer af forskellige led.

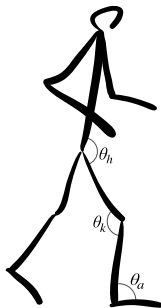
Hvad er menneskelig gang?

Menneskelig gang kan ses som translationer og rotationer af forskellige led.

For at beskrive vinkelrotationerne benytter vi *anatomiske vinkler*:

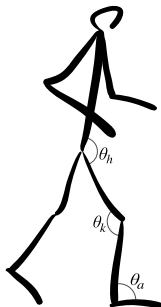


Vi begrænser os til at betragte bevægelse i medianplanen af hofte, knæ og ankel.



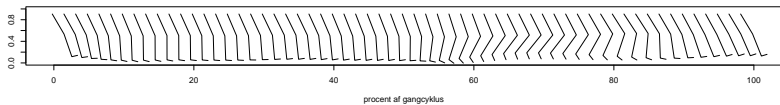
For knæ- og hoftelæddet taler vi om *fleksion/ekstension* og for ankelleddet *dorsifleksion/plantarfleksion*.

Vi begrænser os til at betragte bevægelse i medianplanen af hofte, knæ og ankel.

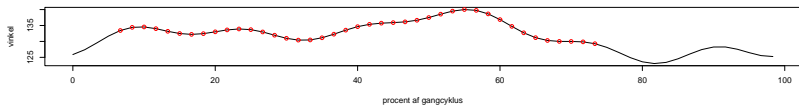


For knæ- og hoftelæddet taler vi om *fleksion/ekstension* og for ankelleddet *dorsifleksion/plantarfleksion*.

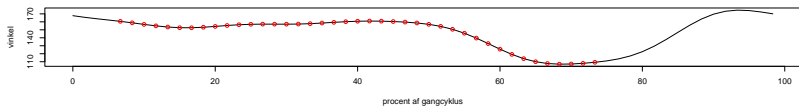
En *gangcyklus* er intervallet fra venstre hæl slipper kontakten med gulvet, og til hælen igen slipper kontakt. *Standfasen* er perioden hvor foden har kontakt med jorden (ca. 3/5 af cyklusen) og *svingfasen*.



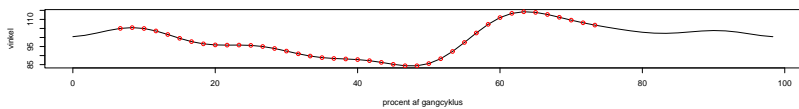
Hofte

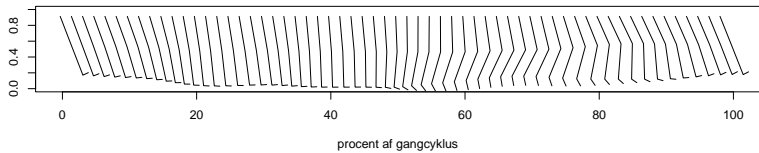
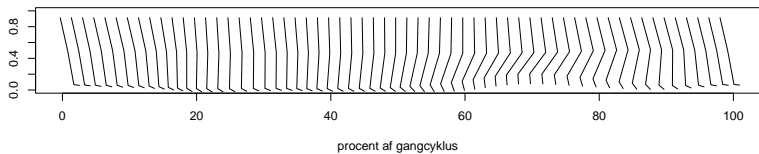
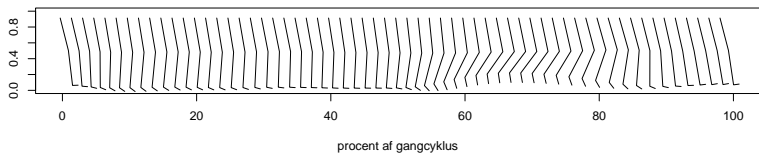


Knæ

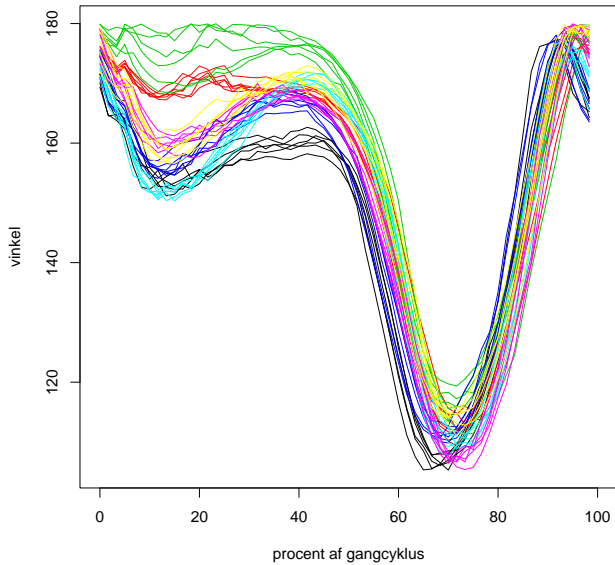


Ankel





Rotationsvinkler for knæ



Model for ledvinkel i en gangcyklus

Model for ledvinkel i en gangcyklus

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Model for ledvinkel i en gangcyklus

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Målefejlen beskrives ved en additiv model

$$\mathbf{y} = \mathbf{a} + \varepsilon$$

hvor \mathbf{a} er de "sande" værdier og ε er målefejlen.

Model for ledvinkel i en gangcyklus

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Målefejlen beskrives ved en additiv model

$$\mathbf{y} = \mathbf{a} + \varepsilon$$

hvor \mathbf{a} er de "sande" værdier og ε er målefejlen.

