universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Determinantal point processes and their application in dimension reduction“

verfasst von / submitted by

lic. Piotr Pawel Kucharski

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik UG2002
Betreut von / Supervisor:	Assoz. Prof. José Luis Romero, Privatdoz. PhD

Abstract

Determinantal point processes are objects from probability theory that have recently risen in popularity in machine learning, quantum physics and various other fields. Their origins reach back to the article [12] from 1975, in which O. Macchi denoted them as fermion processes (a fermion is a type of particle). The model, coming from the theory of quantum particles, shows how a group of fermions is distributed in a state of thermal equilibrium. The fermion process is a time process, which involves recording moments of time at which a detector put at a given point registers any particle. In this model it is important, that the points "tend to exclude one another" - there is a phenomenon of repulsion. Indeed, the defining trait of determinantal point processes is their negative correlation property - chosen elements are more likely to be diverse, which makes these point processes good for modeling diversity. This property is what has made them famous and what has led to wide dissemination in many different applications, such as image and text summarization, matrix factorization, studying the geometry of moduli spaces and dimension reduction. In the last example, thanks to repulsion, a determinantal process can capture directionality in the dependence structure of data. A strong emphasis is put on a recently introduced subclass, called Gaussian determinantal processes, as we can expect large datasets to have Gaussian properties.

In this thesis, we provide an overview of the theory and application of determinantal point processes in dimension reduction with comparison to other methods. In particular chapter 3 provides original insights, selecting the right samples to show different, both preferential and not, outcomes of the so called Gaussian determinantal point process method as well as generalising the estimation of empirical estimator outside of Gaussian scope.

Abstrakt

Determinantaler Punktprozess ist Objekt aus der Wahrscheinlichkeitstheorie, die in letzter Zeit im maschinellen Lernen, in der Quantenphysik und in verschiedenen anderen Bereichen immer beliebter werden. Ihre Ursprünge gehen auf den Artikel [12] aus dem Jahr 1975 zurück, in dem O. Macchi sie als Fermionprozesse (ein Fermion ist eine Teilchenart) bezeichnet. Das aus der Theorie der Quantenteilchen stammende Modell zeigt, wie eine Gruppe von Fermionen im thermischen Gleichgewicht verteilt ist. Der Fermionprozess ist ein Zeitprozess, bei dem Momente aufgezeichnet werden, in denen ein an einem bestimmten Punkt angebrachter Detektor ein Teilchen registriert. In diesem Modell ist es wichtig, dass die Punkte "tend to exclude one another" – es liegt ein Phänomen der Abstoßung vor. Tatsächlich ist das bestimmende Merkmal determinanter Punktprozesse ihre negative Korrelationseigenschaft – ausgewählte Elemente sind mit größerer Wahrscheinlichkeit vielfältig, was diese Punktprozesse gut für die Modellierung von Diversität macht. Diese Eigenschaft hat sie berühmt gemacht und zu einer weiten Verbreitung in vielen verschiedenen Anwendungen geführt, beispielsweise zur Zusammenfassung von Bildern und Texten, zur Matrixfaktorisierung, zur Untersuchung der Geometrie von Modulräumen und zur Dimensionsreduktion. Im letzten Beispiel kann ein determinanter Prozess dank der Abstoßung die Richtung in der Abhängigkeitsstruktur von Daten erfassen. Ein starker Schwerpunkt liegt auf einer kürzlich eingeführten Unterklasse, den sogenannten Gaußschen Determinantenprozessen, da wir davon ausgehen können, dass große Datensätze Gaußsche Eigenschaften aufweisen.

In dieser Arbeit geben wir einen Überblick über die Theorie und Anwendung determinanter Punktprozesse bei der Dimensionsreduktion im Vergleich zu anderen Methoden. Insbesondere Kapitel 3 liefert originelle Erkenntnisse, wählt die richtigen Stichproben aus, um unterschiedliche, sowohl bevorzugte als auch nicht bevorzugte Ergebnisse der sogenannten Gaußschen Determinantenpunktprozessmethode zu zeigen, und verallgemeinert die Schätzung empirischer Schätzer außerhalb des Gaußschen Bereichs.

Acknowledgment

This thesis was co-supervised by Dr. Günther Koliander. The author is very grateful to both advisors, Koliander and Romero, for their guidance.

Contents

1 Background

This thesis deals with the topic of determinantal point processes - a class of probability distributions over subsets of a given space, that appear in various math-related fields, such as machine learning and quantum physics. In order to understand further ideas, we begin by introducing the mathematical framework behind them, including definitions, properties and ways to generate a sample, following the book "Zeros of Gaussian Analytic Functions and Determinantal Point Processes" [6]. The practical application of this model - tackling the curse of dimensionality, will be deferred to a succeeding chapter.

1.1 Determinantal point process

As we are dealing with a concept from measure theory, we need to define the space and measure we are working with.

Definition 1.1.1 *A Polish space is a topological space that is separable and completely metrizable.*

Definition 1.1.2 *Let X be a Hausdorff topological space and μ be a measure on the σ-algebra of Borel sets of X. Then μ is a Radon measure if:*
$\triangleright \mu(K) < \infty$ *for every compact subset* K,
$\triangleright \mu(U) = \sup\{\mu(K); K \subseteq U, K \text{ compact}\}$ *for every open subset* U *and*
$\triangleright \mu(B) = \sup\{\mu(U); B \subseteq U, U \text{ open}\}$ *for every Borel subset* B.

To define a determinantal point process, we need a locally compact Polish space $\mathcal{D}$. In particular $\mathbb{R}^n$ fulfills these conditions. The measure we use is a random integer-valued positive Radon measure defined on $\mathcal{D}$ ($\mathcal{D}$ is a Hausdorff space, since it is a locally compact Polish space).

To properly define determinantal point processes we need to define point processes.

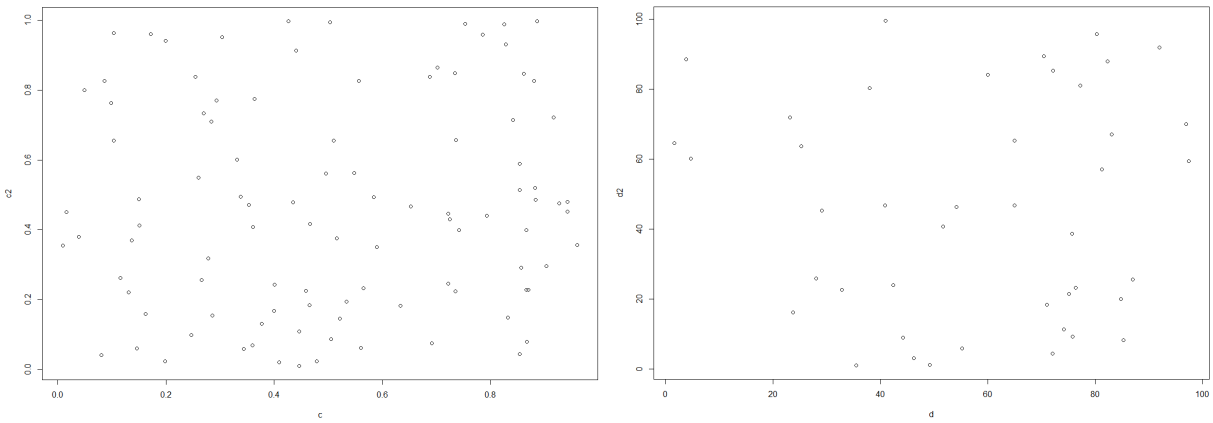
Definition 1.1.3 (Point processes) *A point process $\mathfrak{X}$ on a locally compact Polish space $\mathcal{D}$ is a random integer-valued positive Radon measure on $\mathcal{D}$. If this measure is non greater than* 1 *on singletons we call it a simple point process.*

A classic way to construct a point process is to take a random integer-valued positive Radon measure $\mathfrak{X}$ defined on $\mathcal{D}$ with following properties:

$\triangleright \mathfrak{X}(D_1), ..., \mathfrak{X}(D_k)$ are independent for disjoint sets $D_1, ..., D_k$ of $\mathcal{D}$ and
$\triangleright \mathfrak{X}(D)$ has some discrete distribution with parameter dependent on D for every bounded

Borel set D of $\mathcal{D}$.
A prime example is the Poisson point process, where $\mathcal{D} = \mathbb{R}^d$ and $\mathfrak{X}(D) \sim Poi\,(\lambda|D|)$ for some parameter $\lambda > 0$ and Lebesgue measure $|D|$. Below we have shown some samples of point processes.

Figure 1.1.4



Graphic representation of sample points of a binomial point process (a process where a set number of points is distributed uniformly in the space $\mathcal{D}$) with $n = 100$ on $[0,1]^2$ and Poisson point process with $\lambda = 0.5$ on $[0,100]^2$

A point process can also be defined by the probabilities $\mathbb{P}\left(\mathfrak{X}(D_k) = n_k, k \in [1,m]\right)$, where D_k are subsets of $\mathcal{D}$, however, a more common approach is to define them using functions called joint intensities.

Definition 1.1.5 (Joint intensities) *The joint intensities of simple point process $\mathfrak{X}$, defined on $\mathcal{D}$, with respect to a measure μ are functions $g_k : \mathcal{D}^k \to [0,\infty)$, such that $g_k(x_1,...,x_k) = 0$ if $x_i = x_j$ for some $i \neq j$ and for any mutually disjoint subsets $D_1,...,D_k$ of $\mathcal{D}$:*

$$\mathbb{E}\left[\prod_{i=1}^{k} \mathfrak{X}(D_i)\right] = \int_{\prod_i D_i} g_k(x_1,...,x_k)d\mu(x_1)...d\mu(x_k).$$

In case $k = 1$ we can define $\mu_1(D) = \mathbb{E}\left[\mathfrak{X}(D)\right]$ (i.e. $d\mu_1 = g_1 d\mu$) as the first intensity measure.

The first intensity measure can be understood as averaging a corresponding random measure. Additionally, from the Radon-Nikodym theorem, if $\mu_1 \ll \mu$, then one can prove the existence of g_1 from the definition. The Radon-Nikodym theorem can be applied to introduce the joint intensities in a different way - using moments of linear statistics

Definition 1.1.6 *Given a measurable function $\phi : \mathcal{D} \to \mathbb{R}$, the linear statistic is defined as:*

$$\mathfrak{X}(\phi) = \int_{\mathcal{D}} \phi d\mathfrak{X} = \sum_{k \in \mathcal{D}} \phi(k)\mathfrak{X}(\{k\}).$$

Using the expected value of the linear statistic, we define a positive linear functional $T_1(\phi) = \mathbb{E}(\mathfrak{X}(\phi))$ on the space of continuous functions with compact support on $\mathcal{D}$, denoted by $C_c(\mathcal{D})$. With the aid of the Riesz's representation theorem, we have the following:

$$\mathbb{E}[\mathfrak{X}(\phi)] = \int_{\mathcal{D}} \phi d\nu = \int_{\mathcal{D}} \phi g_1 d\mu,$$

for measure ν. The second equality is a consequence of the Radon-Nikodym theorem - if $\nu \ll \mu$, for the measure μ from the definition 1.1.5, we can rewrite it this way.

If we attempt to extend this method to the n-th intensity measure using the following analogous formula (T is the positive multilinear functional on $C_c(\mathcal{D})^n$):

$$T(\phi_1, ..., \phi_n) = \mathbb{E}\left[\prod_{i=1}^{n} \mathfrak{X}(\phi_i)\right] = \int_{\mathcal{D}^n} \prod_{i=1}^{n} \phi_i d\nu_n,$$

we would run into an obstacle - set $Diag = \{(x_1, ..., x_n); (\exists i, j)\ x_i = x_j\}$. The random measure ν_n will usually have atoms on this set, resulting in the lack of absolute continuity, since $g_n(x) = 0$ for every $x \in Diag$. We define measures $\mu_n, \widehat{\nu}_n$ in the following way:
For every subset $D \subseteq Diag$, $\widehat{\nu}_n(D) = \nu_n(D)$ and $\mu_n(D) = 0$,
For every subset $D \subseteq Diag^c$, $\mu_n(D) = \nu_n(D)$ and $\widehat{\nu}_n(D) = 0$.
From the construction, $\nu_n = \mu_n + \widehat{\nu}_n$ and the measure $\widehat{\nu}_n$ is singular to $\mu^{\otimes^n \mu}$, as for every $D \subseteq Diag$, $\mu^{\otimes^n \mu}(D) = 0$. The measure μ_n is called the n-point intensity measure of the point process $\mathfrak{X}$. If μ_n is absolutely continuous to $\mu^{\otimes^n \mu}$, we can utilize the Radon-Nikodym theorem again, resulting in the following equality:

$$\mathbb{E}\left[\prod_{i=1}^{n} \mathfrak{X}(\phi_i)\right] = \int_{\mathcal{D}^n} \phi g_n d\mu^{\otimes^k \mu},$$

which gives us the n-point intensity function g_n.

We mentioned earlier that a point process can be defined by joint intensities. Natural follow-up question would be which functions g_k can create it, i.e. what properties should they have to create a point process. Turns out, that the necessary and sufficient conditions for the existence are as follows [15, Theorem 1]:

Theorem 1.1.7 *Locally integrable functions $g_k : \mathcal{D}^k \to [0, \infty)$ for positive integers k are joint intensities of some point process iff they are:*
▷ *Symmetric - $g_k(x_1, ..., x_k) = g_k(x_{\sigma(1)}, ..., x_{\sigma(k)})$ for any permutation σ and*
▷ *Positive - for any k from 0 to some integer n and any measurable functions $f_k : \mathcal{D}^k \to$*

$[0,\infty)$ *with compact support the following inequality:*

$$f_0 + \sum_{k=1}^{n} \sum_{i_1 \neq ... \neq i_k} f_k(x_{i_1}, ..., x_{i_k}) \geq 0$$

implies:

$$f_0 + \sum_{k=1}^{n} \int_{\mathcal{D}^k} f_k(x_1, ..., x_k) g_k(x_1, ..., x_k) dx_1 ... dx_k \geq 0.$$

Since proving the sufficiency of these conditions requires an extensive proof (which can be found in the article [11]), it will be ommited in this thesis. It is important to notice that a point process may not have joint intensities, so it is not possible to construct every point process in this way.

Now, we can state the core definition of this chapter.

Definition 1.1.8 *A determinantal process with kernel* $\mathbb{K}$ *(further referred to as DPP) is a simple point process defined on* $\mathcal{D}$*, with joint intensities with respect to a measure* μ *satisfying the following:*

$$g_k(x_1, ..., x_k) = \det\left(\mathbb{K}(x_i, x_j)_{i,j \in [1,k]}\right),$$

for every integer k *and* $x_1, ..., x_k \in \mathcal{D}$*.*

In particular, in this thesis we will utilize the subclass of determinantal point processes, called Gaussian determinantal processes.

Definition 1.1.9 *A determinantal process* $\mathfrak{X}$ *with a Gaussian kernel* $\mathbb{K}$*, i.e.:*

$$\mathbb{K}(x, y) = \Phi(x - y),$$

for $\Phi(x)$ *being the density of a Gaussian vector with mean* 0 *and positive definite covariance matrix, is known as a Gaussian determinantal process (further referred to as GDP). In this case we call the covariance matrix a scattering matrix of* $\mathfrak{X}$*.*

To get better acquaintanced with this topic, we will take a common example from probability theory.

Example 1.1.10 *We define a sequence of i.i.d. one dimensional Markov chains* X_k *for* k *from* 1 *to* n*, so that:*
$\triangleright$ X_k *starts on* $i_k < i_{k+1}$ *for every integer* $k \in [1, n-1]$ *(*X_n *starts at* i_n*),*
$\triangleright$ *The probability of going from* i *to* j *in time* t *is* $P_{i,j}(t)$*,*
$\triangleright$ $P_{i,j}(1) = 0$ *for every* $j \neq i \pm 1$*.*
In this case, for $j_1 < ... < j_n$*, one can show, that the expression for the probability that no chains are intersecting from* 0 *to* t *and that they are at* $j_1, ..., j_n$ *at time* t *is:*

$$\det\left((P_{i_k, j_l}(t))_{k,l \in [1,n]}\right).$$

Using these informations we can find a determinantal point process in the following way. We take the aforedescribed sequence of i.i.d random walks X_k for k from 1 to n, so that its invariant measure π is reversible:

$$\pi(x)P_{x,y}(t) = \pi(y)P_{y,x}(t).$$

We assume that for every integer $p \in [1,n]$ and some sequence $(x_p)_{p\in[1,n]}$ we have the following: $X_p(0) = X_p(2t) = x_p$ and that no random walks intersect one another (we need even number of steps, so we can return to point x_p with positive probability). Then the configuration $\{X_p(t); p \in [1,n]\}$ is a DPP on $\mathbb{Z}$. In particular its kernel is of the form:

$$\mathbb{K}(x_i, x_j) = \sum_{k=1}^{n} \psi_k(x_i)\bar{\psi}_k(x_j),$$

where

$$\psi_k(x) = \sum_{j=1}^{n} \left(A^{-0.5}\right)_{k,j} \frac{P_{x_j,x}(t)}{\pi(x)},$$

$A_{j,k} = \dfrac{P_{x_j,x_k}(2t)}{\pi(x_k)}$ *and $P_{i,j}(t)$ is probability of taking a path from i to j in time t.*

Another example will prove useful later.

Example 1.1.11 *Let $(\phi_k)_{k=1}^n$ be any orthonormal subset of $L^2(\mathcal{D})$. Then the following expression:*

$$\mathbb{K}(x, y) = \sum_{k=1}^{n} \phi_k(x)\bar{\phi}_k(y)$$

generates a DPP.

