Covariance Estimation for Random Surfaces beyond Separability

May 12, 2023

Tomas Masak
Bernoulli Instructor

EPFL

Functional Data Analysis (FDA)

FDA: Statistical inference on the law of a random process $X(u) : [0,1]^D \to \mathbb{R}$ given its multiple realizations $X_1, \ldots, X_N$.

- $D = 1 \quad \longrightarrow \quad X_1, \ldots, X_N$ are random functions
- $D = 2 \quad \longrightarrow \quad X_1, \ldots, X_N$ are **random surfaces**

FDA utilizes ideas from functional analysis

- *functions* (e.g. realizations) are treated as *points* in a Hilbert space
- *operators* (e.g. covariance) are treated as *functions* on these *points*

Functional Data Analysis (FDA)

FDA: Statistical inference on the law of a random process $X(u) : [0,1]^D \to \mathbb{R}$ given its multiple realizations $X_1, \ldots, X_N$.

- $D = 1 \quad \longrightarrow \quad X_1, \ldots, X_N$ are random functions
- $D = 2 \quad \longrightarrow \quad X_1, \ldots, X_N$ are **random surfaces**

FDA utilizes ideas from functional analysis

- *functions* (e.g. realizations) are treated as *points* in a Hilbert space
- *operators* (e.g. covariance) are treated as *functions* on these *points*

in order to generalize multivariate statistical methods, e.g.

- principal component analysis (PCA)
- linear discriminant analysis (LDA)

Functional Data Analysis (FDA)

FDA: Statistical inference on the law of a random process $X(u) : [0,1]^D \to \mathbb{R}$ given its multiple realizations $X_1, \ldots, X_N$.

- $D = 1 \quad \longrightarrow \quad X_1, \ldots, X_N$ are random functions
- $D = 2 \quad \longrightarrow \quad X_1, \ldots, X_N$ are **random surfaces**

FDA utilizes ideas from functional analysis

- *functions* (e.g. realizations) are treated as *points* in a Hilbert space
- *operators* (e.g. covariance) are treated as *functions* on these *points*

in order to generalize multivariate statistical methods, e.g.

- principal component analysis (PCA)
- linear discriminant analysis (LDA)

Key object: the **covariance** operator $C = \mathbb{E}\big[(X_1 - \mathbb{E}X_1) \otimes (X_1 - \mathbb{E}X_1)\big]$

- trace-class
  - unbounded inverse
  - summable eigenvalues
- non-stationary

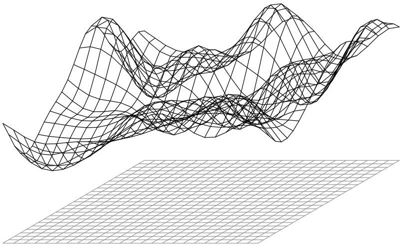
Setup

We have (continuous) surface valued data

$X_1, \ldots, X_N \in \mathcal{H}_1 \otimes \mathcal{H}_2$

I. $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{L}^2[0,1]$

II. $\mathcal{H}_1 = \mathbb{R}^K$ and $\mathcal{H}_2 = \mathbb{R}^K$



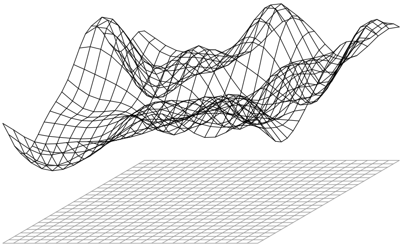
Setup

We have (continuous) surface valued data

$X_1, \ldots, X_N \in \mathcal{H}_1 \otimes \mathcal{H}_2$

I. $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{L}^2[0,1]$

II. $\mathcal{H}_1 = \mathbb{R}^K$ and $\mathcal{H}_2 = \mathbb{R}^K$



Assume $\mathbb{E}X = 0$

Covariance operator:

$$C = \mathbb{E}(X \otimes X) \in \mathcal{S}_1(\mathcal{H}_1 \otimes \mathcal{H}_2) \quad \longleftrightarrow \quad c(t,s,t',s') = \mathrm{Cov}\left(X(t,s), X(t',s')\right)$$

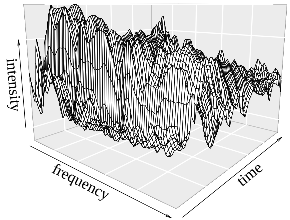
I. $c \in \mathcal{L}^2[0,1]^4$

II. $c \in \mathbb{R}^{K \times K \times K \times K}$

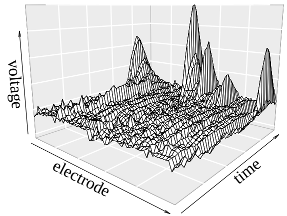
- PCA,
- classification,
- prediction, ...
- linguistics,
- bioimaging,
- finance, ...