Bemærk at \mathbf{a} naturligt kan opfattes som observationer fra en \mathcal{C}^2 -funktion: f .

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Målefejlen beskrives ved en additiv model

$$\mathbf{y} = \mathbf{a} + \varepsilon$$

hvor \mathbf{a} er de "sande" værdier og ε er målefejlen.

Bemærk at \mathbf{a} naturligt kan opfattes som observationer fra en \mathcal{C}^2 -funktion: f .

Metoder til at separere de sande værdier fra støjen:

- Lineær filtrering

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Målefejlen beskrives ved en additiv model

$$\mathbf{y} = \mathbf{a} + \varepsilon$$

hvor \mathbf{a} er de "sande" værdier og ε er målefejlen.

Bemærk at \mathbf{a} naturligt kan opfattes som observationer fra en \mathcal{C}^2 -funktion: f .

Metoder til at separere de sande værdier fra støjen:

- Lineær filtrering
- Repræsentation vha ONB

Data fra en person for et led over en gangcyklus:

$$\mathbf{y} = (y_1, \dots, y_d)$$

Målefejlen beskrives ved en additiv model

$$\mathbf{y} = \mathbf{a} + \varepsilon$$

hvor \mathbf{a} er de "sande" værdier og ε er målefejlen.

Bemærk at \mathbf{a} naturligt kan opfattes som observationer fra en \mathcal{C}^2 -funktion: f .

Metoder til at separere de sande værdier fra støjen:

- Lineær filtrering
- Repræsentation vha ONB
- ...

Finde ONB, $(\varphi_i)_{i=1}^\infty$, så

$$f(t) = \sum_{k=1}^K \beta_k \varphi_k(t), \quad t \in [0, 1].$$

Finde ONB, $(\varphi_i)_{i=1}^{\infty}$, så

$$f(t) = \sum_{k=1}^K \beta_k \varphi_k(t), \quad t \in [0, 1].$$

Egenskaber for f

- Approksimativt periodisk.
- $\mathcal{C}^2([0, 1])$.
- Frekvenskomponenter i intervallet $[0, 6]$.

Finde ONB, $(\varphi_i)_{i=1}^{\infty}$, så

$$f(t) = \sum_{k=1}^K \beta_k \varphi_k(t), \quad t \in [0, 1].$$

Egenskaber for f

- Approksimativt periodisk.
- $\mathcal{C}^2([0, 1])$.
- Frekvenskomponenter i intervallet $[0, 6]$.

Oplagt kandidat: *Fourierrækker*.

Vi kigger altså på basisen

$$\{1, \cos(2\pi x), \sin(2\pi x), \cos(2\pi 2x), \sin(2\pi 2x), \dots\}.$$

Vi definerer matricen

$$\mathbf{A} = \begin{pmatrix} 1 & \cos(2\pi t_1) & \sin(2\pi t_1) & \cdots & \cos(2\pi Nt_1) & \sin(2\pi Nt_1) \\ \vdots & & & \vdots & & \vdots \\ 1 & \cos(2\pi t_d) & \sin(2\pi t_d) & \cdots & \cos(2\pi Nt_d) & \sin(2\pi Nt_d) \end{pmatrix}.$$

Antagelsen er at

$$\mathbf{a} = \mathbf{A}\boldsymbol{\lambda}$$

Vi definerer matricen

$$\mathbf{A} = \begin{pmatrix} 1 & \cos(2\pi t_1) & \sin(2\pi t_1) & \cdots & \cos(2\pi Nt_1) & \sin(2\pi Nt_1) \\ \vdots & & & \vdots & & \vdots \\ 1 & \cos(2\pi t_d) & \sin(2\pi t_d) & \cdots & \cos(2\pi Nt_d) & \sin(2\pi Nt_d) \end{pmatrix}.$$

Antagelsen er at

$$\mathbf{a} = \mathbf{A}\boldsymbol{\lambda}$$

og $\boldsymbol{\lambda}$ kan estimeres ved OLS:

$$\hat{\boldsymbol{\lambda}} = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \mathbf{y} = \boldsymbol{\lambda} + \text{diag}\left(\frac{1}{d}, \frac{2}{d}, \dots, \frac{2}{d}\right) \mathbf{A}^t \boldsymbol{\varepsilon}.$$

Når vi observerer flere forskellige individer, har vi observationer

$$\mathbf{y}_1, \dots, \mathbf{y}_n$$

som hver antages at opfylde modellen

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i.$$

Når vi observerer flere forskellige individer, har vi observationer

$$\mathbf{y}_1, \dots, \mathbf{y}_n$$

som hver antages at opfylde modellen

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i.$$

- $\boldsymbol{\lambda}_i$ 'erne kan nu opfattes som stokastiske, og vi antager, at $\boldsymbol{\lambda}_i$ 'erne er iid og uafhængige af målefejlene $\boldsymbol{\varepsilon}_i$.