These examples are intended to convey an intuitive understanding, so the proofs, which can be found in the aforementioned book [6, Theorem 4.3.1., Corollary 4.3.3., Lemma 4.5.1.], are ommited. Now we will state some properties of determinantal point processes.

Lemma 1.1.12 *The correlation kernel matrix of a DPP is positive semidefinite if it is a Hermitian (i.e. self-adjoint or symmetric if matrix is real) matrix.*

Proof: Since joint intensities g_k of a point process are non-negative for any k and any points from $\mathcal{D}$, all minor determinants need to be non-negative as well. Using Sylvester's criterion finishes the proof. QED

Lemma 1.1.12 was mentioned in the article [15]. It is important to note, that the kernel matrix can be non-Hermitian and still generate a DPP. The book [6] mentions article [8] as a source of interesting examples.

A point process will generally have more than one such kernel: if $\mathbb{K}(x_i, x_j)$ fulfills the equation, one can usually find another satisfying kernel.

Lemma 1.1.13 *Let $\mathfrak{X}$ be a DPP with kernel $\mathbb{K}$ and let $f : \mathcal{D} \to \mathbb{R}/\{0\}$ be any positive function. Then the following expression:*

$$\mathbb{S}(x_i, x_j) = f(x_i)\mathbb{K}(x_i, x_j)f(x_j)^{-1},$$

is a kernel that generates DPP $\mathfrak{X}$, i.e.:

$$\det\left(\mathbb{K}(x_i, x_j)_{i,j\in[1,k]}\right) = \det\left(\mathbb{S}(x_i, x_j)_{i,j\in[1,k]}\right),$$

for any integer k. The kernel $\mathbb{Q}$ is called the equivalent kernel.

In general, the article [16] mentions a way to construct the whole class of such equivalent kernels.

Recalling matrix properties will bring us few additional facts.

Lemma 1.1.14 *For two Hermitian kernels $\mathbb{K}_1, \mathbb{K}_2$, if second kernel dominates the first in a sense, that the matrix:*

$$\mathbb{K}_2 - \mathbb{K}_1 = (\mathbb{K}_2(x_i, x_j) - \mathbb{K}_1(x_i, x_j))_{i,j\in[1,k]}$$

is positive semidefinite, then:

$$\det\left(\mathbb{K}_1(x_i, x_j)_{i,j\in[1,k]}\right) \leq \det\left(\mathbb{K}_2(x_i, x_j)_{i,j\in[1,k]}\right),$$

so, from the definition, intensities of $\mathbb{K}_2$ dominate intensities of $\mathbb{K}_1$.

Another useful fact is connected to scaling.

Lemma 1.1.15 *If $\mathbb{K}_1 = \lambda\mathbb{K}_2$, for some $\lambda \in (0, 1)$, then:*

$$\det\left(\mathbb{K}_1(x_i, x_j)_{i,j\in[1,k]}\right) = \lambda^k \det\left(\mathbb{K}_2(x_i, x_j)_{i,j\in[1,k]}\right),$$

so, from the definition of the joint intensities, for disjoint subsets D_i of $\mathcal{D}$, we have the following;

$$\mathbb{E}\left[\prod_{i=1}^{k} \mathfrak{X}_1(D_i)\right] = \lambda^k \mathbb{E}\left[\prod_{i=1}^{k} \mathfrak{X}_2(D_i)\right],$$

where point processes $\mathfrak{X}_1, \mathfrak{X}_2$ were generated by kernels $\mathbb{K}_1, \mathbb{K}_2$ respectively.

In a similar manner we can demonstrate important connection between specific point processes.

Lemma 1.1.16 *If we independently delete each element of a random set distributed by a DPP $\mathfrak{X}_1$ with kernel $\mathbb{K}_1$ with probability $1 - \lambda$, for $\lambda \in (0, 1)$, we will obtain a random set with distribution according to a DPP $\mathfrak{X}_2$ generated by $\mathbb{K}_2 = \lambda\mathbb{K}_1$.*

Proof: From the construction of a random set with distribution according to point process $\mathfrak{X}_2$, for any set D_i of $\mathcal{D}$, we have the following:

$$\mathbb{E}(\mathfrak{X}_2(D_i)) = \mathbb{E}(\lambda\mathfrak{X}_1(D_i))$$

and we are deleting points independently, so:

$$\mathbb{E}\left(\prod_{i=1}^{k}\mathfrak{X}_2(D_i)\right) = \mathbb{E}\left(\prod_{i=1}^{k}\lambda\mathfrak{X}_1(D_i)\right)$$

for any mutually disjoint subsets $D_1, ..., D_k$ of $\mathcal{D}$. Furthermore:

$$\mathbb{E}(\prod_{i=1}^{k}\lambda\mathfrak{X}_1(D_i)) = \lambda^k\mathbb{E}(\prod_{i=1}^{k}\mathfrak{X}_1(D_i)) = \int_{\prod_i D_i}\lambda^k g_k(x_1, ..., x_k)d\mu(x_1)...d\mu(x_k),$$

so joint intensities of $\mathfrak{X}_2$ are joint intensities of $\mathfrak{X}_1$, scaled by a power of λ. From the definition of DPP it follows, that $\mathbb{K}_2 = \lambda\mathbb{K}_1$. QED
Next, we will show the application of the Fischer's inequality in the determinantal point processes.

Theorem 1.1.17 (Fischer's inequality) *Let M be a positive semidefinite $n \times n$ matrix and let A, B, C be matrices, so that A, C are squared and:*

$$M = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

where B^* is a matrix self-adjoint to matrix B. Then, we have the following inequality:

$$\det(M) \leq \det(A)\det(C).$$

The proof of this theorem can be found in the book [5, Theorem 7.8.5].

Lemma 1.1.18 *Let g_k, g_l, g_{k+l} denote the k, l and $k+l$-joint intensities of a determinantal point process with kernel $\mathbb{K}$. Then, as long as the correlation kernel matrix of g_{k+l} is Hermitian for any $x_1, ..., x_{k+l} \in \mathcal{D}$, we have the following inequality:*

$$g_k(x_1, ..., x_k)g_l(x_{k+1}, ..., x_{k+l}) \geq g_{k+l}(x_1, ..., x_{k+l}).$$

Proof: We take the following matrices:

$$M = \left(\mathbb{K}(x_i, x_j)_{i,j\in[1,k+l]}\right) \qquad A = \left(\mathbb{K}(x_i, x_j)_{i,j\in[1,k]}\right) \qquad C = \left(\mathbb{K}(x_i, x_j)_{i,j\in[k+1,k+l]}\right).$$

Since matrix M is Hermitian, from the Lemma 1.1.12, it is self-adjoint. We can apply the Fischer's inequality, to get the following:

$$\det(M) \leq \det(A)\det(C).$$

From the definition of a DPP:

$$\det(M) = g_{k+l}(x_1, ..., x_{k+l}) \qquad \det(A) = g_k(x_1, ..., x_k) \qquad \det(C) = g_l(x_{k+1}, ..., x_{k+l}),$$

so we get the following:

$$g_k(x_1, ..., x_k) g_l(x_{k+1}, ..., x_{k+l}) \geq g_{k+l}(x_1, ..., x_{k+l}).$$

QED

We will finish this chapter with several lemmas, that will prove useful in the next chapter. Lemma 1.1.19 is a well known fact, while Lemma 1.1.20 can be found in the article [10, Lemma 1].

Lemma 1.1.19 *Let Z be a multivariate normal random variable with mean 0 and covariance matrix Σ and let G_j be i.i.d. univariate standard normal random variables. We denote by $\lambda_1, ..., \lambda_k$ the eigenvalues of Σ in a non-decreasing order. Then the distribution of $||Z||^2$ is the same as the distribution of $\sum_{j=1}^{k} \lambda_j G_j^2$.*

Proof: From the spectral decomposition of matrix Σ, we can write the following:

$$\Sigma = Q \Lambda Q^T,$$

for some orthogonal matrix Q and diagonal matrix Λ, containing every eigenvalue of Σ. Because of that, we can write random variable Z, using the multivariate standard random variable G:

$$Z = \Sigma^{0.5} X = Q \Lambda^{0.5} G.$$

Since $Q^T Q = I$ and $\Lambda = \Lambda^T$ is a diagonal matrix, we have the following:

$$||Z||^2 = Z^T Z = (Q \Lambda^{0.5} G)^T (Q \Lambda^{0.5} G) = G^T \sqrt{\Lambda}^T Q^T Q \sqrt{\Lambda} G = G^T \Lambda G = \sum_k \lambda_k G_k^2.$$

The covariance matrix of G is an identity matrix, so for any $i \neq j$, G_i and G_j are independent, hence, $G_i \sim \mathcal{N}(0, 1)$. QED

Lemma 1.1.20 *[10, Lemma 1] Let $(Y_1, ..., Y_n)$ be i.i.d. Gaussian variables, with mean 0 and variance 1 and b be a nonnegative vector of size n. Denote by Z the following:*

$$Z = \sum_{i=1}^{n} b_i (Y_i^2 - 1).$$

Then, for any positive x we have the following:

$$\mathbb{P}(Z \geq 2||b||_2 \sqrt{x} + 2||b||_\infty x) \leq \exp(-x),$$

$$\mathbb{P}(Z \leq -2||b||_2 \sqrt{x}) \leq \exp(-x).$$

The definition of the norms $|| \cdot ||_2$, $|| \cdot ||_\infty$ is consistent with the usual mathematical notation, i.e.,:

$$||b||_\infty = \sup ||b_i||, \qquad ||b||_2^2 = \sum_{i=1}^{n} b_i^2.$$

1.2 Existence of determinantal point process

To prove the existence of DPP we need to introduce the integral operators of kernels.

Definition 1.2.1 *Let $\mathcal{D}$ and μ be as in Definition* 1.1.3 *and $\mathbb{K}$ be a locally square integrable kernel defined on $\mathcal{D}$, i.e.*

$$\int_{D^2} |\mathbb{K}(x,y)|^2 d\mu(x) d\mu(y) < \infty,$$

for any compact set $D \in \mathcal{D}$. Then the integral operator of $\mathbb{K}$, restricted to a set $D \subseteq \mathcal{D}$, denoted by $\mathcal{K}_D$, is an operator of the following form:

$$\mathcal{K}_D f(x) = \int_D \mathbb{K}(x,y) f(y) d\mu(y),$$

for almost every $x \in D$ and functions $f \in L^2(D,\mu)$ supported on D, aside of a set of measure 0*. If $D = \mathcal{D}$, $\mathcal{K}_{\mathcal{D}} = \mathcal{K}$ is called an integral operator.*

Since $L^2(D,\mu)$ is a Hilbert space, $\mathcal{K}$ has an orthonormal basis (ϕ_j) of the elements (functions) and corresponding eigenvalues (λ_j).

Definition 1.2.2 *For space $L^2(D,\mu)$ we denote by (λ_j) the eigenvalues associated with the orthonormal basis (ϕ_j) of $\mathcal{K}_D$. The integral operator $\mathcal{K}_D$ is of trace class if*

$$\sum_j |\lambda_j| < \infty.$$

The integral operator $\mathcal{K}_D$ is locally of trace class if this condition holds for every compact subset of D instead.

Utilizing the integral operator, we can express various kernels in a different way.

Theorem 1.2.3 *Assume that kernel $\mathbb{K}$ is Hermitian in $L^2(\mathcal{D}^2, \mu \otimes \mu)$ and its integral operator is of trace class. Then, there exists a set A of measure* 0*, so that for every $x, y \in \mathcal{D}/A$ we have the following equality:*

$$\mathbb{K}(x,y) = \sum_j \lambda_j \phi_j(x) \bar{\phi}_j(y),$$

for eigenvalues (λ_j) of orthonormal set (ϕ_j) in $L^2(\mathcal{D},\mu)$. In particular, this series converges absolutely, in $L^2(\mathcal{D},\mu)$ as a function of x or y, and in $L^2(\mathcal{D}^2, \mu \otimes \mu)$ as a function of x and y.

Proof: First, we will prove the convergence of the expression $\sum_j \lambda_j \phi_j(x) \bar{\phi}_j(y)$. Since (ϕ_j) is the orthonormal basis in $L^2(\mathcal{D},\mu)$, we have the following:

$$\int_{\mathcal{D}} \left(\sum_j |\lambda_j| |\phi_j(x)|^2 \right) d\mu(x) = \sum_j |\lambda_j| < \infty,$$

as the integral operator is of the trace class. This means, that aside of some set A of measure 0, this series converges pointwise for every $x \in \mathcal{D}/A$. We can apply the Cauchy-Schwartz inequality, to the original series:

$$\sum_{j\geq N} \lambda_j \phi_j(x)\bar{\phi}_j(y) \leq \left(\sum_{j\geq N} |\lambda_j||\phi_j(x)|^2\right)\left(\sum_{j\geq N} |\lambda_j||\phi_j(y)|^2\right),$$

to get the absolute convergence for any $x, y \in \mathcal{D}/A$. Additionally, if this series is a function of y:

$$h_N(y) = \sum_{j\geq N} \lambda_j \phi_j(x)\bar{\phi}_j(y),$$

for a fixed $x \in \mathcal{D}/A$, by utilizing the Cauchy-Schwartz inequality, we have the following:

$$\int_{\mathcal{D}} |h_N(y)|^2 d\mu(y) = \int_{\mathcal{D}} |\sum_{j\geq N} \lambda_j \phi_j(x)\bar{\phi}_j(y)|^2 d\mu(y) \leq$$

$$\leq \left(\sum_{j} |\lambda_j||\phi_j(x)|^2\right)\left(\sum_{j\geq N} |\lambda_j| \int_{\mathcal{D}} |\phi_j(y)|^2 d\mu(y)\right) = \left(\sum_{j} |\lambda_j||\phi_j(x)|^2\right)\left(\sum_{j\geq N} |\lambda_j|\right),$$

so this series is Cauchy in $L^2(\mathcal{D}, \mu)$. Analogously, we can prove that the function:

$$h_N(x, y) = \sum_{j\geq N} \lambda_j \phi_j(x)\bar{\phi}_j(y),$$

is Cauchy in $L^2(\mathcal{D}^2, \mu)$. We have the following:

$$\int_{\mathcal{D}^2} |h_N(x, y)|^2 d\mu(x) d\mu(y) = \left(\int_{\mathcal{D}} |\sum_{j\geq N} |\lambda_j||\phi_j(x)|^2 d\mu(x)\right)^2 \leq \left(\sum_{j\geq N} |\lambda_j|\right)^2.$$

Now, we will show that this series is equal to the kernel $\mathbb{K}$. From the definition of the integral operator we have the following:

$$\mathcal{K}f(x) = \int_{\mathcal{D}} \mathbb{K}(x, y) f(y) d\mu(y),$$

but, since (ϕ_j) is an orthonormal basis, we get the following:

$$\mathcal{K}f(x) = \sum_{j} \left(\int_{\mathcal{D}} f(y)\bar{\phi}_j(y) d\mu(y)\right) \lambda_j \phi_j(x),$$

as we can write any element of $L^2(\mathcal{D}, \mu)$ in terms of an orthonormal basis. For any fixed $x \in \mathcal{D}/A$, the expression:

$$\sum_{j} \lambda_j \phi_j(x)\bar{\phi}_j(y)$$

is finite, so we can exchange the sum with the integral in the following way:

$$\sum_j \left(\int_{\mathcal{D}} f(y)\bar{\phi}_j(y)d\mu(y) \right) \lambda_j \phi_j(x) = \int_{\mathcal{D}} \left(\sum_j f(y)\bar{\phi}_j(y)\lambda_j\phi_j(x) \right) d\mu(y) =$$

$$= \int_{\mathcal{D}} \left(\sum_j \lambda_j \phi_j(x)\bar{\phi}_j(y) \right) f(y)d\mu(y) = \int_{\mathcal{D}} \mathbb{K}(x,y)f(y)d\mu(y).$$

The last equality comes from the definition of an integral operator. Since this equality holds for any function in $L^2(\mathcal{D},\mu)$, we get that:

$$\mathbb{K}(x,y) = \sum_j \lambda_j \phi_j(x)\bar{\phi}_j(y),$$

for $\mu \otimes \mu - a.e.(x,y)$. QED

A determinantal point process can often be generated by various kernels. As such, we can alter given kernel to obtain important properties. The following change of kernel will prove vital for the remaining properties in this chapter.

Theorem 1.2.4 *Let $\mathfrak{X}$ be a DPP generated by Hermitian kernel $\mathbb{K}$ with integral operator $\mathcal{K}$ of trace class. We denote the orthonormal eigenvectors of $\mathcal{K}$ as (ϕ_j) and their corresponding eigenvalues as (λ_j), with $\lambda_j \in [0,1]$ for every j. From the Theorem* 1.2.3*, we have the following ($n \leq \infty$):*

$$\mathbb{K}(x,y) = \sum_{j=1}^{n} \lambda_j \phi_j(x)\bar{\phi}_j(y).$$

Let (I_j) be independent Bernoulli random variables, with corresponding parameters (λ_j). Then the following kernel:

$$\mathbb{K}_I(x,y) = \sum_{j=1}^{n} I_j \phi_j(x)\bar{\phi}_j(y),$$

generates a DPP $\mathfrak{X}_I$, that has the same distribution as $\mathfrak{X}$, i.e.:

$$\mathfrak{X} \stackrel{d}{=} \mathfrak{X}_I.$$

Proof: Joint intensities uniquely determine a point process [6, Lemma 4.2.6.], so to prove the equality $\mathfrak{X} \stackrel{d}{=} \mathfrak{X}_I$, we will prove that the joint intensities generated by $\mathbb{K}_I$ and $\mathbb{K}$ are equal.