Examples

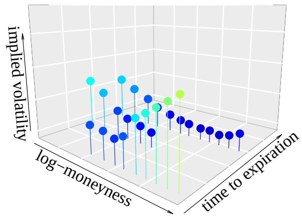
spectograms



EEG



implied volatility
(sparse measurements)



Separability

$c \in \mathbb{R}^{K \times K \times K \times K}$

$$\widehat{c}_N(t,s,t',s') = \frac{1}{N}\sum_{n=1}^{N} X(t,s)X(t',s')$$

Observed df: $\mathcal{O}(NK^2)$

$(X_1,\ldots,X_N \in \mathbb{R}^{K\times K})$

Complexity	Memory	Time		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$

Separability

$c \in \mathbb{R}^{K \times K \times K \times K}$ $\qquad$ $\widehat{c}_N(t,s,t',s') = \frac{1}{N}\sum_{n=1}^{N} X(t,s)X(t',s')$

Observed df: $\mathcal{O}(NK^2)$ $\qquad$ $(X_1, \ldots, X_N \in \mathbb{R}^{K \times K})$

Separability: $c(t,s,t',s') = c_1(t,t')c_2(s,s')$

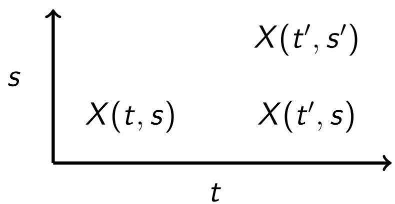
Complexity	Memory	Time		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$

$$c \in \mathbb{R}^{K\times K\times K\times K} \qquad \widehat{c}_N(t,s,t',s') = \frac{1}{N}\sum_{n=1}^{N} X(t,s)X(t',s')$$

Observed df: $\mathcal{O}(NK^2)$ $\qquad (X_1,\ldots,X_N \in \mathbb{R}^{K\times K})$

Separability: $c(t,s,t',s') = c_1(t,t')c_2(s,s')$

Complexity	Memory	Time		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$



$$\mathrm{Cov}\Big(X(t,s), X(t',s') \Big| X(t',s)\Big) = 0$$

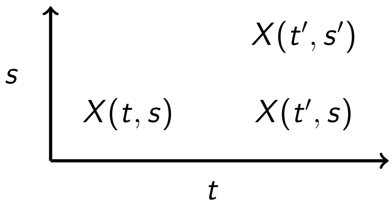
Separability

$$c \in \mathbb{R}^{K \times K \times K \times K} \qquad \widehat{c}_N(t,s,t',s') = \frac{1}{N}\sum_{n=1}^{N} X(t,s)X(t',s')$$

Observed df: $\mathcal{O}(NK^2)$ $\qquad (X_1, \ldots, X_N \in \mathbb{R}^{K \times K})$

Separability: $c(t,s,t',s') = c_1(t,t')c_2(s,s')$

Complexity	Memory	Time		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$



$$\mathrm{Cov}\Big(X(t,s), X(t',s') \Big| X(t',s)\Big) = 0$$

$$Y(t,s) = X(t,s) + \epsilon_{t,s} \qquad t,s = 1,\ldots,K$$

$$\mathrm{Cov}(Y) = \mathrm{Cov}(X) + Id$$ cannot be separable

Motivation

Motivation coming from a line of papers on linguistic data:

1. Pigoli, Aston, Dryden & Secchi (2014). Distances and inference for covariance operators. *Biometrika*, 101(2), 409-422.
   - a preliminary analysis of the linguistic data **using separability**
2. Aston, Pigoli & Tavakoli (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 1431-1461.
   - arguably the go-to test for separability
   - **rejects separability** for the linguistic data
3. Pigoli, Hadjipantelis, Coleman & Aston (2018). The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages. *JRSSC*, 67(5), 1103-1145.
   - full linguistic data analysis, still **using separability**

An example of the need for generalizations of separability.

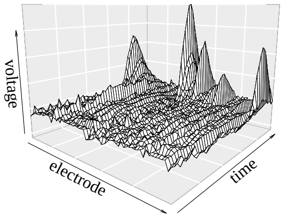
Outline

Part I: Separable-plus-banded Model

- generalization of separability
- short-range entanglement
- shifted partial tracing

Part II: Separable Component Decomposition

- canonical generalization of separability
- partial inner product
- classification of EEG signals



Part I: Masak & Panaretos (2022). Random surface covariance estimation by shifted partial tracing. *Journal of the American Statistical Association*.

Part II: Masak, Sarkar & Panaretos (2023). Separable expansions for covariance estimation via the partial inner product. *Biometrika*, 110(1), 225-247.