Når vi observerer flere forskellige individer, har vi observationer

$$\mathbf{y}_1, \dots, \mathbf{y}_n$$

som hver antages at opfylde modellen

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i.$$

- $\boldsymbol{\lambda}_i$ 'erne kan nu opfattes som stokastiske, og vi antager, at $\boldsymbol{\lambda}_i$ 'erne er iid og uafhængige af målefejlene $\boldsymbol{\varepsilon}_i$. Vi antager at

$$\boldsymbol{\lambda}_i \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}).$$

Når vi observerer flere forskellige individer, har vi observationer

$$\mathbf{y}_1, \dots, \mathbf{y}_n$$

som hver antages at opfylde modellen

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i.$$

- $\boldsymbol{\lambda}_i$ 'erne kan nu opfattes som stokastiske, og vi antager, at $\boldsymbol{\lambda}_i$ 'erne er iid og uafhængige af målefejlene $\boldsymbol{\varepsilon}_i$. Vi antager at

$$\boldsymbol{\lambda}_i \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}).$$

- Indledende modelkontrol giver ikke anledning til at tro på at \mathbf{y}_i 'erne eller $\boldsymbol{\lambda}_i$ 'erne følger normalfordelinger.

På *mixed model* form

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\beta}$: faste virkninger.

\mathbf{u}_i : iid tilfældige virkninger.

På *mixed model* form

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\beta}$: faste virkninger.

\mathbf{u}_i : iid tilfældige virkninger.

$$\mathbf{u}_i \sim (0, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}). \quad (3.1)$$

$$\mathbf{u}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_i \quad (3.2)$$

På *mixed model* form

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\beta}$: faste virkninger.

\mathbf{u}_i : iid tilfældige virkninger.

$$\mathbf{u}_i \sim (0, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}). \quad (3.1)$$

$$\mathbf{u}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_i \quad (3.2)$$

$$\mathbf{V} = \text{var}(\mathbf{y}_i) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2 \mathbf{I}.$$

På *mixed model* form

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\beta}$: faste virkninger.

\mathbf{u}_i : iid tilfældige virkninger.

$$\mathbf{u}_i \sim (0, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}). \quad (3.1)$$

$$\mathbf{u}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_i \quad (3.2)$$

$$\mathbf{V} = \text{var}(\mathbf{y}_i) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2 \mathbf{I}.$$

Estimation af de faste virkninger:

$$\hat{\boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \mathbf{y}_i$$

På *mixed model* form

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\beta} + \mathbf{A}\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

$\boldsymbol{\beta}$: faste virkninger.

\mathbf{u}_i : iid tilfældige virkninger.

$$\mathbf{u}_i \sim (0, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim (0, \sigma^2 \mathbf{I}). \quad (3.1)$$

$$\mathbf{u}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_i \quad (3.2)$$

$$\mathbf{V} = \text{var}(\mathbf{y}_i) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2 \mathbf{I}.$$

Estimation af de faste virkninger:

$$\hat{\boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^n (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \mathbf{y}_i$$

Assymptotisk har vi

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{\Sigma} + \sigma^2(\mathbf{A}^t \mathbf{A})^{-1}), \quad n \rightarrow \infty.$$

For at estimere σ^2 beregnes

$$\hat{\beta}_i = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \mathbf{y}_i$$

og

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{A}\hat{\beta}_i.$$

Vi definere estimatet

$$\hat{\sigma}_{(i)}^2 = \frac{1}{d-p} \mathbf{e}_i^t \mathbf{e}_i$$

For at estimere σ^2 beregnes

$$\hat{\beta}_i = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \mathbf{y}_i$$

og

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{A}\hat{\beta}_i.$$

Vi definere estimatet

$$\hat{\sigma}_{(i)}^2 = \frac{1}{d-p} \mathbf{e}_i^t \mathbf{e}_i$$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_{(i)}^2$$

Variansen af \mathbf{y}_i 'erne estimeres ved:

$$\widehat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})^t$$

som er konsistent

$$\widehat{\mathbf{V}}_n \xrightarrow{P} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2\mathbf{I}, \quad n \rightarrow \infty.$$

Variansen af \mathbf{y}_i 'erne estimeres ved:

$$\widehat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})^t$$

som er konsistent

$$\widehat{\mathbf{V}}_n \xrightarrow{P} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2\mathbf{I}, \quad n \rightarrow \infty.$$

Endelig kan vi estimere variansen af \mathbf{u}_i 'erne:

$$\widehat{\boldsymbol{\Sigma}} = (\mathbf{A}^t\mathbf{A})^{-1} \mathbf{A}^t(\widehat{\mathbf{V}}_n - \widehat{\sigma}_n^2\mathbf{I})\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}$$

som også er konsistent.

Variansen af \mathbf{y}_i 'erne estimeres ved:

$$\widehat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})(\mathbf{y}_i - \mathbf{A}\widehat{\boldsymbol{\beta}})^t$$

som er konsistent

$$\widehat{\mathbf{V}}_n \xrightarrow{P} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t + \sigma^2\mathbf{I}, \quad n \rightarrow \infty.$$

Endelig kan vi estimere variansen af \mathbf{u}_i 'erne:

$$\widehat{\boldsymbol{\Sigma}} = (\mathbf{A}^t\mathbf{A})^{-1} \mathbf{A}^t(\widehat{\mathbf{V}}_n - \widehat{\sigma}_n^2\mathbf{I})\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}$$

som også er konsistent.

Hvis $\widehat{\boldsymbol{\Sigma}}$ ikke positiv semi-definit, kan vi projicere ned på den nærmeste symmetriske, positiv semidefinitte matrix.