Let us assume that $n < \infty$. From the example 1.1.11 we know that the process $\mathfrak{X}_I$ exists. We take an integer $m \leq n$ (in case $m > n$ the joint intensities g_m of both

processes are equal to 0). From the definition of DPP, to show the equality between joint intensities of two processes, it is enough to show that for any points $x_1, ..., x_m \in \mathcal{D}$

$$\mathbb{E}\left[\det\left(\mathbb{K}_I(x_i, x_j)_{i,j\in[1,m]}\right)\right] = \det\left(\mathbb{K}(x_i, x_j)_{i,j\in[1,m]}\right).$$

In order to prove this equality, we construct two matrices: A, B, so that $A_{i,k} = I_k\phi_k(x_i)$, $B_{k,j} = \bar{\phi}_k(x_j)$. Clearly $AB = \mathbb{K}_I(x_i, x_j)_{i,j\in[1,m]}$, so:

$$\mathbb{E}\left[\det\left(\mathbb{K}_I(x_i, x_j)_{i,j\in[1,m]}\right)\right] = \mathbb{E}\left[\det\left(AB\right)\right].$$

To finish this part of the proof we will utilize the Cauchy-Binet formula. We take a submatrix of A, by taking certain chosen columns and submatrix of B, by taking certain chosen rows:

$$\mathbb{E}\left[\det(AB)\right] = \sum_{i_1,...,i_m\in[1,n]} \mathbb{E}\left[\det(A[i_1, ..., i_m])\det(B\{i_1, ..., i_m\})\right] =$$

$$= \sum_{i_1,...,i_m\in[1,n]} \mathbb{E}\left[\det(A[i_1, ..., i_m])\right]\det(B\{i_1, ..., i_m\}),$$

since matrix B is deterministic. Now we choose an integer $l \in [1, m]$ and expand the determinant of the matrix $A[i_1, ..., i_m]$ by the column l (we denote the matrix $A[i_1, ..., i_m]$, without row i and column k by $A[i_1, ..., i_m]_{-i,-k}$):

$$\det(A[i_1, ..., i_m]) = \sum_{i=1}^{m}(-1)^{i+l}A[i_1, ..., i_m]_{i,l}\det(A[i_1, ..., i_m]_{-i,-l}).$$

Using this technique we can replace I_{i_l} with λ_{i_l}:

$$\mathbb{E}\left[\det\left(A[i_1, ..., i_m]\right)\right] = \sum_{i=1}^{m}(-1)^{i+l}\mathbb{E}\left[A[i_1, ..., i_m]_{i,l}\det(A[i_1, ..., i_m]_{-i,-l})\right] =$$

$$= \sum_{i=1}^{m}(-1)^{i+l}\mathbb{E}\left[A[i_1, ..., i_m]_{i,l}\right]\mathbb{E}\left[\det(A[i_1, ..., i_m]_{-i,-l})\right] =$$

$$= \sum_{i=1}^{m}\lambda_{i_l}\phi_l(x_i)(-1)^{i+l}\mathbb{E}\left[\det(A[i_1, ..., i_m]_{-i,-l})\right] = \mathbb{E}\left[\det\left(A[i_1, ..., i_m]_{\lambda_{i_l}}\right)\right],$$

where matrix $A[i_1, ..., i_m]_{\lambda_l}$ has the following elements:
$\triangleright \left(A[i_1, ..., i_m]_{\lambda_{i_l}}\right)_{j,k} = I_{i_k}\phi_{i_k}(x_j)$ if $k \neq l$ and
$\triangleright \left(A[i_1, ..., i_m]_{\lambda_{i_l}}\right)_{j,k} = \lambda_{i_k}\phi_{i_k}(x_j)$ otherwise.
We repeat this procedure for every $l \in [1, m]$ and every element of the sum. By multiplying the elements by $\det(B\{i_1, ..., i_m\})$ we obtain the equality:

$$\mathbb{E}\left[\det\left(\mathbb{K}_I(x_i, x_j)_{i,j\in[1,m]}\right)\right] = \det\left(\mathbb{K}(x_i, x_j)_{i,j\in[1,m]}\right).$$

We can apply the same proof to the case $n = \infty$, however, as we are dealing with infinite sums and infinite size matrices, we have to check few conditions. First, every element of matrix $\mathbb{K}(x_i, x_j)_{i,j\in[1,m]}$ is finite, as $\sum_k \lambda_k < \infty$. Because of that, we can apply example 1.1.11 to prove that process $\mathfrak{X}_I$ exists. The only obstacle left is the formula of Cauchy-Binet, since we should sum only finitely many elements. We take kernels of the form:

$$\mathbb{K}_N(x, y) = \sum_{k=1}^{N} \lambda_k \phi_k(x) \bar{\phi}_k(y)$$

and apply the formula to them. Since they converge to $\mathbb{K}(x, y)$ as $N \to \infty$, almost surely we have the following equation:

$$\mathbb{E}\left[\det\left(\mathbb{K}_I(x_i, x_j)_{i,j\in[1,m]}\right)\right] = \det\left(\mathbb{K}(x_i, x_j)_{i,j\in[1,m]}\right).$$

QED

From Definition 1.1.8 it is clear, that using kernel $\mathbb{K}$ one can construct functions g_k, which can create a point process as long as some conditions are met. From Theorem 1.1.7 we know what are sufficient and necessary conditions for functions g_k to be joint intensities of a point process, so it should suffice to check for which kernels $\mathbb{K}$, functions g_k are positive and symmetric. Using Theorem 1.2.4 we can state the conditions in a more general way.

Theorem 1.2.5 *We consider a Hermitian kernel $\mathbb{K}$ on $\mathcal{D}$ with integral operator $\mathcal{K}$ that is of locally trace class and self-adjoint. Then the kernel $\mathbb{K}$ generates a DPP on $\mathcal{D}$ iff the spectrum of $\mathcal{K}$ is contained in $[0, 1]$.*

Proof: Suppose first that $\mathcal{D}$ is a compact space and $\mathcal{K}$ is an operator of trace class. From the Lemma 1.2.3, we can write $\mathbb{K}$ in the following form:

$$\mathbb{K}(x, y) = \sum_{k} \lambda_k \phi_k(x) \bar{\phi}_k(y),$$

where (ϕ_k) are orthonormal and (λ_k) are the corresponding eigenvalues. Let $\mathfrak{X}$ be DPP with kernel $\mathbb{K}$. We will show, that every eigenvalue λ_k is contained in $[0, 1]$. Kernel $\mathbb{K}$ is positive semidefinite, because every joint intensity of process $\mathfrak{X}$ is non-negative. As such, every eigenvalue λ_k is non-negative. Let λ_{big} denote the biggest eigenvalue of $\mathcal{K}$ and assume, that $\lambda_{big} > 1$. We can utilize Lemma 1.1.16 with:

$$\lambda = \frac{1}{\lambda_{big}},$$

to create a DPP $\mathfrak{X}_1$ with every eigenvalue of the integral operator almost surely contained in $[0, 1]$. As operator $\mathcal{K}$ is of trace class, $\sum_j \lambda_j < \infty$, so $\mathfrak{X}(\mathcal{D})$ has finitely many points (we will denote the number of points by n). From the construction of $\mathfrak{X}_1$, we can infer, that:

$$\mathbb{P}\left(\mathfrak{X}_1(\mathcal{D}) = 0\right) = \left(1 - \frac{1}{\lambda_{big}}\right)^n > 0.$$

On the other hand, we can apply Theorem 1.2.4 to the process $\mathfrak{X}_1$, to create the process $\mathfrak{X}_{1,I} \stackrel{d}{=} \mathfrak{X}_1$. By the Theorem 1.3.1, number of points in $\mathfrak{X}_{1,I}(\mathcal{D})$ is equal to $\sum_k I_k$, for Bernoulli random variables I_k with parameter λ_k/λ_{big} from the Theorem 1.2.4. Since at least one eigenvalue is equal to 1, there exists a random variable I_k that is equal to 1 with probability 1 in this sum. As such:

$$\mathbb{P}\left(\mathfrak{X}_1(\mathcal{D}) \geq 1\right) = \mathbb{P}\left(\mathfrak{X}_{1,I}(\mathcal{D}) \geq 1\right) = 1,$$

since $\mathfrak{X}_{1,I} \stackrel{d}{=} \mathfrak{X}_1$. We have obtained a contradiction, so every value of λ has to be in $[0, 1]$.

Now we can proceed to prove the sufficiency of this statement. We take the kernel $\mathbb{K}$:

$$\mathbb{K}(x, y) = \sum_k \lambda_k \phi_k(x) \bar{\phi}_k(y)$$

and construct the kernel $\mathbb{K}_I$ using the construction from the proof of Theorem 1.2.4. In this way we have obtained:

$$\mathbb{K}_I = \sum_k I_k \phi_k(x) \bar{\phi}_k(y),$$

for Bernoulli random variables I_k with parameter λ_k each. As random variables I_k are either 0 or 1, kernel K_I is a projection kernel - kernel of the form:

$$\mathbb{K}_I = \sum_l \phi_l(x) \bar{\phi}_l(y).$$

From the Example 1.1.11 we can ascertain, that kernels of this form generate a DPP, so $\mathbb{K}_I$ generates a DPP $\mathfrak{X}_I$. From the proof of the Theorem 1.2.4, $\mathbb{K}_I$ generates the same joint intensities as $\mathbb{K}$, so the kernel $\mathbb{K}$ generates a DPP $\mathfrak{X} \stackrel{d}{=} \mathfrak{X}_I$.

We now consider non-compact space $\mathcal{D}$. From the assumptions on the space $\mathcal{D}$ (it is a locally compact Polish space), there exists an increasing sequence of compact subsets $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq ...$ that converges to $\mathcal{D}$. We define a kernel on $\mathcal{C}_N$, by:

$$\mathbb{K}_N(x, y) = \mathbb{1}_{x,y \in \mathcal{C}_N} \sum_k \lambda_k \phi_k(x) \bar{\phi}_k(y).$$

As its integral operator is of trace class, from the first part of the proof, we can conclude that $\mathbb{K}_N$ generates a DPP $\mathfrak{X}_N$ on $\mathcal{C}_N$. Note, that for any integer $k > 0$, $\mathfrak{X}_{N+k}$, restricted to space $\mathcal{C}_N$, generates the same DPP as $\mathbb{K}_N$, as for every $x, y \in \mathcal{C}_N$, $\mathbb{K}_{N+k}(x, y) = \mathbb{K}_N(x, y)$. From the definition of the joint intensities, for any DPP $\mathfrak{X}_N$ and disjoint sets $D_1, ..., D_k \subseteq \mathcal{C}_N$ we have the following:

$$\mathbb{E}_{\mathbb{P}_N}\left(\prod_{j=1}^{k} \mathfrak{X}_N(D_j)\right) = \int_{\prod_{j=1} D_j} \det\left(\mathbb{K}_N(x_i, x_j)_{i,j \in [1,k]}\right) d\mu(x_1)...d\mu(x_k),$$

where $\mathbb{P}_N$ is a measure associated with $\mathfrak{X}_N$. From the construction of the kernels $\mathbb{K}_N$:

$$\int_{\prod_{j=1} D_j} \det\left(\mathbb{K}_N(x_i,x_j)_{i,j\in[1,k]}\right) d\mu(x_1)...d\mu(x_k) =$$

$$= \int_{\prod_{j=1} D_j} \det\left(\mathbb{K}(x_i,x_j)_{i,j\in[1,k]}\right) d\mu(x_1)...d\mu(x_k).$$

On the other hand, from the Kolmogorov extension theorem we have the following:

$$\mathbb{E}_{\mathbb{P}_N}\left(\prod_{j=1}^{k}\mathfrak{X}_N(D_j)\right) \overset{N\to\infty}{\longrightarrow} \mathbb{E}_{\mathbb{P}}\left(\prod_{j=1}^{k}\mathfrak{X}(D_j)\right),$$

so for any N, for any disjoint sets $D_1,...,D_k \subseteq \mathcal{C}_N$:

$$\mathbb{E}_{\mathbb{P}}\left(\prod_{j=1}^{k}\mathfrak{X}(D_j)\right) = \int_{\prod_{j=1} D_j} \det\left(\mathbb{K}(x_i,x_j)_{i,j\in[1,k]}\right) d\mu(x_1)...d\mu(x_k).$$

Since it holds for any disjoint subsets of compact set $\mathcal{C}_N$, we can conclude, that this equation holds for any disjoint subsets of $\mathcal{D}$, as it is the limit of $\mathcal{C}_N$ as N goes to infinity. From this equation, we can ascertain that the process $\mathfrak{X}$ has the joint intensities:

$$g_k(x_1,...,x_k) = \det\left(\mathbb{K}(x_i,x_j)_{i,j\in[1,k]}\right),$$

so $\mathfrak{X}$ is a DPP generated by:

$$\mathbb{K}(x,y) = \mathbb{1}_{x,y\in\mathcal{D}}\sum_k \lambda_k\phi_k(x)\bar{\phi}_k(y) = \sum_k \lambda_k\phi_k(x)\bar{\phi}_k(y).$$

If there exists a DPP $\mathfrak{X}$ on $\mathcal{D}$, we can construct a sequence $(\mathfrak{X}_N)_{N=1}^{\infty}$ in the afore-described way. We will utilize the following fact from the functional analysis. Let the kernels $\mathbb{K}_1,...,\mathbb{K}$ be Hermitian, self-adjoint and defined on the Hilbert space. Then, for every kernel the following holds:

$$\lambda_{N,big} = \sup_{||f||=1}\{\langle\mathcal{K}_N f, f\rangle\},$$

$$\lambda_{N,small} = \inf_{||f||=1}\{\langle\mathcal{K}_N f, f\rangle\},$$

where $\lambda_{N,big}, \lambda_{N,small}$ are the largest and the smallest eigenvalues and $f \in L^2(\mathcal{D},\mu)$. Since kernels $\mathbb{K}_1,...,\mathbb{K}$ are Hermitian, self-adjoint and defined on the Hilbert space, we can utilize this fact. For every N, kernel $\mathbb{K}_N$ has every eigenvalue in the interval $[0,1]$. As such:

$$\langle\mathcal{K}_N f, f\rangle \in [0,1],$$

for every $f \in L^2(\mathcal{D},\mu)$. We have the following:

$$\langle\mathcal{K}_N f, f\rangle = \int\int_{\mathcal{D}^2} \mathbb{K}_N(x,y)f(y)f(x)d\mu(x)d\mu(y) =$$

$$= \int \int_{\mathcal{D}^2} \mathbb{1}_{x,y \in \mathcal{C}_N} \mathbb{K}(x,y) f(y) f(x) d\mu(x) d\mu(y) = \int \int_{\mathcal{D}^2} \mathbb{K}(x,y) f_N(y) f_N(x) d\mu(x) d\mu(y) =$$
$$= \langle \mathcal{K} f_N, f_N \rangle ,$$

where $f_N(x) = f(x)\mathbb{1}_{x \in \mathcal{C}_N}$. From the construction of f_N, f_N converges to f in norm, so by the continuity of the integral bounded operator, we have the convergence:

$$\langle \mathcal{K} f_N, f_N \rangle \longrightarrow \langle \mathcal{K} f, f \rangle ,$$

for any $f \in L^2(\mathcal{D}, \mu)$. Since $\langle \mathcal{K} f_N, f_N \rangle \in [0,1]$, it follows, that the limit is also contained in this interval.

QED

Using Theorem 1.2.4 we can make another interesting observation regarding a sequence of determinantal point processes $\mathfrak{X}_n$ — given few easy to fulfill assumptions, we can state a central limit theorem for DPPs. First, we need to introduce triangular arrays. We consider an array of random variables $X_{i,j}$ over $\mathbb{R}$ of the form:

$$X_{1,1}, X_{1,2}, ..., X_{1,N_1}$$
$$X_{2,1}, X_{2,2}, ..., X_{2,N_2}$$
$$X_{3,1}, X_{3,2}, ..., X_{3,N_3}$$
$$...$$

so that
▷Random variables from the same row are mutually independent,
▷expected value of each random variable is 0 and
▷sum of the variances in the row is finite for every row.
While often the relation $N_1 < N_2 < ...$ is implied (making the array triangular), it is not required. We then take sums of the form

$$\sum_{j=1}^{N_i} X_{i,j},$$

and obtain some properties by the following theorem, which can be found in [14, Theorem 3].

Theorem 1.2.6 (Lindebergs Theorem) *Let $X_{i,j}$ be a triangular array, that satisfies the Lindeberg condition:*

$$(\forall \epsilon > 0) s_i^{-2} \sum_{j=1}^{N_i} \mathbb{E}\left[X_{i,j}^2 \mathbb{1}_{|X_{i,j}| \geq \epsilon s_i} \right] \overset{i \to \infty}{\longrightarrow} 0,$$

where $s_i^2 = Var\left(\sum_{j=1}^{N_i} X_{i,j} \right)$. Then:

$$s_i^{-1} \sum_{j=1}^{N_i} X_{i,j} \overset{d}{\to} \mathcal{N}(0,1).$$

The proof of this theorem, found in the book [1, Theorem 27.2], relies on the Lyapunov central limit theorem, the proof of which can also be found in this book.