I + II $\Rightarrow$ **surfcov** package in **R**

A Generalization of Separability

$$c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s') \qquad (\star)$$

where $b(t,s,t',s') = 0$ for $\max(|t-t'|,|s-s'|) \geq \delta$

A Generalization of Separability

$$c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s') \qquad (\star)$$

where $b(t,s,t',s') = 0$ for $\max(|t-t'|,|s-s'|) \geq \delta$

Under continuity, the δ-shifted partial traces $\mathrm{Tr}_1^\delta(C)$ and $\mathrm{Tr}_2^\delta(C)$ are operators with kernels

$$k_{\mathrm{Tr}_1^\delta(C)}(t,t') = \int_0^{1-\delta} c(t,s,t',s+\delta)ds \quad \& \quad k_{\mathrm{Tr}_2^\delta(C)}(s,s') = \int_0^{1-\delta} c(t,s,t+\delta,s')dt$$

and the δ-shifted trace is defined as $\mathrm{Tr}^\delta(C) = \int_0^{1-\delta}\int_0^{1-\delta} c(t,s,t+\delta,s+\delta)dtds$

A Generalization of Separability

$$c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s') \qquad (\star)$$

where $b(t,s,t',s') = 0$ for $\max(|t-t'|,|s-s'|) \geq \delta$

Under continuity, the δ-shifted partial traces $\mathrm{Tr}_1^\delta(C)$ and $\mathrm{Tr}_2^\delta(C)$ are operators with kernels

$$k_{\mathrm{Tr}_1^\delta(C)}(t,t') = \int_0^{1-\delta} c(t,s,t',s+\delta)ds \quad \& \quad k_{\mathrm{Tr}_2^\delta(C)}(s,s') = \int_0^{1-\delta} c(t,s,t+\delta,s')dt$$

and the δ-shifted trace is defined as $\mathrm{Tr}^\delta(C) = \int_0^{1-\delta}\int_0^{1-\delta} c(t,s,t+\delta,s+\delta)dtds$

For model $(\star)$, we have:

$$k_{\mathrm{Tr}_1^\delta(C)}(t,t') = a_1(t,t')\int_0^{1-\delta} a_2(s,s+\delta)ds + \int_0^{1-\delta} b(t,s,t',s+\delta)ds \propto a_1(t,t')$$

Estimators:

$$a_1(t,t') = \frac{\mathrm{Tr}_1^\delta(C)}{\sqrt{\mathrm{Tr}^\delta(C)}} \quad \& \quad a_2(s,s') = \frac{\mathrm{Tr}_2^\delta(C)}{\sqrt{\mathrm{Tr}^\delta(C)}}$$

A Generalization of Separability

$$c(t,s,t',s') = a_1(t,t')a_2(s,s') + b(t,s,t',s') \qquad (\star)$$

where $b(t,s,t',s') = 0$ for $\max(|t-t'|,|s-s'|) \geq \delta$

Under continuity, the δ-shifted partial traces $\mathrm{Tr}_1^\delta(C)$ and $\mathrm{Tr}_2^\delta(C)$ are operators with kernels

$$k_{\mathrm{Tr}_1^\delta(C)}(t,t') = \int_0^{1-\delta} c(t,s,t',s+\delta)ds \quad \& \quad k_{\mathrm{Tr}_2^\delta(C)}(s,s') = \int_0^{1-\delta} c(t,s,t+\delta,s')dt$$

and the δ-shifted trace is defined as $\mathrm{Tr}^\delta(C) = \int_0^{1-\delta}\int_0^{1-\delta} c(t,s,t+\delta,s+\delta)dtds$

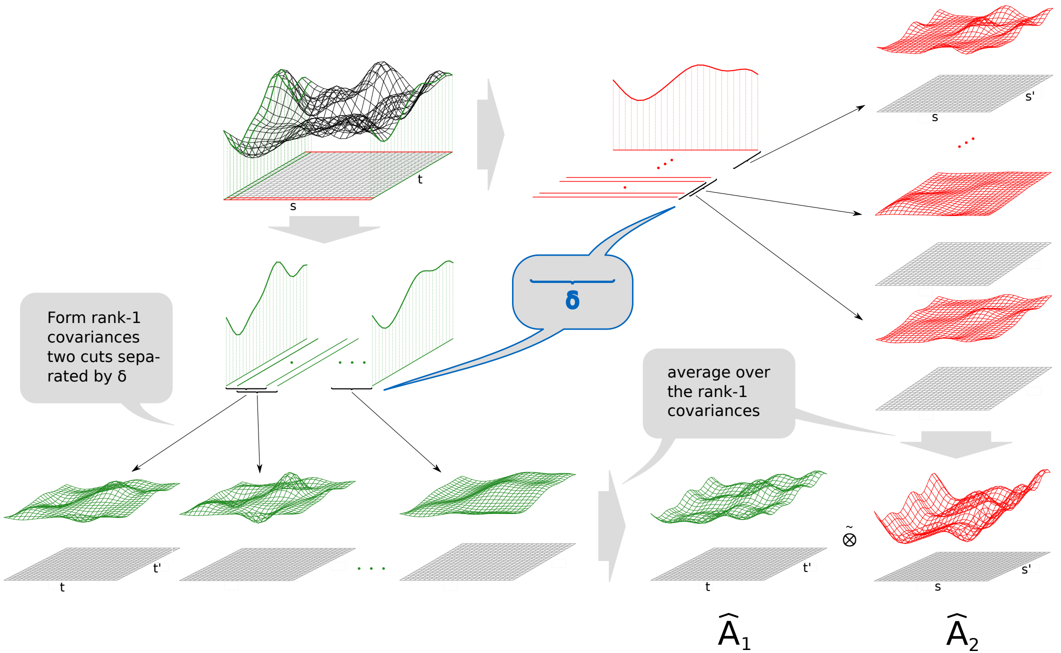
For model $(\star)$, we have:

$$k_{\mathrm{Tr}_1^\delta(C)}(t,t') = a_1(t,t')\int_0^{1-\delta} a_2(s,s+\delta)ds + \int_0^{1-\delta} b(t,s,t',s+\delta)ds \propto a_1(t,t')$$

Estimators:

$$\widehat{a}_1(t,t') = \frac{\mathrm{Tr}_1^\delta(\widehat{C}_N)}{\sqrt{\mathrm{Tr}^\delta(\widehat{C}_N)}} \quad \& \quad \widehat{a}_2(s,s') = \frac{\mathrm{Tr}_2^\delta(\widehat{C}_N)}{\sqrt{\mathrm{Tr}^\delta(\widehat{C}_N)}}$$

Shifted Partial Tracing Visualized



Asymptotics

Theorem (Discrete Observations)

Assume:

1. $Cov(X_n) = A_1 \tilde{\otimes} A_2 + B$, *where* B *is banded by* $\delta^\star$,
2. $\delta \geq \delta^\star$ *such that* $Tr^\delta(A_1 \tilde{\otimes} A_2) \neq 0$,
3. $\mathbb{E}\|X_n\|^4 < \infty$ *and* A_1, A_2 *be Lipschitz with constant* L,
4. $Y_n(t,s) = X_n(t,s) + E_n(t,s)$, $t,s = 1,\ldots,K$, $n = 1,\ldots,N$ *be a random sample, where* E_n *is white noise with variance* $\leq \sigma^2 = \mathcal{O}(\sqrt{K})$.

Then

$$\left|\left|\left| \widehat{A}_1^K \tilde{\otimes} \widehat{A}_1^K - A_1 \tilde{\otimes} A_2 \right|\right|\right|_2^2 = \mathcal{O}_P(N^{-1}) + 2K^{-2}L^2$$

where the $\mathcal{O}_P(N^{-1})$ *term is uniform in the grid size* K.

Asymptotics

Theorem (Discrete Observations)

Assume:

1. $Cov(X_n) = A_1 \tilde{\otimes} A_2 + B$, *where* B *is banded by* $\delta^\star$,
2. $\delta \geq \delta^\star$ *such that* $Tr^\delta(A_1 \tilde{\otimes} A_2) \neq 0$,
3. $\mathbb{E}\|X_n\|^4 < \infty$ *and* A_1, A_2 *be Lipschitz with constant* L,
4. $Y_n(t,s) = X_n(t,s) + E_n(t,s)$, $t, s = 1, \ldots, K$, $n = 1, \ldots, N$ *be a random sample, where* E_n *is white noise with variance* $\leq \sigma^2 = \mathcal{O}(\sqrt{K})$.

Then

$$\left|\left|\left| \widehat{A}_1^K \tilde{\otimes} \widehat{A}_1^K - A_1 \tilde{\otimes} A_2 \right|\right|\right|_2^2 = \mathcal{O}_P(N^{-1}) + 2K^{-2}L^2$$

where the $\mathcal{O}_P(N^{-1})$ *term is uniform in the grid size* K.

- provably consistent data-driven choice of δ
- rates of convergence also for
  - the data-driven choice of δ
  - uniform (instead of L^2) rates
  - rates for eigenfunctions
- goodness of fit test for the model

Computational Considerations

Complexity	Memory	Time		
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$
separable+banded	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$

Banded part has $\mathcal{O}(K^2)$ parameters if it is

- diagonal $\Rightarrow$ separable+diagonal model (separability under heteroscedastic white noise), or
- stationary $\Rightarrow$ separable+stationary model (separability under weakly dependent white noise)

Inversion: alternating direction implicit (ADI) method

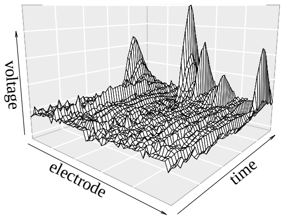
Outline

Part I: Separable-plus-banded Model

- generalization of separability
- shifted partial tracing
- PCA of mortality surfaces

Part II: Separable Component Decomposition

- canonical generalization of separability
- partial inner product
- classification of EEG signals



Covariance Operator

The following spaces are isometrically isomorphic:

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \simeq \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2) \simeq \mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1) \simeq \mathcal{S}_2(\mathcal{H}_1) \otimes \mathcal{S}_2(\mathcal{H}_2)$$

Eigen: $C = \sum_{j=1}^{\infty} \lambda_j g_j \otimes g_j, \quad g_j \in \mathcal{H}_1 \otimes \mathcal{H}_2$

SVD: $C = \sum_{j=1}^{\infty} \sigma_j e_j \otimes_2 f_j, \quad e_j \in \mathcal{H}_1 \otimes \mathcal{H}_1, f_j \in \mathcal{H}_2 \otimes \mathcal{H}_2$

SCD: $C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j, \quad A_j \in \mathcal{S}_2(\mathcal{H}_1), B_j \in \mathcal{S}_2(\mathcal{H}_2)$

$(e_j \otimes_2 f_j)x := \langle x, e \rangle f$

Covariance Operator

The following spaces are isometrically isomorphic:

$$\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2 \simeq \mathcal{S}_2(\mathcal{H}_1 \otimes \mathcal{H}_2) \simeq \mathcal{S}_2(\mathcal{H}_2 \otimes \mathcal{H}_2, \mathcal{H}_1 \otimes \mathcal{H}_1) \simeq \mathcal{S}_2(\mathcal{H}_1) \otimes \mathcal{S}_2(\mathcal{H}_2)$$

$$\text{Eigen:} \quad C = \sum_{j=1}^{\infty} \lambda_j g_j \otimes g_j, \quad g_j \in \mathcal{H}_1 \otimes \mathcal{H}_2$$

$$\text{SVD:} \quad C = \sum_{j=1}^{\infty} \sigma_j e_j \otimes_2 f_j, \quad e_j \in \mathcal{H}_1 \otimes \mathcal{H}_1, f_j \in \mathcal{H}_2 \otimes \mathcal{H}_2$$

$$\text{SCD:} \quad C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j, \quad A_j \in \mathcal{S}_2(\mathcal{H}_1), B_j \in \mathcal{S}_2(\mathcal{H}_2)$$

Definition: Covariance $C \in \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2$ is called ***R*-separable** if

$$C = \sum_{j=1}^{R} \sigma_j A_j \otimes B_j, \quad A_j \in \mathcal{H}_1 \otimes \mathcal{H}_1, B_j \in \mathcal{H}_2 \otimes \mathcal{H}_2$$

Is R-separability a useful notion?

$$\text{SCD:} \quad C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j, \quad A_j \in \mathcal{S}_2(\mathcal{H}_1), B_j \in \mathcal{S}_2(\mathcal{H}_2)$$

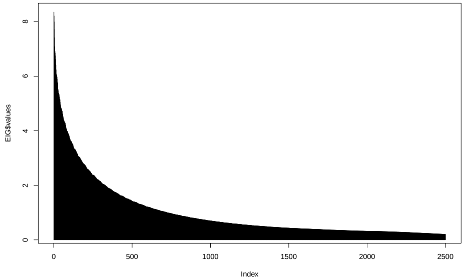
Any covariance $C \in \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2$ can be approximated to arbitrary precision as $\sum_{j=1}^{R} \sigma_j A_j \otimes B_j$ for R large enough.

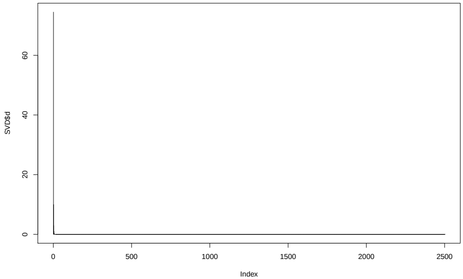
Is R-separability a useful notion?

$$\text{SCD:} \quad C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j, \quad A_j \in \mathcal{S}_2(\mathcal{H}_1), B_j \in \mathcal{S}_2(\mathcal{H}_2)$$

Any covariance $C \in \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_1 \otimes \mathcal{H}_2$ can be approximated to arbitrary precision as $\sum_{j=1}^{R} \sigma_j A_j \otimes B_j$ for R large enough.

$\mathbf{C} \in \mathbb{R}^{50\times 50\times 50\times 50}$ fitted (via a parametric model) to the Irish Wind data is considered the ground truth here.