Prediktion af de tilfældige virkninger

Best Predictor (BP):

$$\hat{\mathbf{u}}_i^{(BP)} = E(\mathbf{u}_i | \mathbf{y}_i)$$

Best Predictor (BP):

$$\hat{\mathbf{u}}_i^{(\text{BP})} = E(\mathbf{u}_i \mid \mathbf{y}_i)$$

Best Linear Unbiased Predictor (BLUP):

$$\hat{\mathbf{u}}_i^{(\text{BLUP})} = \mathbf{\Sigma} \mathbf{A}^t \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{A} \boldsymbol{\beta})$$

Best Predictor (BP):

$$\hat{\mathbf{u}}_i^{(\text{BP})} = E(\mathbf{u}_i \mid \mathbf{y}_i)$$

Best Linear Unbiased Predictor (BLUP):

$$\hat{\mathbf{u}}_i^{(\text{BLUP})} = \mathbf{\Sigma} \mathbf{A}^t \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{A} \boldsymbol{\beta})$$

Estimated Best Linear Unbiased Predictor (EBLUP):

$$\hat{\mathbf{u}}_i = \hat{\mathbf{\Sigma}} \mathbf{A}^t \hat{\mathbf{V}}^{-1} (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}})$$

Best Predictor (BP):

$$\hat{\mathbf{u}}_i^{(\text{BP})} = E(\mathbf{u}_i \mid \mathbf{y}_i)$$

Best Linear Unbiased Predictor (BLUP):

$$\hat{\mathbf{u}}_i^{(\text{BLUP})} = \mathbf{\Sigma} \mathbf{A}^t \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{A} \boldsymbol{\beta})$$

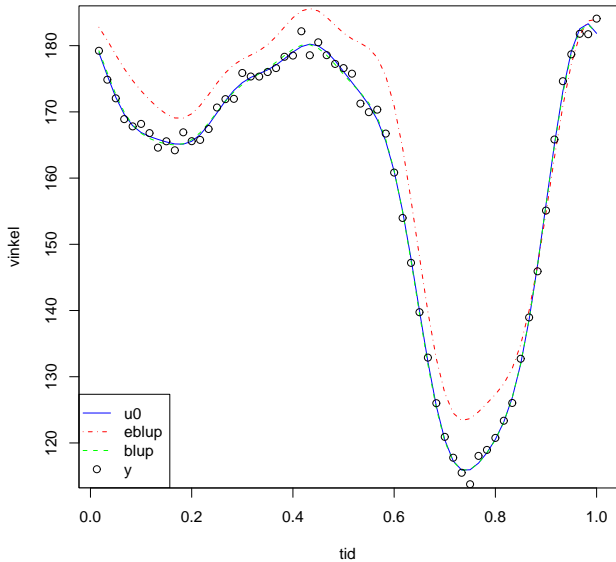
Estimated Best Linear Unbiased Predictor (EBLUP):

$$\hat{\mathbf{u}}_i = \hat{\mathbf{\Sigma}} \mathbf{A}^t \hat{\mathbf{V}}^{-1} (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}})$$

En pragmatisk prediktor:

$$\begin{aligned} \hat{\mathbf{u}}_i^0 &= (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}}) \\ &= \mathbf{u}_i + (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \boldsymbol{\varepsilon}_i - \frac{1}{n} \sum_{j=1}^n (\mathbf{u}_j + (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \boldsymbol{\varepsilon}_j) \end{aligned}$$

10th observation of 100



Konklusion: De tilfældige virkninger kan med fordel predikteres ved

$$\hat{\mathbf{u}}_i^0 = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}})$$

og variansen estimeres ved

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^n \hat{\mathbf{u}}_i^0 \hat{\mathbf{u}}_i^{0t}$$

Konklusion: De tilfældige virkninger kan med fordel predikteres ved

$$\hat{\mathbf{u}}_i^0 = (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}})$$

og variansen estimeres ved

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^n \hat{\mathbf{u}}_i^0 \hat{\mathbf{u}}_i^{0t}$$

Modellen kan udvides til at tage højde for flere gangcykler for hvert individ, ved at benytte en additiv model for de tilfældige virkninger:

$$\mathbf{u}_{ij} = \boldsymbol{\nu}_i + \boldsymbol{\gamma}_{ij}.$$

Principalkomponentanalyse (PCA)

- Reducering i dimension.

Principalkomponentanalyse (PCA)

- Reducering i dimension.
- Fortolkning af variationen i data.

Principalkomponentanalyse (PCA)

- Reducering i dimension.
- Fortolkning af variationen i data.
- Ukorrelerede koordinater.

Principalkomponentanalyse (PCA)

- Reducering i dimension.
- Fortolkning af variationen i data.
- Ukorrelerede koordinater.
- Tegn på at gang kan fungere som biometrisk mål.

Lad $\mathbf{X} = (X_1, \dots, X_n)$ være stokastisk variable med værdier i \mathbb{R}^n og middelværdi 0 og kovariansmatrix

$$\text{var}(\mathbf{X}) = \mathbf{\Sigma}_X = \mathbf{P}_X \sigma(\mathbf{\Sigma}_X) \mathbf{P}_X^t,$$

Lad $\mathbf{X} = (X_1, \dots, X_n)$ være stokastisk variable med værdier i \mathbb{R}^n og middelværdi 0 og kovariansmatrix

$$\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}} \boldsymbol{\sigma}(\boldsymbol{\Sigma}_{\mathbf{X}}) \mathbf{P}_{\mathbf{X}}^t,$$

hvor $\mathbf{P}_{\mathbf{X}} = (\mathbf{e}_1 \ \cdots \ \mathbf{e}_n)$ og $\boldsymbol{\sigma}(\boldsymbol{\Sigma}_{\mathbf{X}}) = \text{diag}(\lambda_1, \dots, \lambda_n)$ og egenverdierne antages at være sorteret: $\lambda_1 \geq \dots \geq \lambda_n \geq 0$.