Theorem 1.2.7 (Lyapunov Central Limit Theorem) *Let $X_1, ..., X_n, ...$ be a sequence of independent random variables with expected values $\mu_n < \infty$ and variances $\sigma_n^2 < \infty$ respectively. Assume, that the Lyapunov condition is satisfied, i.e.,:*

$$\frac{\sum_{i=1}^{n} \mathbb{E}\left(|X_i - \mu_i|^{2+\delta}\right)}{s_n^{2+\delta}} \xrightarrow{n\to\infty} 0,$$

for some $\delta > 0$ and $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$. Then we have the following convergence in distribution:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i - \mu_i)}{s_n} \overset{d}{=} \mathcal{N}(0,1).$$

Theorem 1.2.8 *Assume we have a sequence of DPPs $\mathfrak{X}_n$ on $\mathcal{D}_n$ with kernels $\mathbb{K}_n$ and its integral operators $\mathcal{K}_n$ locally of trace class. Denote by D_n compact Borel subsets of $\mathcal{D}_n$, such that $\mathfrak{X}_n(D_n)$ is finite a.s. and $Var(\mathfrak{X}_n(D_n)) \xrightarrow{n\to\infty} \infty$. Under those circumstances, we have the following:*

$$\lim_{n\to\infty} \frac{\mathfrak{X}_n(D_n) - \mathbb{E}\left[\mathfrak{X}_n(D_n)\right]}{\sqrt{Var(\mathfrak{X}_n(D_n))}} \overset{d}{=} \mathcal{N}(0,1).$$

Proof: We have proven Lindebergs theorem, however, the remaining question is how to apply it to prove Theorem 1.2.8. We have elements of the form $\mathfrak{X}_n(D_n)$. From Theorem 1.2.5 we can express this object in a different way:

$$\mathfrak{X}_n(D_n) \overset{d}{=} \mathfrak{X}_{n,I}(D_n)$$

and since $\mathfrak{X}_{n,I}$ is a projection process, from Theorem 1.3.1 we have the following:

$$\mathfrak{X}_{n,I}(D_n) \overset{d}{=} \sum_{k=1}^{m_n} I_{n,k},$$

where $I_{n,k}$ is a Bernoulli random variable with parameter $\lambda_{n,k}$. We can calculate the expected value and variance of $\sum_{k=1}^{m_n} I_{n,k}$:

$$\mathbb{E}\left[\sum_{k=1}^{m_n} I_{n,k}\right] = \sum_{k=1}^{m_n} \lambda_{n,k},$$

$$Var\left[\sum_{k=1}^{m_n} I_{n,k}\right] = \sum_{k=1}^{m_n} \lambda_{n,k}(1 - \lambda_{n,k}).$$

We create the aforedescribed array, where the sum of each row is equal to $\sum_{k=1}^{m_n} I_{n,k} - \lambda_{n,k}$ ($\mathbb{E}\left[I_{n,k} - \lambda_{n,k}\right] = 0$). In order to apply the Lindebergs theorem, we have to check that the Lindebergs condition holds, i.e.:

$$(\forall \epsilon > 0) s_n^{-2} \sum_{k=1}^{m_n} \mathbb{E}\left[(I_{n,k} - \lambda_{n,k})^2 \mathbb{1}_{|(I_{n,k} - \lambda_{n,k})| \geq \epsilon s_n}\right] \to^{n\to\infty} 0,$$

where $s_n = Var\left(\sum_{k=1}^{m_n} I_{n,k} - \lambda_{n,k}\right)$. We get the following:

$$Var\left(\sum_{k=1}^{m_n} I_{n,k} - \lambda_{n,k}\right) = Var\left(\sum_{k=1}^{m_n} I_{n,k}\right),$$

as $\lambda_{n,k}$ is deterministic for every n, k. Since

$$Var(\mathfrak{X}_n(D_n)) \xrightarrow{n\to\infty} \infty$$

and

$$\mathfrak{X}_{n,I}(D_n) \stackrel{d}{=} \sum_{k=1}^{m_n} I_{n,k},$$

the variance: $Var\left(\sum_{k=1}^{m_n} I_{n,k}\right)$ goes to infinity. As a consequence, for every $\epsilon > 0$ there exists n, so that for every $p > n$, we have the following:

$$s_p\epsilon > |I_{p,k} - \lambda_{p,k}|$$

almost surely. This implies that the expression:

$$\mathbb{E}\left[(I_{p,k} - \lambda_{p,k})^2 \mathbb{1}_{|(I_{p,k}-\lambda_{p,k})|\geq\epsilon s_p}\right]$$

is equal to 0. In conclusion, for every ϵ, there exists n, so that for every $p > n$:

$$s_p^{-2}\sum_{k=1}^{m_p} \mathbb{E}\left[(I_{p,k} - \lambda_{p,k})^2 \mathbb{1}_{|(I_{p,k}-\lambda_{p,k})|\geq\epsilon s_p}\right] = 0,$$

hence, as $p \to \infty$, this expression converges to 0 and the Lindebergs condition holds.

From the Theorem 1.2.6, we have the following:

$$\frac{\sum_{k=1}^{m_n} I_{n,k} - \lambda_{n,k}}{\sqrt{\sum_{k=1}^{m_n} \lambda_{n,k}(1-\lambda_{n,k})}} \xrightarrow{d} \mathcal{N}(0,1),$$

as n goes to infinity, but:

$$\frac{\sum_{k=1}^{m_n} I_{n,k} - \lambda_{n,k}}{\sqrt{\sum_{k=1}^{m_n} \lambda_{n,k}(1-\lambda_{n,k})}} \stackrel{d}{=} \frac{\mathfrak{X}_n(D_n) - \mathbb{E}\left[\mathfrak{X}_n(D_n)\right]}{\sqrt{Var(\mathfrak{X}_n(D_n))}},$$

so:

$$\frac{\mathfrak{X}_n(D_n) - \mathbb{E}\left[\mathfrak{X}_n(D_n)\right]}{\sqrt{Var(\mathfrak{X}_n(D_n))}} \xrightarrow{d} \mathcal{N}(0,1).$$

QED

1.3 Generating determinantal point process

Since we are working with probabilistic object and we have already checked when it exists, our next step is to generate a sample from a determinantal point process. How do we do that? In the Figure 1.1.4 we have already shown examples of a binomial and a homogenous Poisson point process. They were easy to implement - we simply distributed n points randomly, where n was either given (binomial), or followed Poisson distribution in the latter. In the determinantal case the number of points will be our first obstacle, however we will prove that with some additional conditions the number of points on a set almost surely coincides with its expectation.

Theorem 1.3.1 *Let $\mathfrak{X}$ be a DPP on $\mathcal{D}$ with finite projection kernel $\mathbb{K}$, i.e., a kernel of the form*

$$\mathbb{K}(x,y) = \sum_{k=1}^{n} \phi_k(x)\bar{\phi}_k(y),$$

for orthonormal $(\phi_k)_{k=1}^n$ and finite n. Then:

$$\mathbb{P}\left(\mathfrak{X}(\mathcal{D}) = n\right) = 1.$$

From Theorem 1.2.4 we can deduce that as long as $\mathbb{K}$ is Hermitian, with integral operator $\mathcal{K}$ of trace class, every DPP is equal in distribution to the DPP with the following kernel:

$$\mathbb{K}_I(x,y) = \sum_{j=1}^{n} I_j \phi_j(x)\bar{\phi}_j(y),$$

for Bernoulli random variables I_k. Since these random variables take the value 0 or 1, this determinantal process is a mixture of DPPs with projection kernels.
Proof: From the joint intensities and kernel assumption, we know that:

$$\mathbb{E}(\mathfrak{X}(\mathcal{D})) = \int_{\mathcal{D}} \mathbb{K}(x,x)d\mu(x) = \sum_{k=1}^{n} 1 = n.$$

On the other hand we have the following:

$$(\mathbb{K}(x_i,x_j))_{i,j\in[1,k]} = \left(\sum_{k=1}^{n} \phi_k(x_i)\bar{\phi}_k(x_j)\right)_{i,j\in[1,k]} = (\phi_j(x_i))_{i\leq k, j\leq n}\left(\bar{\phi}_i(x_j)\right)_{i\leq n, j\leq k}.$$

This factorization shows, that $(\mathbb{K}(x_i,x_j))_{i,j\in[1,k]}$ is a matrix of rank no greater than n, since the rank of a product of matrices is at most the rank of either factor. From the Definition 1.1.8 for any $k > n$, $g_k = 0$, as the determinant of k size matrix of rank n is equal to 0. Assume, that with positive probability $\mathfrak{X}(\mathcal{D}) > n$. Then we can find $n+1$ disjoint subsets D_i of $\mathcal{D}$, so that:

$$\mathbb{P}\left(\forall i \in \{1,...,n+1\} \quad \mathfrak{X}(D_i) > 0\right) > 0.$$

Since $\mathfrak{X}(D_i)$ is non-negative, we get the following:

$$\mathbb{E}\left[\prod_{i=1}^{n+1}\mathfrak{X}(D_i)\right] > 0.$$

On the other hand $g_{n+1} = 0$, so we get the contradiction from the definition of joint intensities. If $\mathfrak{X}(\mathcal{D}) \leq n - \epsilon$ with positive probability for some $\epsilon > 0$, then the following occurs:

$$n = \mathbb{E}(\mathfrak{X}(\mathcal{D})) \leq \mathbb{P}(\mathfrak{X}(\mathcal{D}) = n)n + \mathbb{P}(\mathfrak{X}(\mathcal{D}) < n)(n - \epsilon) = n - \epsilon\mathbb{P}(\mathfrak{X}(\mathcal{D}) < n) < n,$$

which gives us a contradiction. This means, that for any $\epsilon > 0$, $\mathbb{P}(\mathfrak{X}(\mathcal{D}) \leq n - \epsilon) = 0$, so $\mathbb{P}(\mathfrak{X}(\mathcal{D}) = n) = 1$. QED

We know how many random points we have in our sample, so the only thing left to find out is what is the distribution of these points. We will restrict ourselves to the case with a DPP with a Hermitian kernel.

Theorem 1.3.2 *Let $H \subseteq L^2(\mathcal{D}, \mu)$ be a space spanned by the orthonormal set $\{\phi_1, ..., \phi_k\}$ and $\mathbb{K}$ be a Hermitian kernel with real values for every $x_i, x_j \in \mathcal{D}$. The following algorithm generates a sample from the DPP $\mathfrak{X}$ with a projection kernel $\mathbb{K}$ on the space H.*
▷Input: space $H = H_n$, kernel $\mathbb{K}$
▷$n := dim(H)$
▷While $n > 0$
▷Take a sample of $X_n \sim \frac{\mu_{H_n}}{n}$
▷Take the orthogonal complement of the function $\mathcal{K}_{H_n}\delta_{X_n} := \mathbb{K}(\cdot, X_n)$ in H_n
▷Decrease n by 1
▷Output: $(X_1, ..., X_n)$

Proof: We start by defining the measure μ_{H_k} for k from 1 to n. As in every iteration we are restricted to the space H_k, we have to use a probabilistic measure defined on this space. We can write the following from the definition of the first intensity measure g_1 and kernel $\mathbb{K}$:

$$d\mu_{H_k}(x) = g_1(x)d\mu(x) = \mathbb{K}(x, x)d\mu(x).$$

Next, using an integral operator $\mathcal{K}_{H_k}$ of $\mathbb{K}$ restricted to H_k, we can introduce a new object: $\mathcal{K}_{H_k}\delta_x$. If μ is an atomic measure and $\mu(x) > 0$, we define δ_x as a delta function, i.e.,:
$\delta_x(y) = \frac{1}{\mu(x)}$ if $y = x$ and
$\delta_x(y) = 0$ otherwise.
We get the following:

$$\mathcal{K}_{H_k}\delta_x(z) = \int_{H_k}\mathbb{K}(z, y)\delta_x(y)d\mu(y) = \mathbb{K}(z, x)\delta_x(x)\mu(x) = \mathbb{K}(z, x),$$

for $x \in H_k$. If $x \notin H_k$, this expression is equal to 0, as for every point $y \in H_k$ $\delta_x(y) = 0$. If the measure μ is non-atomic, we cannot define a delta function this way, so instead we use the notation $\mathcal{K}_{H_k}\delta_x$ for $\mathbb{K}(\cdot, x)$ and treat these objects as equal, as $\mathcal{K}_{H_k}\delta_x = \mathbb{K}(\cdot, x)$.

The kernel $\mathbb{K}$ is Hermitian, so $\mathbb{K}(x, y) = \mathbb{K}(y, x)$. We can write the following:

$$\mathbb{K}(x,x) = \int \mathbb{K}(x,y)\mathbb{K}(y,x)d\mu(y) =$$

$$= \int \mathbb{K}^2(x,y)d\mu(y) = \int \left(\mathcal{K}_{H_k}\delta_x(y)\right)^2 d\mu(y) = ||\mathcal{K}_{H_k}\delta_x||^2_{L^2(\mu)},$$

which gives us the expression for $d\mu_{H_k}$:

$$d\mu_{H_k}(x) = ||\mathcal{K}_{H_k}\delta_x||^2_{L^2(\mu)}d\mu(x).$$

Since μ_{H_k} is a finite measure, we can divide it by the measure of the whole space to obtain the probability measure μ_{H_k}/k. Knowing the space H_k, we define space $H_{k-1} \subseteq H_k$ as the orthogonal complement of the function $\mathcal{K}_{H_k}\delta_{X_k}$ ($dim(H_{k-1}) = k-1$).

Next, we will check that the points $X_1, ..., X_n$ generated by this algorithm have the correct $n-$point joint intensity. From the definition of μ_{H_k}, the density of the random vector $(X_1, ..., X_n)$ is the following (since $H_k \subseteq H$, the projection onto H_k is equivalent to the projection onto H and then onto H_k):

$$f_n(x_1, ..., x_n) = \prod_{k=1}^{n} \frac{||\mathcal{K}_{H_k}\delta_{x_k}||^2}{k} = \frac{\prod_{k=1}^{n} ||\mathcal{K}_{H_k}\mathcal{K}_H\delta_{x_k}||^2}{n!}.$$

We define $\psi_k = \mathcal{K}_H\delta_{x_j}$. From the construction, H_k is the intersection of H and the orthogonal complement of the space spanned by $\psi_{k+1}, \ldots, \psi_n$. Since $\mathcal{K}_{H_j}\psi_k$ is contained in H_j, it is the projection of ψ_j to the orthogonal complement of the space spanned by vectors $\psi_{j+1}, \ldots, \psi_n$. Hence, the expression $\prod_{k=1}^{n}\|\mathcal{K}_{H_k}\psi_k\|$ can be interpreted as a repeated "base time height" formula for the volume of the parallelepiped, determined by vectors $\psi_1, \ldots, \psi_n$ in the space H, in other words:

$$\prod_{k=1}^{n}\|\mathcal{K}_{H_k}\psi_k\| = |\det(\psi_1, ..., \psi_n)|.$$

Now, we will use the fact from linear algebra, that $|\det(\psi_1, ..., \psi_n)|^2$ is equal to the determinant of the Gramm matrix $G(i,j) = \langle\psi_i, \psi_j\rangle$. The proof can be found in the book [13, page 214 of the second version]. We get the following:

$$f_n(x_1, ..., x_n) = \frac{\det\left(\left(\langle\psi_i, \psi_j\rangle\right)_{i,j\in[1,n]}\right)}{n!} = \frac{\det\left(\mathbb{K}(x_i, x_j)_{i,j\in[1,n]}\right)}{n!}.$$

The last equality comes from the definition of ψ_j. Since

$$\psi_j = \mathcal{K}_H\delta_{x_j} = \mathbb{K}(\cdot, x_j),$$

and kernel $\mathbb{K}$ is symmetric, we have the following:

$$\langle \psi_i, \psi_j \rangle = \int \mathbb{K}(x_i, y)\mathbb{K}(y, x_j)d\mu(y) = \mathbb{K}(x_i, x_j).$$

Next, we will analyze the expression $\prod_{i=1}^{n} \mathfrak{X}(D_i)$ for the disjoint sets $D_i \subseteq \mathcal{D}$. Since almost surely the number of points is n, this expresion is only positive if each set D_i contains exactly one point. We denote the set of all possible permutations of integers from 1 to n by S_n. We have the following:

$$\mathbb{E}\left[\prod_{i=1}^{n} \mathfrak{X}(D_i)\right] = 1 \cdot \sum_{\sigma \in S_n} \mathbb{P}(x_{\sigma(1)} \in D_1, ..., x_{\sigma(n)} \in D_n) =$$

$$= \sum_{\sigma \in S_n} \int_{\prod_i D_i} f_n(x_{\sigma(1)}, ..., x_{\sigma(n)})d\mu(x_1)...d\mu(x_n).$$

Since every permutation has the same value of f_n and S_n contains $n!$ elements, we have the following:

$$\sum_{\sigma \in S_n} \int_{\prod_i D_i} f_n(x_{\sigma(1)}, ..., x_{\sigma(n)})d\mu(x_1)...d\mu(x_n) = \int_{\prod_i D_i} n! f_n(x_{\sigma(1)}, ..., x_{\sigma(n)})d\mu(x_1)...d\mu(x_n).$$

On the other hand, from the definition of joint intensities:

$$\mathbb{E}\left[\prod_{i=1}^{n} \mathfrak{X}(D_i)\right] = \int_{\prod_i D_i} g_n(x_1, ..., x_k)d\mu(x_1)...d\mu(x_n),$$

so $g_n = n! f_n$. In conclusion:

$$g_n(x_1, ..., x_n) = n! f_n(x_1, ..., x_n) = n! \frac{\det(\mathbb{K}(x_i, x_j)_{i,j \in [1,n]})}{n!} = \det(\mathbb{K}(x_i, x_j)_{i,j \in [1,n]}).$$

We have obtained the joint intensity of the DPP generated by $\mathbb{K}$, so this joint intensity generates the determinantal point process $\mathfrak{X}$. QED

2 Omitting curse of dimensionality

While looking at the graphic representation of samples in the previous chapter, one may notice a hurdle in representing data - showing higher dimensional samples is difficult. One may project to reduce the number of dimensions (usually to 2), but then we need a lot of few-dimensional plots to see all the correlations ($n(n-1)/2$ in 2-dimensional scenario). This is one of many problems when dealing with high-dimensional data, *Richard Bellman* coined a term *curse of dimensionality* to describe this kind of problems. Some examples may include the following:

▷In combinatorics, given d variables with n values, the number of all the combinations is n^d,
▷In machine learning, the number of data points required for good performance usually grows exponentially.