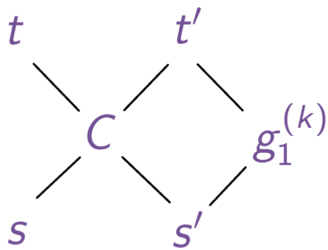
Left: 2031 (out of 2500) eigenvectors needed to explain 95% of variance
Right: only **two** singular vector pairs needed to explain 95% of variance

Generalized Power Iteration

Eigendecomposition:

$$C = \sum_{j=1}^{\infty} \lambda_j g_j \otimes g_j$$

$$g_j \in \mathcal{H}_1 \otimes \mathcal{H}_2$$

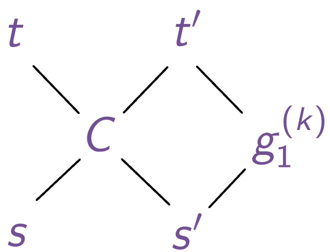


Generalized Power Iteration

Eigendecomposition:

$$C = \sum_{j=1}^{\infty} \lambda_j g_j \otimes g_j$$

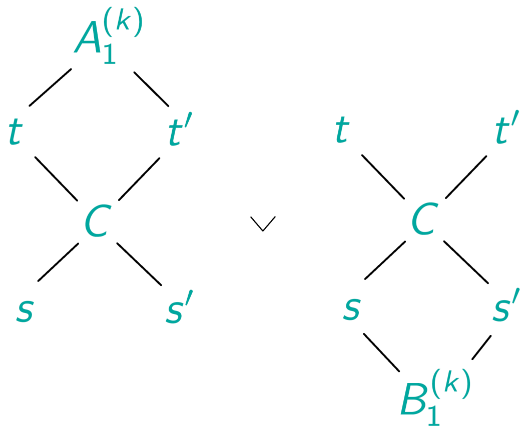
$$g_j \in \mathcal{H}_1 \otimes \mathcal{H}_2$$



SCD:

$$C = \sum_{j=1}^{\infty} \sigma_j A_j \otimes B_j$$

$$A_j \in \mathcal{S}_2(\mathcal{H}_1), B_j \in \mathcal{S}_2(\mathcal{H}_2)$$



Partial Inner Product

Tensor product operators:

$$(x \otimes_1 y)f = \langle y, f \rangle x$$
$$(x \otimes_2 y)e = \langle x, e \rangle y \qquad x, e \in H_1, y, f \in H_2$$

$$\otimes_1 : [H_1 \times H_2] \times H_2 \to H_1,$$
$$\otimes_2 : [H_1 \times H_2] \times H_1 \to H_2.$$

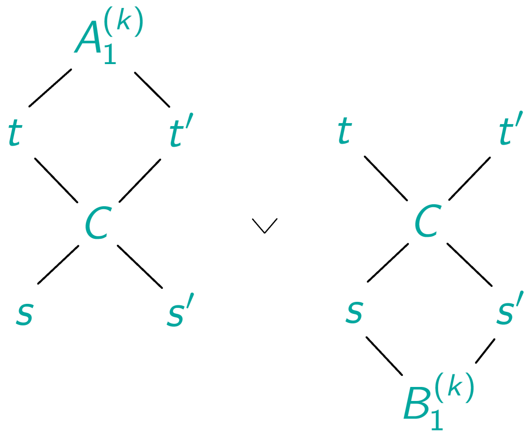
Definition: The partial inner product w.r.t. the first argument is the unique bi-linear operator $T_1 : [H_1 \otimes H_2] \times H_2 \to H_1$ defined by

$$T_1(x \otimes y, f) = (x \otimes_1 y)f, \quad x \in H_1,\ y, f \in H_2.$$

Similarly, the partial inner product w.r.t. the second argument is the unique bi-linear operator $T_2 : [H_1 \otimes H_2] \times H_1 \to H_2$ defined by

$$T_2(x \otimes y, e) = (x \otimes_2 y)e, \quad x, e \in H_1,\ y \in H_2.$$

Generalized Power Iteration



$T_2(C, A_1^{(k)})$ $\qquad$ $T_1(C, B_1^{(k)})$

Proposition: This alternating scheme is provably convergent with a linear rate.

R-separable estimator

$$\widehat{\mathbf{C}}_{N,R} = \arg\min_{\mathbf{C}} \left|\left|\left| \widehat{\mathbf{C}}_N - \mathbf{C} \right|\right|\right|_2^2 \quad \text{s.t.} \quad \mathbf{C} = \sum_{j=1}^{R} \sigma_j \mathbf{A}_j \otimes \mathbf{B}_j.$$

This leads to estimator

$$\widehat{\mathbf{C}}_{N,R} = \sum_{r=1}^{R} \widehat{\sigma}_r \widehat{\mathbf{A}}_r \otimes \widehat{\mathbf{B}}_r,$$

which is computed via the **generalized power iteration**. For the first term

repeat

$\mathbf{A} := T_1(\widehat{\mathbf{C}}_N, \mathbf{B})$

$\mathbf{B} := T_2(\widehat{\mathbf{C}}_N, \mathbf{A})$

until convergence

R-separable estimator

$$\widehat{\mathbf{C}}_{N,R} = \arg\min_{\mathbf{C}} \left|\left|\left| \widehat{\mathbf{C}}_N - \mathbf{C} \right|\right|\right|_2^2 \quad \text{s.t.} \quad \mathbf{C} = \sum_{j=1}^{R} \sigma_j \mathbf{A}_j \otimes \mathbf{B}_j .$$