Lad $\mathbf{X} = (X_1, \dots, X_n)$ være stokastisk variable med værdier i \mathbb{R}^n og middelværdi 0 og kovariansmatrix

$$\text{var}(\mathbf{X}) = \mathbf{\Sigma}_X = \mathbf{P}_X \sigma(\mathbf{\Sigma}_X) \mathbf{P}_X^t,$$

hvor $\mathbf{P}_X = (\mathbf{e}_1 \ \cdots \ \mathbf{e}_n)$ og $\sigma(\mathbf{\Sigma}_X) = \text{diag}(\lambda_1, \dots, \lambda_n)$ og egenverdierne antages at være sorteret: $\lambda_1 \geq \dots \geq \lambda_n \geq 0$.

Definition

Principalkomponenterne hørende til \mathbf{X} er de stokastiske variable

$$\mathbf{Y} = (Y_1 \ \cdots \ Y_n)^t = \mathbf{P}_X^t \mathbf{X}$$

$$\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathbf{X}})$$

$$\mathbf{P}_{\mathbf{X}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\sigma(\boldsymbol{\Sigma}_{\mathbf{X}}) = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}.$$

Dvs. at

$$\boldsymbol{\Sigma}_{\mathbf{X}} = \begin{pmatrix} \frac{5}{2} & \frac{3}{2} \\ \frac{3}{2} & \frac{5}{2} \end{pmatrix}.$$

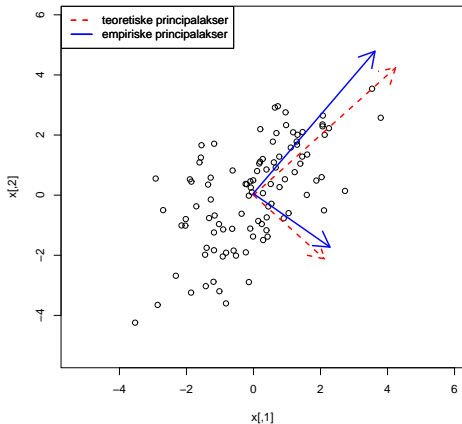
$$\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_X)$$

$$\mathbf{P}_X = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\sigma(\boldsymbol{\Sigma}_X) = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}.$$

Dvs. at

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}.$$



Sætning (Karhunen-Loève - diskret version)

For principalkomponenterne, $\mathbf{Y} = \mathbf{P}_X^t \mathbf{X}$, gælder at

- $\text{var}(\mathbf{Y}) = \sigma(\boldsymbol{\Sigma}_X)$.
- $\text{tr}[\text{var}(\mathbf{Y})] = \text{tr}[\text{var}(\mathbf{X})] = \text{tr}(\boldsymbol{\Sigma}_X)$.
- $\mathbf{X} = \mathbf{P}_X \mathbf{Y}$.

Vi kan *rekonstruere* \mathbf{X} vha. de første $k < n$ egenvektorer. Vi sætter

$$\mathbf{P}_X^k = (\mathbf{e}_1 \quad \cdots \quad \mathbf{e}_k \quad 0 \cdots 0,)$$

og lader $\mathbf{Y}_k = \mathbf{P}_X^{k,t} \mathbf{X}$. Approksimationen er så

$$\mathbf{X}_k = \mathbf{P}_X^k \mathbf{Y}_k$$

Vi kan *rekonstruere* \mathbf{X} vha. de første $k < n$ egenvektorer. Vi sætter

$$\mathbf{P}_X^k = (\mathbf{e}_1 \quad \cdots \quad \mathbf{e}_k \quad 0 \cdots 0),$$

og lader $\mathbf{Y}_k = \mathbf{P}_X^k{}^t \mathbf{X}$. Approksimationen er så

$$\mathbf{X}_k = \mathbf{P}_X^k \mathbf{Y}_k$$

Sætning (Approksimation vha. PCA)

Forholdet mellem den totale variation af \mathbf{X}_k og \mathbf{X} er

$$\rho_k = \frac{\text{tr}[\text{var}(\mathbf{X}_k)]}{\text{tr}[\text{var}(\mathbf{X})]} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}$$

og middel-kvadratfejlen er givet ved

$$\varepsilon(\mathbf{P}_X) = E[(\mathbf{X} - \mathbf{X}_k)^t (\mathbf{X} - \mathbf{X}_k)] = \sum_{i=k+1}^n \lambda_i.$$

Dette er den bedste approksimation: $\varepsilon(\Phi) \geq \sum_{i=k+1}^n \lambda_i$ for alle ONB, Φ .

Lad $(X_t)_{t \in [0,1]}$ være stokastisk proces på baggrundsrummet (Ω, \mathcal{F}, P) , med $t \mapsto X_t(\omega) \in L^2([0, 1])$. Desuden antages

$$\mu_t = \mathbb{E} X_t = \int_{\Omega} X_t dP < \infty,$$

$$V(s, t) = \text{cov}(X_t, X_s) = \int_{\Omega} (X_t - \mu_t)(X_s - \mu_s) dP < \infty, \quad s, t \in [0, 1]$$

og at V er kontinuert.