Curse of dimensionality is a term well-known in a vast number of branches of mathematics, so, naturally, many ways to avoid it were invented. In this chapter we will explore methods to reduce the number of dimensions in sample data, often to 2 in order to represent it graphically. We will briefly summarize few popular methods, then move on to the main topic of this thesis - the method that requires Gaussian determinantal processes.

2.1 Simple projection

The first method is the most trivial one - we simply remove the unnecesary dimensions, until 2 most interesting remain. For example, if we have a table of n students and their scores in k subjects, we would leave two of them, e.g. math and physics and represent them on a graph. While there is usually some correlation between scores in those two subjects, one can easily pinpoint the flaw of this method - only a small part of our data and correlations is being shown. We could treat every two subjects in the same way, but processing and showing everything in this way may prove tedious. One way to resolve this problem is to find benefitial for us spatial dependency of data first and then reduce the number of dimensions in some way. This approach will be shown in the main method.

2.2 Factor analysis

In our next method we will implement some spatial dependency, thus representing the data after dimension reduction more effectively. The matrix X of k, n-dimensional samples can be expressed in a following way:

$$X_i = \sum_{j=1}^{k} Z_{i,j} F_j + \epsilon_i,$$

for matrixes F, Z of sizes $m \times n, k \times m$ respectively and noise vector ϵ. This way we can reduce the number of dimensions to m. F is called factor matrix and Z - factor loading matrix. The question that remains is how to find matrixes, that will preserve spatial dependency good enough. We need to deduce some dependency beforehand, but basically we try to represent each element as weighted sum of factors, with weights depending on relevancy. If factors are uncorrelated (i.e. $cov(F) = I$), we cannot reduce the number of dimensions without a heavy loss of information.

There are different methods of finding factors, classified into two categories - exploratory and confirmatory factor analysis. The first one comes from analysing values of random variables, while the latter - from assuming values and then checking their validity and estimation. The main methods are: principal factor analysis and principal component analysis. In the next chapter we will take a brief look at the latter. Different methods from factor analysis, as well as other additional informations can be found in the book The Essentials of Factor Analysis [2].

2.3 Principal component analysis

As mentioned in this chapter, we can use spatial dependency to try to minimize the amount of lost information. When hearing the term 'spatial dependency', one would likely think of covariation or correlation. Indeed, principal component analysis bases the analysis on eigenvalues of covariation (or correlation) matrix. This way we can rotate the coordinate system and by that, maximize the variance of first coordinate, then that of the second one and the rest in turn. We reduce the number of dimensions by removing the most dependent data. The algorithm of PCA works in a following way:

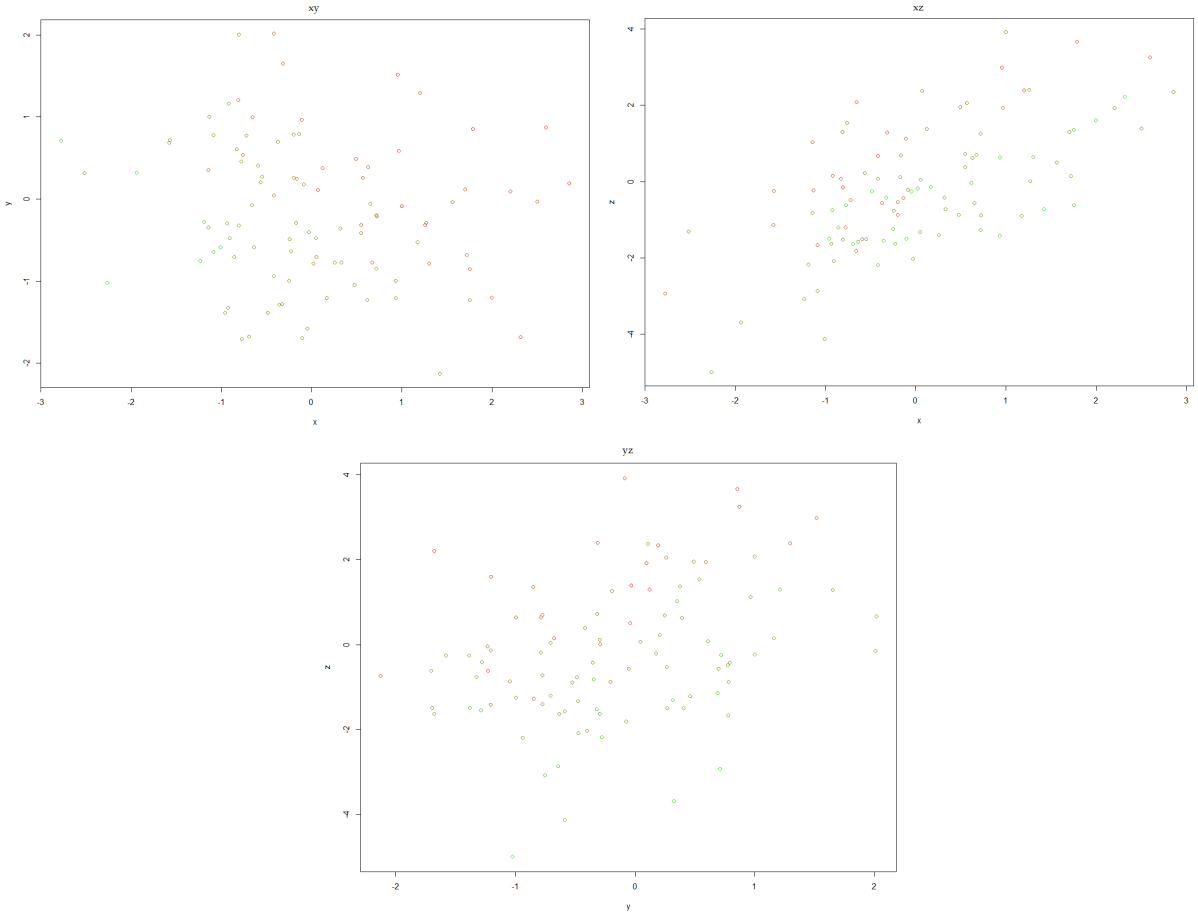
▷Calculate the mean of each row of the matrix
▷Remove the row mean from each element
▷Calculate the covariance (or correlation) matrix
▷Calculate the eigenvalues
▷Choose the right eigenvalues
▷Calculate the eigenvectors
▷Project onto the new less-dimensional space

While looking at this algorithm one may have two questions - what are the 'right eigenvalues' and what is the difference between choosing covariance and correlation matrix.

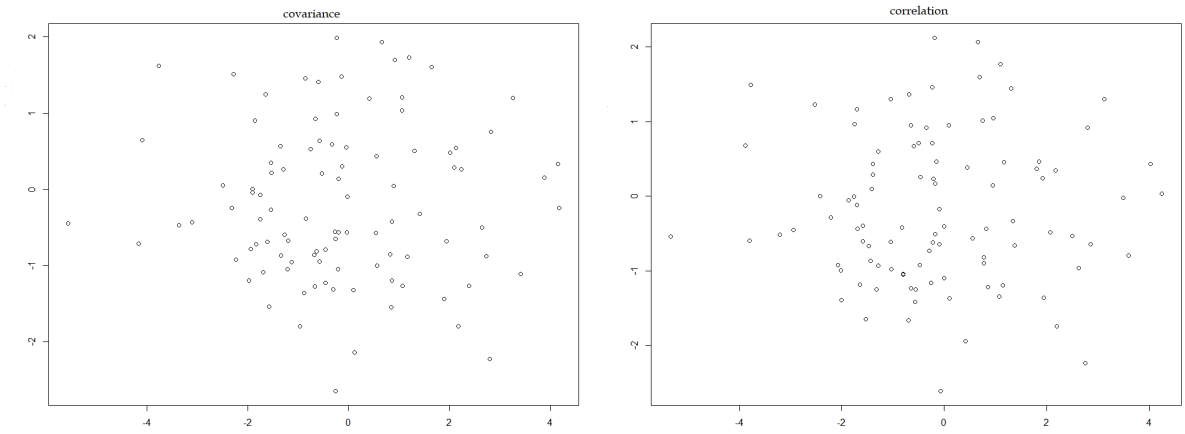
The answer to the first question is simple - we reduce the number of dimensions by removing the smallest eigenvalues. This way we remove the most correlated data. The number of removed eigenvalues is the number of lost dimensions.

While the algorithm is identical irregardless of the matrix we choose, the results we obtain will usually differ. In the case with covariation matrix, the variables with biggest variance will have more influence on the results, so this method should be used on data of similar sizes or when we want to put an emphasis on the variables with bigger variance. In the other case, the values will be scaled, to manage the difference in size.

Let the data consist of 100 sample points of the form (x, y, z), where x and y are sample points of $i.i.d$ random variables $X, Y \sim \mathcal{N}(0, 1)$ and z is a sample of $X + Y + Z$, for $Z \sim \mathcal{N}(0, 1)$. We can represent the data on $3d$ plot, however for clarity we will look at 3 $2-$ dimensional projections, with the third coordinate as a colour (going from green to red).



Using PCA method we can represent it in the form of one $2d$ plot. First figure was made using covariance matrix and second - using correlation matrix. While the graphics seem similar, one is visibly more stretched. Indeed, the third coordinate of the original sample is the sum of three variables $\mathcal{N}(0,1)$, so corresponding values in covariance matrix have visibly larger values. The method which requires correlation matrix regulates this, by rescaling the values.



This section contains the summary of the principal component analysis method. One can explore it further in the book "Principal component analysis" [9].

2.4 Gaussian determinantal process method

Finally we take a look at the main dimension reduction method of this chapter, which is based on an article of Subhroshekhar Ghosh and Philippe Rigollet [4] and utilizes knowledge gained in the previous section, regarding Gaussian determinantal process. While this method also requires spatial dependency, it produces different results than PCA, we will compare those methods in another section.

How do we capture spatial dependency in this method? We will reiterate the Definition 1.1.9:

Definition 2.4.1 *A determinantal process $\mathfrak{X}$ with a Gaussian kernel $\mathbb{K}$, i.e.:*

$$\mathbb{K}(x, y) = \Phi(x - y),$$

for $\Phi(x)$ being the density of a Gaussian vector with mean 0 and positive definite covariance matrix, is known as a Gaussian determinantal process (further referred to as GDP). In this case we call the covariance matrix a scattering matrix of $\mathfrak{X}$.

Since kernel $\mathbb{K}$ has to follow Gaussian density, the only parameter capable of catching the correlation is the covariance matrix. While the previous method reduces number of dimensions by reducing correlation, this method uses the scattering matrix to capture the directionality of data. Its eigenvectors, associated to its largest eigenvalues, span low-dimensional spaces along which the DPP exhibits the most repulsion.

To better understand the connection between scattering matrix and repulsion, let us take a look at two-dimensional case. We can pinpoint correlation rate by taking

$\bar{g}(x,y) = g(x,y) - g(x)g(y)$ (g denotes the intensity measure). From the definition of GDP, we have the following:

$$\bar{g}(x,y) = \mathbb{K}(y,y)\mathbb{K}(x,x) - \mathbb{K}(x,y)^2 - \mathbb{K}(x,x)\mathbb{K}(y,y) = -|\mathbb{K}(x,y)|^2 =$$

$$= -\frac{\exp(-(x-y)^T\Sigma^{-1}(x-y))}{(2\pi)^2 \det\Sigma}.$$

Since the expression is non-positive, we will capture spatial repulsion. In particular the smaller the expression, the bigger the repulsion.

In this thesis we are more interested in spatial dependency rather than spatial scaling, so we will assume the normalization $\mathbb{K}(x,x) = 1$. What happens if we do not have the normalization? Let us take a look at two covariance matrices: A and $B = cA$, for some $c > 1$. Then:

$$\bar{g}_B(x,y) = -\frac{\exp(-(x-y)^T B^{-1}(x-y))}{(2\pi)^2 \det B} = -\frac{\exp(-(x-y)^T c^{-1}A^{-1}(x-y))}{(2\pi)^2 c \det A},$$

hence, in most cases we will obtain expression bigger in magnitude, resulting in stronger repulsion, while most likely preserving the spatial dependency.

Let us take a look at the trivial case $\Sigma = I$. Then our expression gets visibly simplified:

$$\bar{g}(x,y) = -\frac{\exp\left(-||x-y||^2\right)}{4\pi^2}$$

and we can easily observe the repulsion rate using the distance between variables. The bigger the distance, the smaller the function $\bar{g}$, resulting in smaller dependency. With very small values of $||x-y||^2$ we can expect $\bar{g}(x,y)$ to be very close to $-1/(4\pi^2)$. We can expect similar results as long as $x-y$ is eigenvector of the inverse of scattering matrix. Indeed, since for the matrix A, its eigenvector v and the corresponding eigenvalue λ, $Av = \lambda v$ and the scattering matrix is symmetric, we have the following:

$$-\frac{\exp(-(x-y)^T\Sigma^{-1}(x-y))}{(2\pi)^2 \det\Sigma} = -\frac{\exp(-\lambda||x-y||^2)}{(2\pi)^2 \det\Sigma},$$

meaning, that the rate of repulsion depends, as before, only on the distance between x and y. While the repulsion is lesser for bigger values of $||x-y||^2$, it can still be significant along some directions. In this method we will analyze this spike. We will try to observe if it exists and, if so, how big it is.

2.5 Statistical estimation

In order to identify the aforementioned spike, first we need to construct the scattering matrix from the given dataset. By stationarity, the sample of GDP over $\mathbb{R}^d$ will almost

surely have an infinite amount of points, so, since datasets contain finite number of points, we will take the realization of GDP at the $d-$dimensional ball of radius R around 0. In this approach, we will calculate the estimator $\widehat{\Sigma}$ on local neighbourhoods of points (balls of radius r) and then average it. By translation invariance of the point process, in this way we can obtain a very good approximation of the scattering matrix of GDP over $\mathbb{R}^d$. From the proof of the Theorem 1.3.1, we can conclude that the expected number of points in the ball of radius R around 0 is $|B(R)|$. Because of that, our empirical estimator will contain identity matrix multiplied by $|B(R)|$. By inserting the aforementioned average, we obtain the following equation:

Equation 2.5.1

$$\widehat{\Sigma} = \frac{|B(1)|r^{d+2}}{d+2} I - \frac{1}{|B(R-r)|} \sum_{i \in \mathcal{N}_0} \sum_{j \in \mathcal{N}_i} (X_i - X_j)(X_i - X_j)^T .$$

$$\mathcal{N}_i = \{j, j \neq i, ||X_i - X_j|| < r\},$$
$$\mathcal{N}_0 = \{j, ||X_j|| < R - r\}.$$

To avoid taking points from the boundary we restricted ourselves to the set $\mathcal{N}_0$. The set $\mathcal{N}_i$ contains indexes of points neighbouring point X_i.

From the explanation above it seems plausible that $\widehat{\Sigma}$ is a good estimator. Below we will prove, that, on average, our estimator is very close to the scattering matrix in Frobenius norm.

Theorem 2.5.2 *[4, Theorem 3] For $d < \infty$, let $X_1, ..., X_N$ be the d-dimensional data points, with $\mathbb{E}(N) = n$, and R, r be the parameters determining the empirical estimator $\widehat{\Sigma}$ from the Equation 2.5.1. Assume $r = c_r / c^{d/(2d+4)} \sqrt{d \log n} \geq \max(\sqrt{d}, \sqrt{5Tr(\Sigma)/2})$ and that every eigenvalue of matrix Σ is bounded by $2c_r^2 c^{2d/(2d+4)}/3$ for constants $c = 16\pi^2 e^2$ and some $c_r > 0$. Then we have the following inequality:*

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||_F \leq \frac{d^2 (c_r \sqrt{\log n})^{d+1}}{\sqrt{n}} = t_{n,d}$$

Although the difference grows exponentially with the number of dimensions, our estimation is good - $\log n$ grows very slowly, so the rate coincides with the one in independent setting. We will prove this estimation by finding the optimal bias-variance tradeoff.

Proof: Since from the Jensen's inequality and the triange inequality we have the following:

$$\left(\mathbb{E}||\widehat{\Sigma} - \Sigma||_F\right)^2 \leq \mathbb{E}||\widehat{\Sigma} - \Sigma||_F^2 \leq ||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2 + \mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2,$$

bounding the bias and variance will play a big role in the proof.