This leads to estimator

$$\widehat{\mathbf{C}}_{N,R} = \sum_{r=1}^{R} \widehat{\sigma}_r \widehat{\mathbf{A}}_r \otimes \widehat{\mathbf{B}}_r ,$$

which is computed via the **generalized power iteration**. For the first term

repeat

$\mathbf{A} := T_1(\widehat{\mathbf{C}}_N, \mathbf{B})$

$\mathbf{B} := T_2(\widehat{\mathbf{C}}_N, \mathbf{A})$

until convergence

Partial inner products can be calculated on the level of data:

$$T_1(\widehat{\mathbf{C}}_N, \mathbf{B}) = \frac{1}{N} \sum_{n=1}^{N} T_1(\mathbf{X}_n \otimes \mathbf{X}_n, \mathbf{B}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{X}_n \mathbf{B} \mathbf{X}_n^\top ,$$

$$T_2(\widehat{\mathbf{C}}_N, \mathbf{A}) = \frac{1}{N} \sum_{n=1}^{N} T_2(\mathbf{X}_n \otimes \mathbf{X}_n, \mathbf{A}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{X}_n^\top \mathbf{A} \mathbf{X}_n .$$

Properties

- estimation computationally efficient
- degree-of-separability R selection via CV
  - bias vs. variance
- numerical inversion computationally efficient

$\Rightarrow$ estimation beyond separability without excessive computational costs, for a small fixed R:

Complexity	Memory		Time	
		Estimation	Application	Inversion
empirical	$\mathcal{O}(K^4)$	$\mathcal{O}(NK^4)$	$\mathcal{O}(K^4)$	$\mathcal{O}(K^6)$
separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$
separable+banded	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$
R-separable	$\mathcal{O}(K^2)$	$\mathcal{O}(NK^3)$	$\mathcal{O}(K^3)$	$\mathcal{O}(K^3)$

Asymptotic Properties

Theorem (Complete Observations)

For $X_1, \ldots, X_N \sim X$ i.i.d. on $H = H_1 \otimes H_2$ with $\mathbb{E}\|C\|^4 < \infty$, denote

- $C = \sum_{i \geq 1} \sigma_i A_i \otimes B_i$ *the SCD of the covariance of* X,
- $\alpha_i = \min\{\sigma_{i-1}^2 - \sigma_i^2, \sigma_i^2 - \sigma_{i+1}^2\}$ *the gap between the squared scores,*
- $a_R = |||C|||_2 \sum_{i \geq 1} (\sigma_i / \alpha_i)$.

Then

$$\left|\left|\left| \widehat{C}_{R,N} - C \right|\right|\right|_2 \leq \sqrt{\sum_{i=R+1}^{\infty} \sigma_i^2} + \mathcal{O}_P\left(\frac{a_R}{\sqrt{N}}\right).$$

Theorem (Discrete Observations)

Under setup the above, but with measurements on a $K_1 \times K_2$ grid, and L-Lipschitz continuity of C, we have

$$\left|\left|\left| \widehat{C}_{R,N} - C \right|\right|\right|_2 \leq \sqrt{\sum_{i=R+1}^{\infty} \sigma_i^2} + \mathcal{O}_P\left(\frac{a_R}{\sqrt{N}}\right)$$

$$+ (16 a_R + \sqrt{2}) L \sqrt{\frac{1}{K_1^2} + \frac{1}{K_2^2}} + \frac{8\sqrt{2} L^2}{|||C|||_2} \left(\frac{1}{K_1^2} + \frac{1}{K_2^2}\right) a_R.$$

Classification of EEG Signals

$X_1, \ldots, X_{121} \in \mathbb{R}^{256 \times 64}$ $N = 121$ patients (76 alcoholic + 45 control)
$K_1 = 256$ time points (= 1 second)
$K_2 = 64$ electrodes

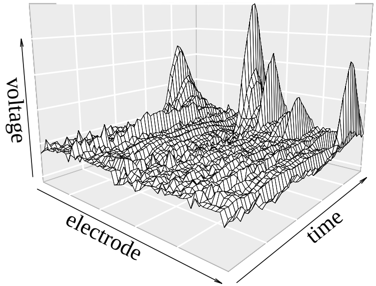
3 different trial conditions $\Rightarrow$ 3 separate data sets

fLDA classifier

$$\mathbb{1}_{[\langle X - m_0, v\rangle^2 > \langle X - m_1, v\rangle^2]}$$

with v given by

$$C\, v = m_1 - m_0$$



Classification of EEG Signals

$X_1, \dots, X_{121} \in \mathbb{R}^{256 \times 64}$ $\quad N = 121$ patients (77 alcoholic + 45 control)
$K_1 = 256$ time points ($= 1$ second)
$K_2 = 64$ electrodes