Lad $(X_t)_{t \in [0,1]}$ være stokastisk proces på baggrundsrummet (Ω, \mathcal{F}, P) , med $t \mapsto X_t(\omega) \in L^2([0, 1])$. Desuden antages

$$\mu_t = \mathbb{E} X_t = \int_{\Omega} X_t dP < \infty,$$

$$V(s, t) = \text{cov}(X_t, X_s) = \int_{\Omega} (X_t - \mu_t)(X_s - \mu_s) dP < \infty, \quad s, t \in [0, 1]$$

og at V er kontinuert. *Autokovariansfunktionen* sættes til

$$A_V f(t) = \int_I f(u) V(u, t) du, \quad f \in L^2(I). \quad (4.1)$$

Lad $(X_t)_{t \in [0,1]}$ være stokastisk proces på baggrundsrummet (Ω, \mathcal{F}, P) , med $t \mapsto X_t(\omega) \in L^2([0, 1])$. Desuden antages

$$\mu_t = \mathbb{E} X_t = \int_{\Omega} X_t dP < \infty,$$

$$V(s, t) = \text{cov}(X_t, X_s) = \int_{\Omega} (X_t - \mu_t)(X_s - \mu_s) dP < \infty, \quad s, t \in [0, 1]$$

og at V er kontinuert. *Autokovariansfunktionen* sættes til

$$A_V f(t) = \int_I f(u) V(u, t) du, \quad f \in L^2(I). \quad (4.1)$$

Spørgsmålet er nu om vi kan opstille resultater analogt til resultaterne vedr. PCA i det endeligt dimensionale tilfælde. Kan A_V diagonaliseres?

Lad $(X_t)_{t \in [0,1]}$ være stokastisk proces på baggrundsrummet (Ω, \mathcal{F}, P) , med $t \mapsto X_t(\omega) \in L^2([0, 1])$. Desuden antages

$$\mu_t = \mathbb{E} X_t = \int_{\Omega} X_t dP < \infty,$$

$$V(s, t) = \text{cov}(X_t, X_s) = \int_{\Omega} (X_t - \mu_t)(X_s - \mu_s) dP < \infty, \quad s, t \in [0, 1]$$

og at V er kontinuert. *Autokovariansfunktionen* sættes til

$$A_V f(t) = \int_I f(u) V(u, t) du, \quad f \in L^2(I). \quad (4.1)$$

Spørgsmålet er nu om vi kan opstille resultater analogt til resultaterne vedr. PCA i det endeligt dimensionale tilfælde. Kan A_V diagonaliseres? **Ja, da A_V er en kompakt normal operator.**

A_V har tælleligt spektrum, der netop er lig egenværdierne. Egenværdierne er ikke-negative og egenfunktionerne er kontinuerte. Lad $\lambda_1, \lambda_2, \dots$ være de positive egenværdier og e_1, e_2, \dots de tilhørende egenfunktioner, da er

$$V(s, t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_i(t) = \varphi(t) \mathbf{\Lambda} \varphi(s)^t$$

A_V har tælleligt spektrum, der netop er lig egenværdierne. Egenværdierne er ikke-negative og egenfunktionerne er kontinuerte. Lad $\lambda_1, \lambda_2, \dots$ være de positive egenværdier og e_1, e_2, \dots de tilhørende egenfunktioner, da er

$$V(s, t) = \sum_{i=1}^{\infty} \lambda_i e_i(s) e_i(t) = \varphi(t) \Lambda \varphi(s)^t$$

Sætning (Karhunen-Loève)

$(X_t)_{t \in [0,1]}$ har dekompositionen

$$X_t \stackrel{L^2}{=} \mu_t + \sum_{i=1}^{\infty} Z_i e_i$$

med

$$Z_i = \int_0^1 (X_t - \mu_t) e_i(t) dt.$$

som er parvist uafhængige.

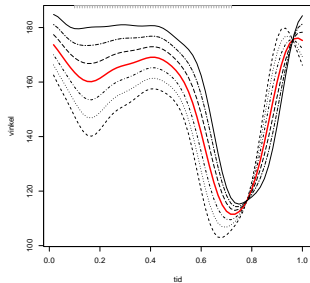
Lad $x_i(t) \in L^2([0, 1])$, $i = 1, \dots, n$, være realisationer af $(X_t)_{t \in [0, 1]}$, med empirisk kovariansfunktion

$$V(s, t) = \frac{1}{n} \sum_{i=1}^n [x_i(s) - \hat{\mu}(t)][x_i(t) - \hat{\mu}(t)]$$

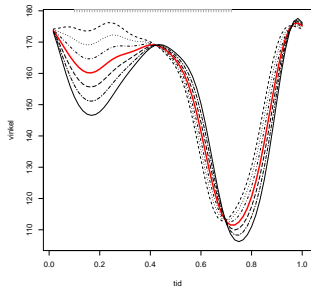
hvor $\hat{\mu}(t) = \overline{x_i(t)}$. Under antagelse at x_i 'erne såvel som egenfunktionerne hørende til V har en endelig fourierekspansion, da findes egenfunktionerne og egenværdierne for A_V ved at finde egenvektorer og egenværdier for den empiriske kovariansmatrix for fourierkoefficienterne.