Let us take a look at $||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2$ first. We can express it in a different way:

$$||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2 \leq d||\mathbb{E}\widehat{\Sigma} - \Sigma||_{op}^2 = d \sup_{||v||=1} [v^T \Sigma v - v^T \mathbb{E}\widehat{\Sigma} v]^2,$$

where d denotes the dimension. It follows from the definition of $\widehat{\Sigma}$, that

$$v^T\widehat{\Sigma}v = \frac{|B(1)|r^{d+2}}{d+2}||v||^2 - \frac{1}{|B(R-r)|}\sum_{i\in\mathcal{N}_0}\sum_{j\in\mathcal{N}_i}\langle X_i - X_j, v\rangle^2 .$$

From the definition of the intensity function we have the following:

$$\mathbb{E}\left(\sum_{i\neq j}\psi\left(X_i, X_j\right)\right) = \int\int_{B(R)\times B(R)}\psi(x,y)g_2(x,y)dxdy,$$

so for the right test function ψ:

$$\psi(x,y) = -\frac{1}{|B(R-r)|}\langle x-y, v\rangle^2 \mathbb{1}\{||x-y||\leq r, ||x||\leq R-r\},$$

the expression $\mathbb{E}\left(v^T\widehat{\Sigma}v\right)$ can be expressed as:

$$\mathbb{E}\left(v^T\widehat{\Sigma}v\right) = \frac{|B(1)|r^{d+2}}{d+2}||v||^2 - \sum_{i\neq j}\mathbb{E}\left(\psi\left(X_i, X_j\right)\right) = \frac{|B(1)|r^{d+2}}{d+2} -$$

$$-\frac{1}{B(R-r)}\int\int_{||x-y||\leq r, ||x||\leq R-r}\langle x-y, v\rangle^2\left(1-\exp(-(x-y)^T\Sigma^{-1}(x-y))\right)dydx.$$

Next, to calculate the integral we will insert $u = x - y$. We get the following expression:

$$\frac{|B(1)|r^{d+2}}{d+2} - \frac{1}{B(R-r)}\int_{||x||\leq R-r}\left(\int_{||u||\leq r}\langle u, v\rangle^2 du - \int_{||u||\leq r}\langle u, v\rangle^2\exp(-u^T\Sigma^{-1}u)du\right)dx.$$

Both inner integrals are the expected values of uniform and normal distribution respectively, so for $U\sim\mathcal{U}(B(1))$ and $Z\sim\mathcal{N}(0,\Sigma)$ we can rewrite the expression in the following way:

$$\frac{|B(1)|r^{d+2}}{d+2} - \frac{1}{B(R-r)}\int_{||x||\leq R-r}\left(r^2|B(r)|\mathbb{E}[\langle U, v\rangle^2] - \mathbb{E}[\langle Z, v\rangle^2\mathbb{1}\{||Z||\leq r\}]\right)dx =$$

$$= \frac{|B(1)|r^{d+2}}{d+2} - \frac{B(R-r)}{B(R-r)}\left(\frac{|B(r)|r^2}{d+2} - \mathbb{E}[\langle Z, v\rangle^2\mathbb{1}\{||Z||\leq r\}]\right).$$

The last line is the consequence of the fact that $\mathbb{E}[UU^T]$ is a scaled identity matrix.

This way, we have managed to simplify the formula for $\mathbb{E}[v^T\widehat{\Sigma}v]$:

$$\mathbb{E}\left[v^T\widehat{\Sigma}v\right] = \mathbb{E}\left[\langle Z, v\rangle^2\mathbb{1}\{||Z||\leq r\}\right]\xrightarrow{r\to\infty}\mathbb{E}\left[\langle Z, v\rangle^2\right] = v^T\Sigma v.$$

Since the only element depending on r is the indicator function, we can conclude that $\mathbb{E}[v^T\widehat{\Sigma}v] \leq v^T\Sigma v$. In order to obtain the bound on the aforedescribed expression:

$$\sup_{||v||=1}\left[v^T\Sigma v - v^T\mathbb{E}\widehat{\Sigma}v\right]^2 = \sup_{||v||=1}\left(\mathbb{E}\left[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}\right] - \mathbb{E}\left[\langle Z,v\rangle^2\right]\right)^2,$$

we will look at the convergence rate. Applying Cauchy-Schwarz and Fubini theorems will prove fruitful:

$$\mathbb{E}\left[\langle Z,v\rangle^2\right] - \mathbb{E}\left[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}\right] \leq \mathbb{E}\left[||Z||^2\mathbb{1}\{||Z||>r\}\right] = \int_{r^2}^{\infty}\mathbb{P}\left(||Z||^2>t\right)dt.$$

Next, we will utilize the Lemma 1.1.19 to imply, that $||Z||^2$ is equal in distribution to $\sum_{j=1}^k \lambda_j G_j^2$, for i.i.d. univariate standard normal random variables G_j and the eigenvalues of the scattering matrix Σ denoted by $\lambda_1,...,\lambda_k$. We will apply the form $\sum_{j=1}^k \lambda_j G_j^2$ in the Lemma 1.1.20, to obtain the following inequality:

$$\mathbb{P}\left(\sum_{j=1}^k \lambda_j G_j^2 - \sum_{j=1}^k \lambda_j \geq 2||\lambda||_2 x + 2||\lambda||_\infty x^2\right) \leq \exp(-x^2),$$

for λ denoting the vector of $\lambda_1,...,\lambda_k$. We utilized x^2 instead of x, as x^2 is non-negative. Furthermore, since λ_j are eigenvalues of the covariance matrix Σ, we can write $Tr(\Sigma)$ instead of $\sum_{j=1}^k \lambda_j$:

$$\mathbb{P}\left(\sum_{j=1}^k \lambda_j G_j^2 \geq Tr(\Sigma) + 2||\lambda||x + 2\lambda_1 x^2\right) \leq \exp(-x^2),$$

for $\lambda_1 > 0$ denoting the highest eigenvalue. Instead of $||\cdot||_2$, we will write $||\cdot||$. Since $\sum_{j=1}^k \lambda_j G_j^2$ has the same distribution as $||Z||^2$, we have the following:

$$\mathbb{P}\left(||Z||^2 \geq Tr(\Sigma) + 2||\lambda||x + 2\lambda_1 x^2\right) \leq \exp(-x^2).$$

To obtain the expression that we want, we need to find the bound t, such that

$$t = Tr(\Sigma) + 2||\lambda||x + 2\lambda_1 x^2.$$

We can solve this equation using Δ formula for quadratic equations (x is a variable).

$$\Delta = 4||\lambda||^2 - 8\lambda_1(Tr(\Sigma)-t) \geq 0 \iff t \geq Tr(\Sigma) - \frac{||\lambda||^2}{2\lambda_1}$$

Which gives us the restriction for x:

$$x = \frac{-2||\lambda|| \pm \sqrt{4||\lambda||^2 - 8\lambda_1\left(Tr(\Sigma)-t\right)}}{4\lambda_1}.$$

Since x needs to be non-negative, the nominator needs to be non-negative and, hence, the square root has to be of positive sign and $Tr(\Sigma) \leq t$ (otherwise the root is smaller than $2||\lambda||$, so $x < 0$). While this gives us an expression we can put inside the integral, we can simplify it even further, as long as $t \geq \frac{5Tr(\Sigma)}{2}$. Then, using the fact that $||\lambda|| \leq \lambda_1 Tr(\Sigma)$, as well as the inequality of arithmetic and geometric means, will bring us the following:

$$t = Tr(\Sigma) + 2||\lambda||x + 2\lambda_1 x^2 \leq Tr(\Sigma) + (Tr(\Sigma) + \lambda_1 x^2) + 2\lambda_1 x^2 \Rightarrow x^2 \geq \frac{t - 2Tr(\Sigma)}{3\lambda_1}.$$

This gives us the following inequality:

$$\int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt \leq \int_{r^2}^{\infty} \exp\left(\frac{2Tr(\Sigma) - t}{3\lambda_1}\right) dt.$$

Exponential of this expression can be easily integrated:

$$\int_{r^2}^{\infty} \exp\left(\frac{2Tr(\Sigma) - t}{3\lambda_1}\right) dt \leq 3\lambda_1 \exp\left(\frac{2Tr(\Sigma) - r^2}{3\lambda_1}\right),$$

bringing us the desired bound (as long as r^2 fulfills the same restriction as t):

$$||\mathbb{E}\widehat{\Sigma} - \Sigma||^2 \leq 9d\lambda_1^2 \exp\left(\frac{4Tr(\Sigma) - 2r^2}{3\lambda_1}\right).$$

In the above expression we can swap λ_1 with norm of the operator, as $\lambda_1 \leq ||\widehat{\Sigma}||_{op}$.

Now we will bound the variance $\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||^2$. We can write the following:

$$\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||^2 = \sum_{i,j \leq d} Var\left(e_i^T \widehat{\Sigma} e_j\right),$$

for vectors e_i from canonical basis of $\mathbb{R}^d$. Analogously to the first part of this proof, for normalized vectors u, v, we utilize the definition of $\widehat{\Sigma}$, to get the following:

$$Var\left(u^T \widehat{\Sigma} v\right) = Var\left(\frac{|B(1)|r^{d+2}}{d+2} \langle u, v \rangle\right) +$$

$$+ Var\left(\frac{1}{|B(R-r)|} \sum_{i \in \mathcal{N}_0} \sum_{j \in \mathcal{N}_i} \langle X_i - X_j, u \rangle \langle X_i - X_j, v \rangle\right).$$

As the first variance contains non-probabilistic expression, we can ommit it:

$$Var\left(u^T \widehat{\Sigma} v\right) = Var\left(\frac{1}{|B(R-r)|} \sum_{i \in \mathcal{N}_0} \sum_{j \in \mathcal{N}_i} \langle X_i - X_j, u \rangle \langle X_i - X_j, v \rangle\right) =$$

$$= \frac{1}{|B(R-r)|^2} Var\left(\sum_{i\neq j} \langle X_i - X_j, u\rangle \langle X_i - X_j, v\rangle \mathbb{1}\{||X_i - X_j|| \leq r, ||X_i|| \leq R - r\}\right).$$

For the right function ψ:

$$\psi(x,y) = \langle x-y, v\rangle \langle x-y, u\rangle \mathbb{1}\{||x-y|| \leq r, ||x|| \leq R-r\},$$

the variance becomes the following expression:

$$\frac{1}{|B(R-r)|^2} Var\left(\sum_{i\neq j} \psi(X_i, X_j)\right),$$

resulting in the following inequality:

$$Var\left(u^T \widehat{\Sigma} v\right) \leq \frac{1}{|B(R-r)|^2} Var\left(\sum_{i\neq j} \psi(X_i, X_j)\right).$$

As before we will apply the joint intensity function to the expected value and use its symmetric property:

Lemma 2.5.3 *For the aforementioned variables the following holds:*

$$Var\left(\sum_{i\neq j} \psi(X_i, X_j)\right) \leq 4\int\int\int \psi(x,y)\psi(x,z) g_3(x,y,z)dxdydz + \\ + \int\int \psi(x,y)^2 g_2(x,y)dxdy.$$

Proof: From the definition of the variance, we have the following:

$$Var\left(\sum_{i\neq j} \psi(X_i, X_j)\right) = \sum_{i\neq j, k\neq l} \mathbb{E}(\psi(X_i, X_j)\psi(X_k, X_l)) - \left(\mathbb{E}\left(\sum_{i\neq j} \psi(X_i, X_j)\right)\right)^2 =$$

$$= \int\int\int\int \psi(x,y)\psi(z,w) g_4(x,y,z,w)dxdydzdw + 4\int\int\int \psi(x,y)\psi(x,z) g_3(x,y,z)dxdydz + \\ + \int\int \psi(x,y)^2 g_2(x,y)dxdy - \left(\int\int \psi(x,y) g_2(x,y)dxdy\right)^2.$$

We can utilize the Lemma 1.1.18 to get the inequality:

$$g_4(x,y,z,w) \leq g_2(x,y) g_2(z,w),$$

to obtain the following:

$$\int\int\int\int \psi(x,y)\psi(z,w) g_4(x,y,z,w)dxdydzdw - \left(\int\int \psi(x,y) g_2(x,y)dxdy\right)^2 \leq 0,$$

which brings us the following bound:

$$Var\left(\sum_{i\neq j}\psi(X_i,X_j)\right) \leq 4\int\int\int \psi(x,y)\psi(x,z)g_3(x,y,z)dxdydz+$$

$$+\int\int \psi(x,y)^2 g_2(x,y)dxdy.$$

QED

At this point, we will eliminate the integrals. Since $||u|| = 1$ and $||x-y|| \leq r$, we obtain the following from the Cauchy-Schwarz inequality:

$$\langle x-y,u\rangle \leq ||x-y||\cdot||u|| \leq r.$$

This means, that for the test function ψ we have the following inequality:

$$\psi(x,y) \leq r^2 \mathbb{1}\{||x-y|| \leq r, ||x|| \leq R-r\}.$$

Additionally, it follows from the Fischer's inequality (Lemma 1.1.18) and normalization, that:

$$g_k(x_1,...,x_k) \leq \prod_{j=1}^{k} g_1(x_j,x_j) = 1,$$

so we have the following:

$$4\int\int\int \psi(x,y)\psi(x,z)g_3(x,y,z)dxdydz + \int\int \psi(x,y)^2 g_2(x,y)dxdy \leq$$

$$\leq 4\int\int\int r^4\mathbb{1}\{||x-y|| \leq r, ||x-z|| \leq r, ||x|| \leq R-r\}dxdydz+$$

$$+\int\int r^4\mathbb{1}\{||x-y|| \leq r, ||x|| \leq R-r\}dxdy =$$

$$= r^4\left(|B(r)|^2|B(R-r)| + |B(r)||B(R-r)|\right) \leq 5r^{2d+4}(R-r)^d|B(1)|^3.$$

The last inequality holds if we take $r \geq \sqrt{d} \geq |B(1)|^{-1/d}$. In conclusion we have bounded the variance by the following expression:

$$Var\left(u^T\widehat{\Sigma}v\right) \leq \frac{5r^{2d+4}(R-r)^d|B(1)|^3}{|B(R-r)|^2} \leq \frac{5r^{2d+4}(R-r)^d|B(1)|^3}{|B(1)|^2(R-r)^{2d}} = \frac{5r^{2d+4}|B(1)|}{(R-r)^d}.$$

Furthermore, since the magnitude of the parameter r should be smaller than that of the parameter R, we can assume that $r \leq R/2$, resulting in:

$$\frac{5r^{2d+4}|B(1)|}{(R-r)^d} \leq \frac{5r^{2d+4}2^d|B(1)|^2}{|B(1)|R^d} = \frac{5r^{2d+4}2^d|B(1)|^2}{n},$$

where n denotes the expected number of points in the ball $B(R)$. Using the Stirling's formula we can remove the ball $|B(1)|$ from the equation, since $|B(1)|^{1/d} \leq 2\pi e/\sqrt{d}$:

$$\frac{5r^{2d+4}2^d|B(1)|^2}{n} \leq \frac{5r^{2d+4}2^d(2\pi e)^{2d}}{nd^d}.$$

Obtained this way bound:

$$Var\left(u^T\widehat{\Sigma}v\right) \leq \frac{5r^{2d+4}2^d(2\pi e)^{2d}}{nd^d},$$

brings us the final restriction on $\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||^2$:

$$\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||^2 \leq \sum_{i,j\leq d} \frac{5r^{2d+4}(8\pi^2e^2)^d}{nd^d} = \frac{d^2 5r^{2d+4}(8\pi^2e^2)^d}{nd^d} \leq \frac{c^d r^{2d+4}}{d^{d-2}n},$$

for $c = 16\pi^2e^2$ (since $2^d > 5$ for $d \geq 3$).

To finish the proof we will combine those expressions. For $r \geq \sqrt{d}$, $r \geq \sqrt{5Tr(\Sigma)/2}$, $r \leq R/2$ we have the following:

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||^2 \leq 9d\lambda_1^2 \exp\left(\frac{4Tr(\Sigma) - 2r^2}{3\lambda_1}\right) + \frac{c^d r^{2d+4}}{d^{d-2}n}.$$

We can simplify this expression. Since $r \geq \sqrt{5Tr(\Sigma)/2}$, $4Tr(\Sigma) - 2r^2 \leq 0$, and therefore, the exponential is no greater than 1. In the second expression, choosing the right value of r will be of great importance. When $r = c_r/c^{d/(2d+4)}\sqrt{d\log n}$, the whole inequality becomes the following:

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||^2 \leq \frac{9d\lambda_1^2 \exp\left(4Tr(\Sigma)/3\lambda_1\right)}{n^{2dc_r^2c^{-2d/(2d+4)}/(3\lambda_1)}} + \frac{c_r^{2d+4}d^4\log^{d+2} n}{n}.$$

For $0 \leq \lambda_1 \leq 2dc_r^2c^{-2d/(2d+4)}/3$ we can omit the first expression.

QED

In order to complete the proof we required certain Gaussian properties. One may wonder what if we do not have them, how well can we bound this expression if we are not so restrained. We will analyze a more general case in another chapter.

2.6 Spike model - detection and estimation

The simplest model involving spike analysis handles the case, where the scattering matrix is of the form:

$$\Sigma = I + \lambda uu^T, ||u|| = 1, \lambda \geq 0.$$

The direction of the spike is based on the vector u and its strength - on the value of λ.

For the GDP case, slight modification of this equation will suffice. We can assume that

the spike does not have an impact on the mean, otherwise, we can detect the spike by analyzing the mean. This gives us the following alternation:

$$2\pi\Sigma = (1+\lambda)^{-\frac{1}{d-1}}\left(I - uu^T\right) + (1+\lambda)uu^T, ||u|| = 1.$$

We can detect the existence of the spike, by analyzing the parameter λ. If this parameter is too small to detect the difference, we obtain the case of no spike (i.e. $\lambda = 0$) and:

$$2\pi\Sigma = I.$$

Equivalently we can utilize the operator norm for the detection case. If there is no spike, we have the following:

$$\frac{||I||}{2\pi} = \frac{1}{2\pi} = ||\Sigma||_{op}.$$

In other case, this is the result:

$$\frac{||\,(1+\lambda)^{-\frac{1}{d-1}}\left(I - uu^T\right) + (1+\lambda)uu^T||}{2\pi} = \frac{1+\lambda}{2\pi} = ||\Sigma||_{op}.$$

If we detect λ, we can estimate vector u, by utilizing the data from the dataset.