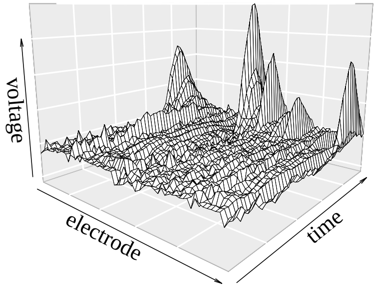
3 different trial conditions $\Rightarrow$ 3 separate data sets

fLDA classifier

$$\mathbb{1}_{[\langle X - \widehat{m}_0, v\rangle^2 > \langle X - \widehat{m}_1, v\rangle^2]}$$

with v given by

$$\widehat{C}_{R,N}\, v = \widehat{m}_1 - \widehat{m}_0$$



Degree-of-separability	Condition 1	Condition 2	Condition 3
$R = 1$	78 %	79 %	77 %
$R = 2$	**90 %**	**84 %**	**91 %**

Table: Accuracy of classification.

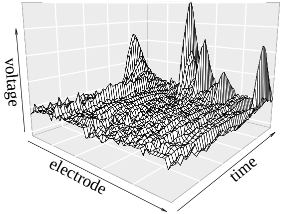
Outline

Part I: Separable-plus-banded Model

- generalization of separability
- shifted partial tracing
- PCA of mortality surfaces

Part II: Separable Component Decomposition

- canonical generalization of separability
- partial inner product
- classification of EEG signals



Extra: Separability under Sparse Measurement

- "sparsified" partial inner product
- prediction of implied volatility

Outline

Part I: Separable-plus-banded Model

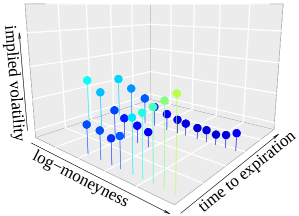
- generalization of separability
- shifted partial tracing
- PCA of mortality surfaces

Part II: Separable Component Decomposition

- canonical generalization of separability
- partial inner product
- classification of EEG signals

Extra: Separability under Sparse Measurement

- "sparsified" partial inner product
- prediction of implied volatility



Prediction

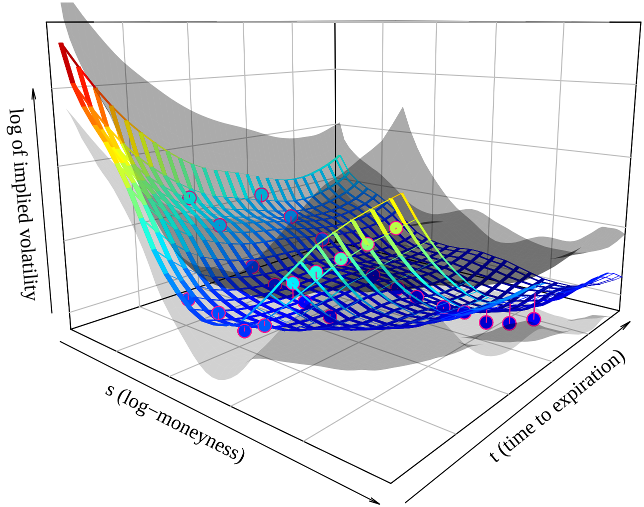


Figure: Predicted implied volatility surface for Dell Technologies with the 95% simultaneous confidence band under Gaussianity.

Future Directions

1. models arising as superposition of simple structures
   - $C = A_1 \otimes A_2 + B$
   - weak separability and weak bandedness

Future Directions

1. models arising as superposition of simple structures
   - $C = A_1 \otimes A_2 + B + C_1 \otimes U + V \otimes C_2$
   - weak separability and weak bandedness

Future Directions

1. models arising as superposition of simple structures
   - $C = A_1 \otimes A_2 + B + C_1 \otimes U + V \otimes C_2$
   - weak separability and weak bandedness
2. sparsely observed surface valued time series
3. domains of higher dimensions
   - EEG data are 2D, but MRI is 3D, fMRI 4D
   - evolution (e.g. of brain connectivity) in time – an extra longitudinal dimension
4. spherical data
5. **applications**
   - classification of words using spectograms
   - classification of languages using covariances as data

References

- Aston, Pigoli & Tavakoli (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 1431-1461.
- Bagchi & Dette (2020). A Test for Separability in Covariance Operators of Random Surfaces. *The Annals of Statistics*, 48(4), 2303-2322.
- **Masak** & Panaretos (2022). Random surface covariance estimation by shifted partial tracing. *Journal of the American Statistical Association*.
- **Masak**, Rubin & Panaretos (2022). Inference and Computation for Sparsely Sampled Random Surfaces. *Journal of Computational and Graphical Statistics*, 31(4), 1361-1374.
- **Masak**, Sarkar & Panaretos (2023). Separable expansions for covariance estimation via the partial inner product. *Biometrika*, 110(1), 225-247.
- Van Loan & Pitsianis (1993) Approximation with Kronecker product. In *Linear algebra for large scale and real-time applications*, pages 293-314. Springer.