PCA for gang-data

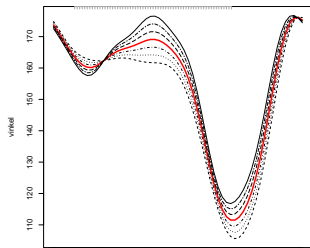
1. principalkomponent (63.5%)



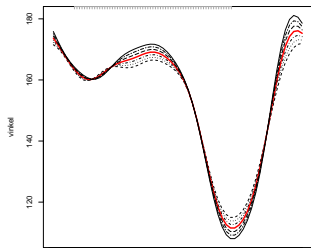
2. principalkomponent (22.8%)



3. principalkomponent (7.5%)

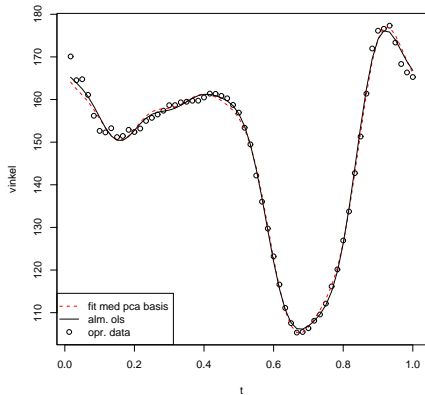


4. principalkomponent (2.4%)



k	ρ_k	λ_k
1	0.635	33.413
2	0.862	11.962
3	0.937	3.922
4	0.961	1.239
5	0.977	0.871
6	0.985	0.412
7	0.992	0.341
8	0.995	0.179
9	0.997	0.1
10	0.998	0.07
11	0.999	0.04
12	0.999	0.018
13	1	0.01

k	ρ_k	λ_k
1	0.635	33.413
2	0.862	11.962
3	0.937	3.922
4	0.961	1.239
5	0.977	0.871
6	0.985	0.412
7	0.992	0.341
8	0.995	0.179
9	0.997	0.1
10	0.998	0.07
11	0.999	0.04
12	0.999	0.018
13	1	0.01



Observationer fra knæ, hofte og ankel:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{y}_h \\ \mathbf{y}_a \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{c}_k \\ \mathbf{A}\mathbf{c}_h \\ \mathbf{A}\mathbf{c}_a \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_h \\ \boldsymbol{\varepsilon}_a \end{pmatrix} .$$

Observationer fra knæ, hofte og ankel:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{y}_h \\ \mathbf{y}_a \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{c}_k \\ \mathbf{A}\mathbf{c}_h \\ \mathbf{A}\mathbf{c}_a \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_h \\ \boldsymbol{\varepsilon}_a \end{pmatrix}.$$

Sættes på en formel

$$\mathbf{y} = \mathbf{A}_3\mathbf{c} + \boldsymbol{\varepsilon}$$

hvor $\mathbf{A}_3 = \mathbf{A} \oplus \mathbf{A} \oplus \mathbf{A}$.

Observationer fra knæ, hofte og ankel:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{y}_h \\ \mathbf{y}_a \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{c}_k \\ \mathbf{A}\mathbf{c}_h \\ \mathbf{A}\mathbf{c}_a \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_h \\ \boldsymbol{\varepsilon}_a \end{pmatrix}.$$

Sættes på en formel

$$\mathbf{y} = \mathbf{A}_3\mathbf{c} + \boldsymbol{\varepsilon}$$

hvor $\mathbf{A}_3 = \mathbf{A} \oplus \mathbf{A} \oplus \mathbf{A}$.

Koefficienterne predikteres ved

$$\hat{\mathbf{c}} = (\mathbf{A}_3^t \mathbf{A}_3)^{-1} \mathbf{A}_3^t \mathbf{y},$$

og de udglattede værdier er så givet ved $\mathbf{A}_3 \hat{\mathbf{c}}$.

Observationer fra knæ, hofte og ankel:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{y}_h \\ \mathbf{y}_a \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{c}_k \\ \mathbf{A}\mathbf{c}_h \\ \mathbf{A}\mathbf{c}_a \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_k \\ \boldsymbol{\varepsilon}_h \\ \boldsymbol{\varepsilon}_a \end{pmatrix}.$$

Sættes på en formel

$$\mathbf{y} = \mathbf{A}_3\mathbf{c} + \boldsymbol{\varepsilon}$$

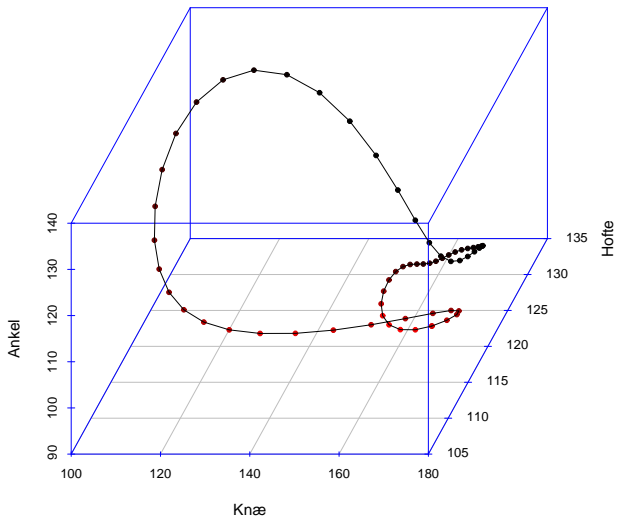
hvor $\mathbf{A}_3 = \mathbf{A} \oplus \mathbf{A} \oplus \mathbf{A}$.

Koefficienterne predikteres ved

$$\hat{\mathbf{c}} = (\mathbf{A}_3^t \mathbf{A}_3)^{-1} \mathbf{A}_3^t \mathbf{y},$$

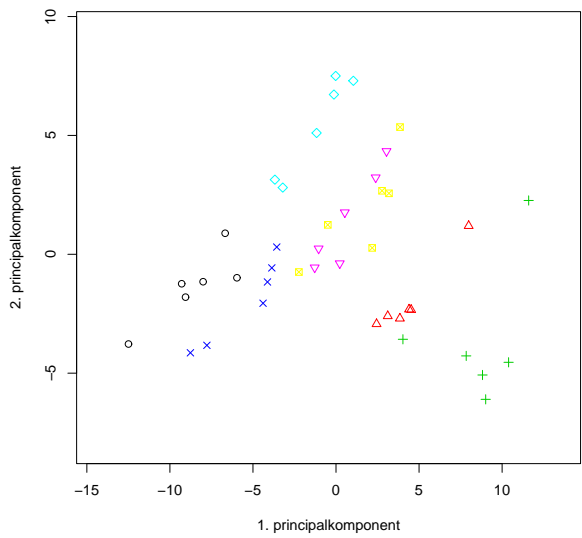
og de udglattede værdier er så givet ved $\mathbf{A}_3 \hat{\mathbf{c}}$.