In the last section, we have bounded the average difference between the empirical scattering matrix and the scattering matrix in the Frobenius norm, so now we can proceed to the detection and estimation of the spike in the empirical matrix.

Theorem 2.6.1 *Let ψ_t be the test function for detecting a spike of strength λ in the empirical matrix $\widehat{\Sigma}$ from the equation 2.5.1, based on the model with scattering matrix Σ. The test function ψ_t has the following form:*

$$\psi_t = \mathbb{1}\{2\pi||\widehat{\Sigma}||_{op} > 1 + t\mathfrak{t}_{n,d}\},$$

for parameter t. Assume that:

$$2\pi||\widehat{\Sigma} - \Sigma||_{op} \leq ||\widehat{\Sigma} - \Sigma||_F.$$

Then as long as:

$$t = \frac{1}{\delta} \text{ and } \lambda > \frac{\mathfrak{t}_{n,d}}{\pi},$$

for parameter $\delta > 0$, we can detect a spike with accuracy $1-\delta$. In other words, let ϕ_{spike} be a Boolean variable equal to 1 if there exists a spike in the matrix Σ and 0 otherwise. We have the following:

$$\mathbb{P}\left(\psi = 1|\phi_{spike} = 0\right) \leq \delta, \quad \mathbb{P}\left(\psi = 0|\phi_{spike} = 1\right) \leq \delta.$$

Proof: First, we consider the case of no spike, i.e..

$$||\Sigma||_{op} = \frac{1}{2\pi}.$$

We have the following:

$$\mathbb{P}\left(\psi = 1 | \phi_{spike} = 0\right) = \mathbb{P}\left(2\pi||\widehat{\Sigma}||_{op} - 1 \geq t\mathfrak{t}_{n,d}\right).$$

We can apply the Markov's inequality, to obtain the following:

$$\mathbb{P}\left(2\pi||\widehat{\Sigma}||_{op} - 1 \geq t\mathfrak{t}_{n,d}\right) \leq \frac{2\pi\mathbb{E}\left[||\widehat{\Sigma}||_{op}\right] - 1}{t\mathfrak{t}_{n,d}} \leq \frac{2\pi\mathbb{E}\left[||\widehat{\Sigma} - \Sigma||_{op} + ||\Sigma||_{op}\right] - 1}{t\mathfrak{t}_{n,d}}.$$

The last inequality is the result of the triangular inequality. Next, utilizing the norm assumption and the fact, that $2\pi||\Sigma||_{op} = 1$, we can alter the expression above:

$$\frac{2\pi\mathbb{E}\left[||\widehat{\Sigma} - \Sigma||_{op} + ||\Sigma||_{op}\right] - 1}{t\mathfrak{t}_{n,d}} \leq \frac{\mathbb{E}\left[||\widehat{\Sigma} - \Sigma||_{F}\right]}{t\mathfrak{t}_{n,d}}.$$

By the Theorem 2.5.2:

$$\frac{\mathbb{E}\left[||\widehat{\Sigma} - \Sigma||_{F}\right]}{t\mathfrak{t}_{n,d}} \leq \frac{\mathfrak{t}_{n,d}}{t\mathfrak{t}_{n,d}} = \frac{1}{t},$$

which brings us the final inequality:

$$\mathbb{P}\left(\psi = 1 | \phi_{spike} = 0\right) \leq \frac{1}{t}.$$

As long as $1/t \leq \delta$, we obtain the desired bound. If the spike exists, we can show the bound of $\mathbb{P}\left(\psi = 0 | \phi_{spike} = 1\right)$ analogously. QED

If the spike exists, we can find vector u, from the formula below:

$$2\pi\Sigma = (1+\lambda)^{-1/(d-1)}\left(I - uu^T\right) + (1+\lambda)uu^T, ||u|| = 1,$$

using matrix perturbation analysis. It is important to note, that the vector u is the eigenvector of this matrix, with the corresponding eigenvalue λ. It follows from the calculations:

$$((1+\lambda)^{-1/(d-1)}\left(I - uu^T\right) + (1+\lambda)uu^T)u = (1+\lambda)^{-1/(d-1)}(u-u) + (1+\lambda)u = (1+\lambda)u.$$

We can approximate the vector $\hat{u}$ using the following theorem.

Theorem 2.6.2 *Let $\Sigma, \widehat{\Sigma}$ be symmetric matrices with eigenvalues $\lambda_1 \geq ... \geq \lambda_d$ and $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_d$ respectively. Let $r, s \in [1, d]$ be integers and $V, \hat{V}$ be $d \times (s - r + 1)$ matrices with orthonormal columns corresponding to these eigenvalues. Finally let $H, \hat{H}$ be subspaces spanned by $V, \hat{V}$. Then, as long as $\delta > 0$, we have the following:*

$$||\sin\angle(H, \hat{H})|| \leq \frac{||\widehat{\Sigma} - \Sigma||_F}{\delta},$$

where δ denotes the eigengap ($\hat{\lambda}_0 = -\infty, \hat{\lambda}_{d+1} = \infty$):

$$\delta = \inf\{|\lambda - \hat{\lambda}|; \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s+1}] \cup [\hat{\lambda}_{r-1}, \infty)\}.$$

The proof of this theorem can be found in the article of C. Davis and W. M. Kahan [3]. We can see why this theorem is relevant here - using this formula we can estimate the angle difference of vectors $u, \hat{u}$, by inserting known variables into the correct places. We know, that $u, \hat{u}$ are eigenvectors associated with largest eigenvalues, also we can estimate eigengap as:

$$1 + \lambda - \frac{1}{1+\lambda}^{1/(d+1)} \geq 1 + \lambda - 1 = \lambda,$$

which brings us the following:

$$\mathbb{E}|\sin(\angle(u, \hat{u}))| \leq \frac{\mathbb{E}||\widehat{\Sigma} - \Sigma||_F}{\lambda} \leq \frac{\mathfrak{t}_{n,d}}{\lambda}.$$

The last inequality follows from the Theorem 2.5.2.

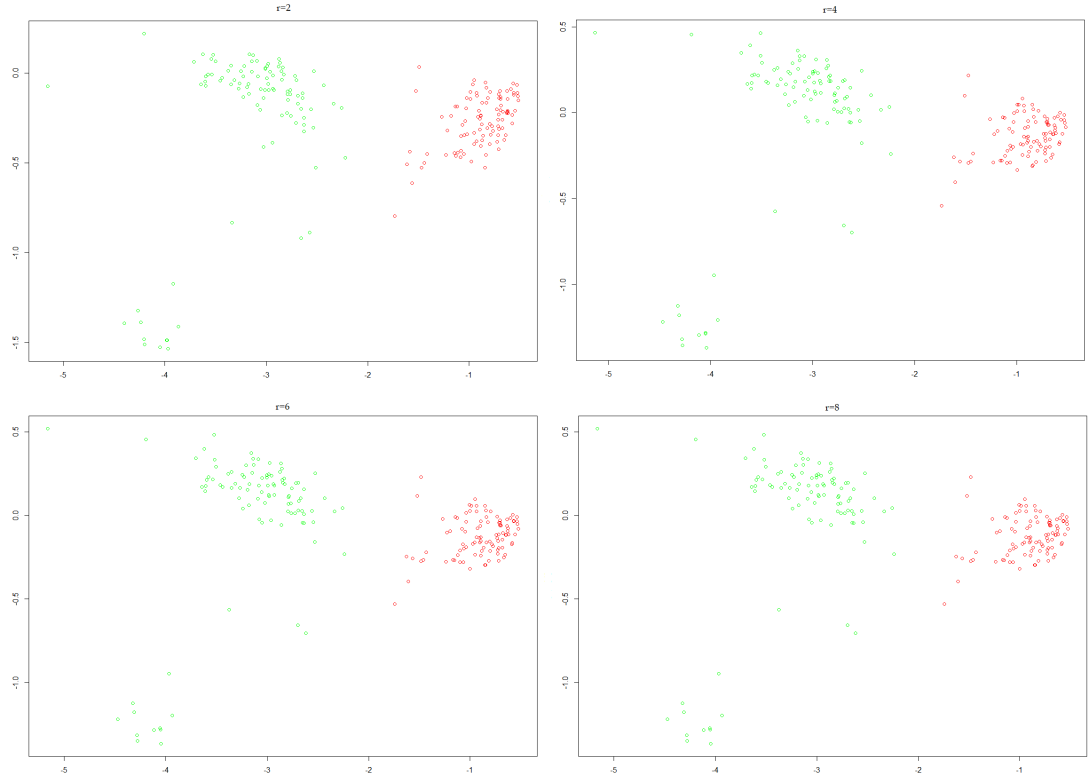
3 Examples and statistical estimation in a more general case

3.1 Selection of the parameters

In the previous chapter we have introduced the method of finding empirical estimator of the scattering matrix of Gaussian determinantal point process. In the Theorem 2.5.2 we have bounded the expected distance of the empirical estimator from the true estimator in the Frobenius norm, so the next logical step is to apply the GDP method to chosen datasets, in order to reduce the number of dimensions. We will compare this method with other aforedescribed methods (projection, simple factor analysis, principal component analysis and Gaussian determinantal process method) on the selected samples.

First, it is necessary to discuss the parameters R and r in GDP model. Since every data point in the sample needs to be inside the ball of radius R centered at 0, it is natural to assume that R needs to be bigger than the norm of the biggest sample. We could verify that an adequate number of data points is distributed at a similar distance from 0, so that the parameter R in the model is not too big, however the choice of R will not influence the outcome in the case of dimension reduction. On the other hand picking too large values of R will not have any positive influence, but may cause some unnecessary calculations, since in the formula for empirical estimator we have the ball of radius $R-r$.

In most cases the selection of the value of r barely have any effect on the outcome. Indeed, as long as r is big enough that for most i, $\mathcal{N}_i$ contains enough points, increasing the value will not have that much of a difference, for the data sharply concetrated on one area. If r is too small, we will not be able to observe enough correlation (even none), making the outcome unreliable. While selecting larger values of r may result in bigger variance, it also provides almost all sets $\mathcal{N}_i$ with majority of points. For this reason in the examples below we will choose big values of parameter r, unless specified otherwise. Below we have shown how a choice of parameter r affects the selected sample.



Plot of cost of living using different values of r for dimension reduction. The distance between points can exceed 5, however in most cases it stays within 4.

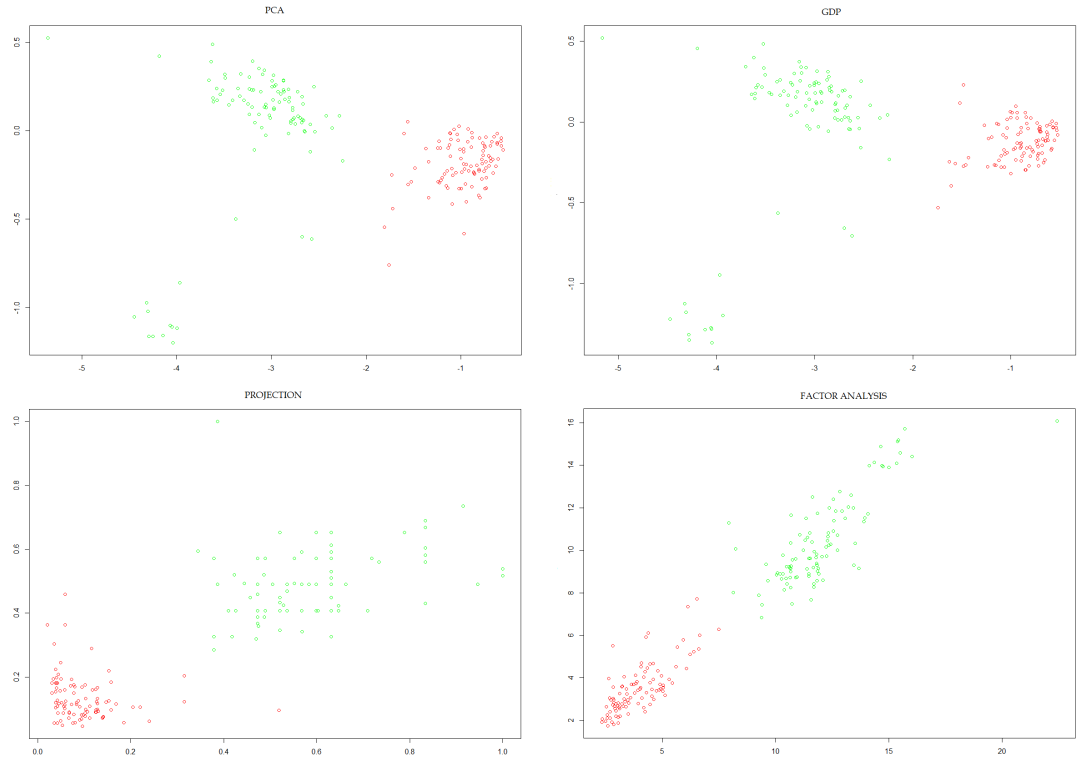
3.2 Examples

The article [4] mentions two examples, that were selected to present the superiority of the GDP method in clustering the similar data points, compared to the PCA method. Nevertheless, the analysis of these examples was mostly ommited and the number of the examples is too small to draw comprehensive conclusions. This section will provide an in-depth analysis of the GDP method.

3.2.1 Example 1A

Data description:
The first example deals with cost of living in different cities. This dataset, provided by the website *numbeo.com* [7], gathers cost of various (54) products (such as bottled water, domestic/imported beer, milk, bread, rice, eggs, apples), meals, services (such as local transport, internet, renting a car, flat) and average wages across 4873 cities in the world. One would expect correlation between wages and prices, for example one can notice that inexpensive meal in India costs less than in USA, with latter having higher wages. After removing rows with empty cells and rescalling each column to have values

from 0 to 1, we will implement four aforedescribed methods on smaller sample of 200 cities - 100 with highest wages and 100 with lowest. Then, by colouring the plots with 2 colours - red for poorer cities and green for richer, we will get the results.



but while GDP and PCA methods were explained in-depth before, we need to specify the other methods.

Methods:
We have utilized four aforementioned methods: projection, simple factor analysis, principal component analysis and Gaussian determinantal process method.

Projection:
In the projection method, we used 2 factors - price of a meal in an inexpensive restaurant and price of a domestic beer. While prices of many products or services can be influenced by culture or import (e.g., the average Japanese consumes more rice than average Polish), in most countries one can find inexpensive restaurants. Also, most cultures produce some sort of alcoholic beverage, since beer is one of the oldest and one of the most widely consumed type of alcoholic drink in the world, we have decided to include these factors.

Simple factor analysis:
While other methods use factor analysis, we have decided to implement a simple factor

analysis method, that is influenced by all the factors - the first coordinate of each data point is the sum of the values of its first 27 factors and the second coordinate is the sum of the values of its remaining factors, resulting in two-dimensional points.

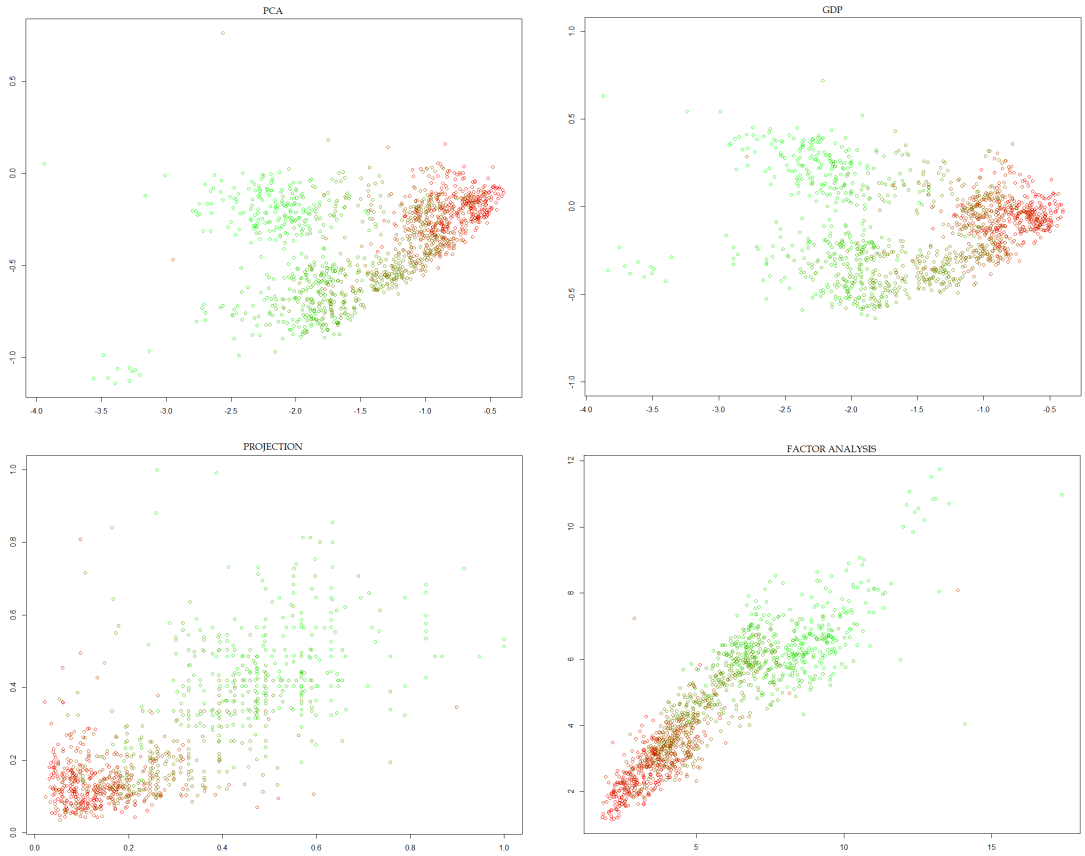
PCA and GDP methods:
In the PCA and GDP methods we have followed the aforementioned algorithms.