Funktional PCA foretages nu ved almindelig flerdimensional PCA på $\hat{\mathbf{c}}$ 'erne. Dækningsgraden $p_k \geq 97\%$ opnås med $k = 10$.

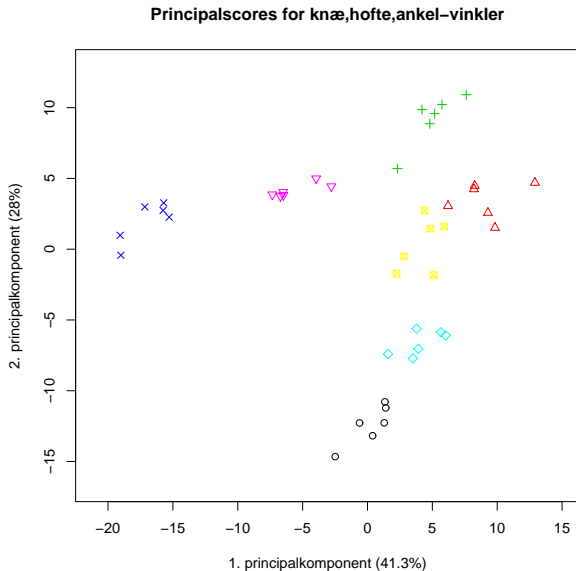


Gang som mulig biometrisk mål?

Gang som mulig biometrisk mål?



Gang som mulig biometrisk mål?



- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.

- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.
- Der er eftervist en pæn sammenhæng mellem flerdimensionale og funktional PCA. Via PCA er der opnået en stor dimensionsreduktion af data.

- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.
- Der er eftervist en pæn sammenhæng mellem flerdimensionale og funktional PCA. Via PCA er der opnået en stor dimensionsreduktion af data.
- Vha. PCA ser vi at gang har potentiale som biometrisk mål.

- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.
- Der er eftervist en pæn sammenhæng mellem flerdimensionale og funktional PCA. Via PCA er der opnået en stor dimensionsreduktion af data.
- Vha. PCA ser vi at gang har potentiale som biometrisk mål.

Bud på fremtidige arbejdsområder:

- Ikke-balancerede data. Mere generelle modeller for målefejlen.

- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.
- Der er eftervist en pæn sammenhæng mellem flerdimensionale og funktional PCA. Via PCA er der opnået en stor dimensionsreduktion af data.
- Vha. PCA ser vi at gang har potentiale som biometrisk mål.

Bud på fremtidige arbejdsområder:

- Ikke-balancerede data. Mere generelle modeller for målefejlen.
- Transformation af data. REML. NPMLE.
- Opstille en decideret teststørrelse for at afgøre om to observationer stammer fra samme individ.
- Diskriminantanalyse.

- Opstillet model for ledvinkels-rotationer med mulighed for at arbejde med flere forskellige gangcykler og observerede led. En ikke-parametriske metode til estimation af varianskomponenterne er foreslået.
- Der er eftervist en pæn sammenhæng mellem flerdimensionale og funktional PCA. Via PCA er der opnået en stor dimensionsreduktion af data.
- Vha. PCA ser vi at gang har potentiale som biometrisk mål.

Bud på fremtidige arbejdsområder:

- Ikke-balancerede data. Mere generelle modeller for målefejlen.
- Transformation af data. REML. NPMLE.
- Opstille en decideret teststørrelse for at afgøre om to observationer stammer fra samme individ.
- Diskriminantanalyse.

<http://www.math.ku.dk/~holst/speciale>



p.36 (prediktion):

$$\begin{aligned}\hat{\mathbf{u}}_i^0 &= (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t (\mathbf{y}_i - \mathbf{A} \hat{\boldsymbol{\beta}}) \\ &= \mathbf{u}_i + (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \boldsymbol{\varepsilon}_i - \frac{1}{n} \sum_{j=1}^n (\mathbf{u}_j + (\mathbf{A}^t \mathbf{A})^{-1} \mathbf{A}^t \boldsymbol{\varepsilon}_j)\end{aligned}$$

p.77 (prediktion):

$$\begin{aligned} y_i - \mathbf{A}\hat{\boldsymbol{\beta}} &= \mathbf{B}\nu_i + \varepsilon_i - \mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\frac{1}{n}\sum_{j=1}^n\mathbf{B}\nu_j \\ &\quad - \mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\frac{1}{n}\sum_{j=1}^n\varepsilon_j \end{aligned}$$

$$\begin{aligned} \hat{\nu}_i - \nu_i &= (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\varepsilon_i - (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\frac{1}{n}\sum_{j=1}^n\mathbf{B}\nu_j \\ &\quad - (\mathbf{B}^t\mathbf{B})^{-1}\mathbf{B}^t\mathbf{A}(\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\frac{1}{n}\sum_{j=1}^n\varepsilon_j \end{aligned}$$