Results:
In the projection method we can see clear distinction between cities with higher and lower wages on the plot. While red dots are densely concentrated around origin, the green dots take bigger values. In the simple factor analysis method one can notice a repulsion between green and red dots, but at the same time it seems that green points are more concentrated than in the projection method and we can observe more directionality. The GDP and PCA methods seems to show similar results. Red points are sharply concentrated in one place, while green dots have two concentration points, with one gathering smaller number of points. Those are cities with highest average wages. While both colours are divided even better than in the last two methods, in the GDP plot red points seem too be more concentrated than in the PCA plot. While one can have the same observation regarding green points, albeit of much lesser magnitude, the difference is too small to have any certain statements.

Conclusion:
While the plots look different, in a way, every graphic presents the same thing - cities with higher average wage, tend to have higher cost of living. All methods managed to show that, but the PCA and GDP methods are better at clustering the same type of data points. In this example, the GDP method is superior at concentrating points of the same category.

3.2.2 Example 1B

Data description:
We utilize the dataset from the previous example. While we concluded, that it shows slight superiority of GDP over PCA method, we have only compared cities with highest and lowest wages. We will take a look at all 1264 samples and colour them with the remaining colours between red and green.



Methods:
We utilize the exact same methods as in the Example 1*A*.

Results:
While all the plots preserved their forms, the spaces between clusters got smaller. Projection and factor analysis plots remained essentially unchanged (the wages get higher as we move towards top right corner), while some new observations can be made about the other two plots. We have lost the sharp concentration of red points, but we have gained 2 new clusters. Even though PCA method shows better concentration of points (albeit one needs to spend some time looking at both plots to observe that), GDP method shows better spacing between clusters, especially the bottom two. We can even notice, that the red cluster seem to be dividing into few smaller ones in the GDP method - each of slightly different colour.

Conclusion:
We arrive at the similar conclusion as in the Example 1*A* - every graphic presents, that the cities with higher average wage, tend to have higher cost of living. All methods managed to show that, but the PCA and GDP methods are better at clustering the same type of data points. Although in this example we can see very slight improvement

of the concentration of points in the PCA method with comparison to the GDP method, the GDP method shows better repulsion between different clusters.

3.2.3 Example 2A

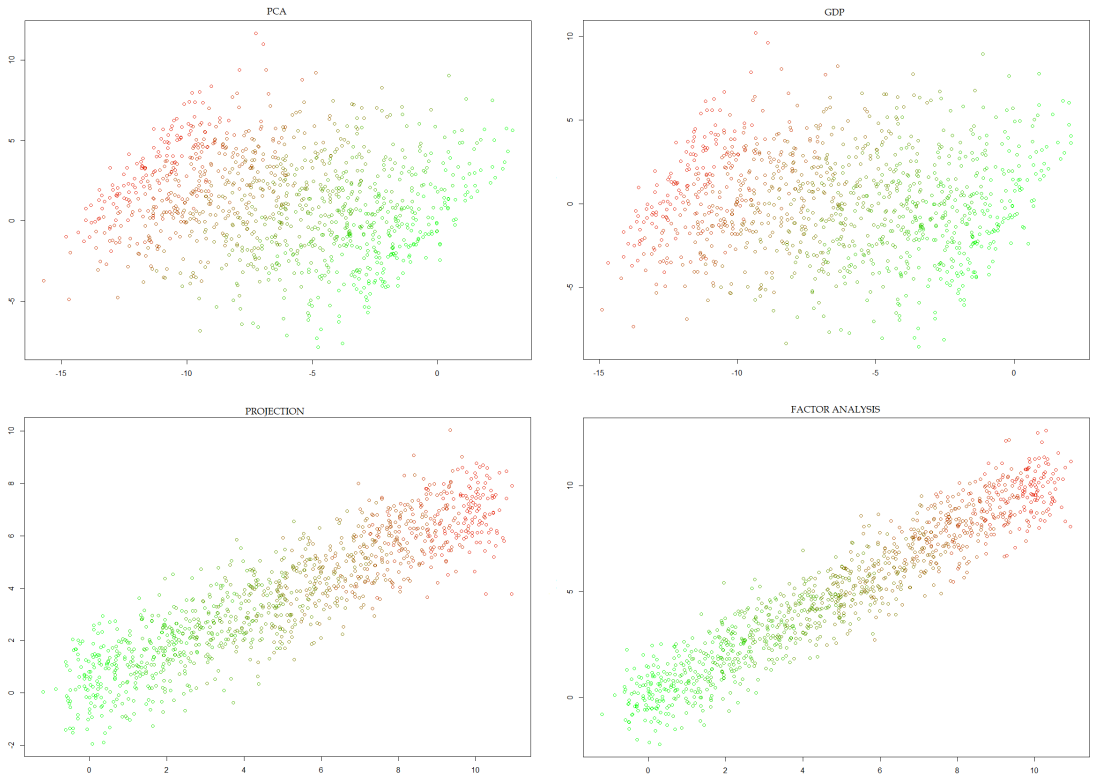
While previous dataset enables us to try various factor analysis methods, in order to represent it we need to lose over 50 dimensions, so a lot of information is lost. In addition the analysis is more tedious - to see the correlation between points we would have to calculate at least second intensity g_2, which is the determinant of the matrix of size 54×54 and then integrate it. To check the influence of factors we would have to check eigenvectors of the same size. Because of that the next dataset will have only 3 factors.

Taking into consideration that a set of samples from the Gaussian random variables, resembles a dataset, we can generate samples instead of utilizing existing data. We need to generate them from more than one random variable, since we want to see correlation between samples with different outcomes (for example in case of trying to predict cancer, positive results will have different values than negative ones, resulting in some repulsion). This approach allows us to tune the parameters.

Data description:
The second dataset contains 1100 samples with 3 factors x, y, z. Each value was generated using random variables X, Y, $(X + Y + Z)/3$ respectively, where:

$$X \sim \mathcal{N}(k, 0.4), \quad Y \sim \mathcal{N}(1, 3), \quad Z \sim \mathcal{N}(k, 1).$$

The integer k starts at 0 and changes by 1 every 100 samples, in order to create different clusters. The values of y do not differ between clusters. We will observe what is the influence of such noise in our dimension reduction methods.



Methods:
We have utilized four aforementioned methods: projection, simple factor analysis, principal component analysis and Gaussian determinantal process method.

Projection and Simple factor analysis:
In the projection and (simple) factor analysis we tried not to rely too much on factor y, by taking the other factors in projection and weighted arithmetic mean in the latter.

PCA and GDP:
In the PCA and GDP methods we have followed the aforementioned algorithms.

Results:
In this example projection and (simple) factor analysis are highly effective - since the expected value changes every 100 samples, we can clearly see clustering around those values. In the PCA and GDP methods we can certainly see the clustering as well, however the spaces between each 100 samples (represented as different colours) are wider. Both approaches resemble each other, while one may find the GDP method as the one with better clustering, the pictures are too similar to make any conclusions. We will study closely the eigenvectors of the correlation matrix of the PCA method and the empirical estimator from the Equation 2.5.1 of the GDP method.

$$
\begin{array}{cc}
\multicolumn{2}{c}{PCA} \\
-0.6368981 & +0.4331657 \\
-0.3092159 & -0.90130550 \\
-0.7062197 & +0.0039868 \\
\multicolumn{2}{c}{GDP} \\
-0.7857839 & +0.3105233 \\
-0.1904525 & -0.9442916 \\
-0.5884485 & -0.1090351
\end{array}
$$

The first eigenvector is the most important, so it is the first one we will put into consideration. Both approaches value factor y less than the other factors. While the third coordinate of these vectors contains the factor y, in total this factor contributes to -0.5446225 in PCA method and -0.386602 in GDP method (we have obtained these numbers by taking the second coordinate and a third of the third coordinate as factor). Since the value is much smaller in GDP case, one may think that this method cares more about the values that differ, in order to make the samples of different distributions more clustered. Indeed, after generating the samples 100 times, the average contribution of factor y was 0.324 in PCA method and 0.228 in GDP method in magnitude to the sum of the eigenvectors. Generating samples with different parameters of X, Y, Z, while preserving factor y as a random noise, brings similar results (in each iteration parameter k was kept only in distributions of X and Z).

The second eigenvector values factor y more in case of GDP, however the difference is usually small and, as we have seen before, this vector is of much lesser importance. The other factors are more balanced in the first case, while the second case values more factor x.

Conclusion:
In conclusion, the GDP method is better than the PCA method in detecting an irrelevant factor.

3.2.4 Example 2B

Data description:
The next example will help us analyze the influence of variance on the first eigenvector. As before, the dataset contains 1100 samples with 3 factors x, y, z. Each value was generated using random variables X_1, Y_1, $(X_1 + Y_1 + Z_1)/3$ in the first case, X_2, Y_2, $(X_2 + Y_2 + Z_2)/3$, in the second case and X_3, Y_3, $(X_3 + Y_3 + Z_3)/3$ in the third case. Random variables have the following distributions:

$$X_1 \sim \mathcal{N}(4 + k, 0.1) \quad X_2 \sim \mathcal{N}(4 + k, 1) \quad X_3 \sim \mathcal{N}(4 + k, 10),$$

$$Y_1 \sim \mathcal{N}(1+k, 0.2) \quad Y_2 \sim \mathcal{N}(1+k, 2) \quad Y_3 \sim \mathcal{N}(1+k, 20),$$
$$Z_1 \sim \mathcal{N}(k, 0.05) \quad Z_2 \sim \mathcal{N}(k, 0.5) \quad Z_3 \sim \mathcal{N}(k, 5).$$

As before, the integer k starts at 0 and changes by 1 every 100 samples. The only difference between variables is the variance - in the second case it increased by 10 and in third - by 100, compared to the first dataset. After generating the dataset 100 times we got the following averages of the first eigenvectors (the number next to eigenvector responds to the index number of random variables):

1	PCA	GDP	2	PCA	GDP	3	PCA	GDP
x	0.5772832	0.5747514	x	0.5706597	0.5368106	x	0.3701357	0.4857652
y	0.5771879	0.5810792	y	0.5645063	0.6306291	y	0.6102790	0.7153781
z	0.5775796	0.5762004	z	0.5963878	0.5603491	z	0.6999850	0.3010035

Methods:
In this example we want to focus on the difference between the eigenvectors of the correlation matrix of the PCA method and the empirical estimator from the Equation 2.5.1 of the GDP method. As such we will consider only the PCA and the GDP methods.
PCA and GDP:
In the PCA and GDP methods we have followed the aforementioned algorithms.

Results:
In the first case eigenvectors have very similar values, but as variance grows, the GDP eigenvector seems to be more unbalanced - in second case the difference on second factor is around 0.07, while in third case it is over 0.1. Different growth rate on this coordinate is linked to the fact, that Y has the biggest variance among random variables. The difference in other coordinates grows as well and the GDP method seems to care little about the mixed factor. While it shows that the PCA method works better on more scattered data, the important observation is that the value of k is lost in the last method among the noise, so it is hard to find the directionality. If we increase the variance by 1000, both methods show similar eigenvectors again (the difference is less than 0.001 on each coordinate).

Conclusion:
The GDP method is more volatile to the changes of the variance than the PCA method. When the directionality is lost, the PCA method works better on the scattered data.

3.2.5 Example 2C

Data description:
In the last example we will show, how specific choice of parameter r may influence the GDP method. We have 1100 samples and the factors x, y, z are based on random variables X, Y, $(Y+Z)/2$, which have the following distributions:

$$X \sim \mathcal{N}(k, 1), \quad Y \sim \mathcal{N}(0.5, 1), \quad Z \sim \mathcal{N}(1, 1),$$

where k is as before, however for the first 100 samples we will use different distributions:

$$X \sim \mathcal{N}(0,5), \quad Y \sim \mathcal{N}(5,10), \quad Z \sim \mathcal{N}(5,10).$$

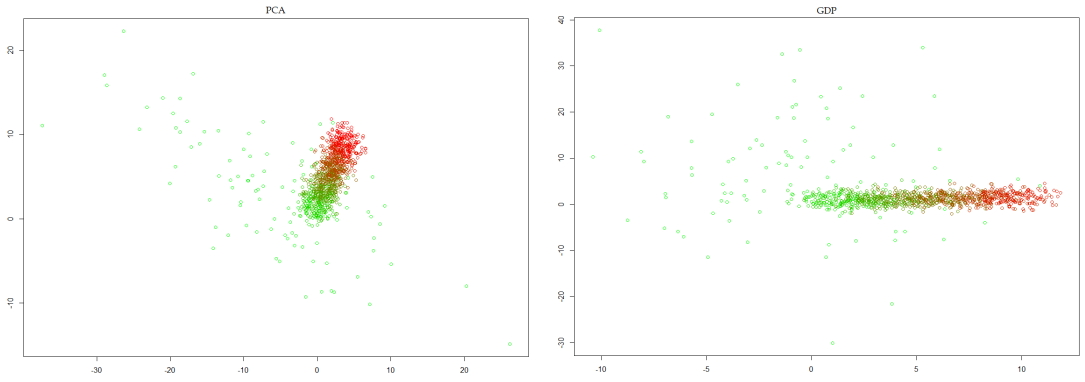
We can treat this 100 samples as noise or anomaly. In this scenario the values differ a lot between first eigenvectors of two methods:

PCA	GDP
0.5241979	0.99766483
0.6860292	0.03511829
0.4851875	0.04918215

Methods:
In this example we want to focus on the difference between the eigenvectors of the correlation matrix of the PCA method and the empirical estimator from the Equation 2.5.1 of the GDP method. As such we will consider only the PCA and the GDP methods.
PCA and GDP:
In the PCA and GDP methods we have followed the aforementioned algorithms.

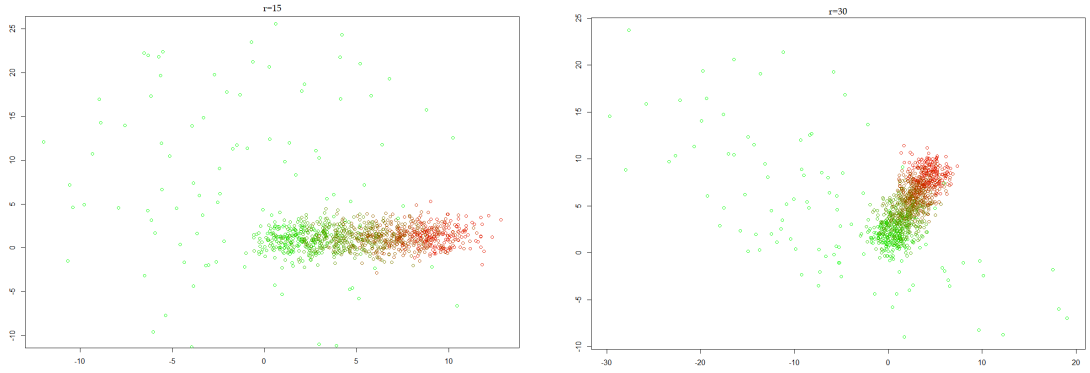
Results:
While PCA method shows almost balanced eigenvector (the second factor is larger, possibly thanks to the anomaly), the GDP method is highly unbalanced. Factor x is almost equal to 1, while the other factors are barely acknowledged. This is good - we wanted to detect directionality and in this case of a dataset with directionality at only one factor and some anomaly, the GDP method is highly efficient at detecting the direction, much better than the PCA method.

Let us take a look at the graphs of this example:



The colours, going from green to red, represent the samples - from 0 to 1100. The first 100 samples are chaotic, since they are the anomaly. We can clearly observe the difference between the graphs - in the second one the points create a figure that is more stretched and thinner than the figure in the first one. The plot of GDP method is better

at presenting the directionality in this example.
One may wonder why these two methods give us so different eigenvectors. The answer lies with the choice of r. Since the value of r controls which points are taken into consideration for each point, we can modify the algorithm to simply ommit the anomaly in order to get better results. Below we present the graphs of this example for $r = 15$ and $r = 30$. As we can see, for r big enough the GDP method gives us very similar results to those of the PCA method.



Conclusion:
The selection of the parameter r is important in the case of irrelevant noise in the dataset. With the right parameter r we can detect stronger directionality with superior accuracy in the GDP method, compared to the PCA method.

3.3 Statistical estimation in a more general case

In chapters 2 and 3 every determinantal point process we used was Gaussian. The natural follow-up question would be what if we take different DPP, how well the empirical parameter estimates the theoretical one then.

Problem 3.3.1 *What properties should a kernel $\mathbb{K}$ possess, in order to satisfy an inequality similar, or even the same, to the one described in Theorem* 2.5.2*?*

In the proof of Theorem 2.5.2 we have utilized Gaussian properties, to represent an integral as expected value of normal random variable with variance Σ, as well as control the decay of $\mathbb{P}\left(||Z||^2 > t\right)$ for increasing values of t and random variable Z. The former reason does not require the random variable to be Gaussian. We can control the decay of the tail by increasing the number of restrictions on the density of random variable Z, by adding tail bound to create artificial restriction. However, we will take a different direction and check how big the bound on the tail can be, so that the bound on $\mathbb{E}||\widehat{\Sigma}-\Sigma||_F^2$ stays the same.

Theorem 3.3.2 *Let function f be a symmetric function, i.e., $f(x) = f(-x)$, such that f^2 is a density function, with covariance matrix Σ, $f(0) = 1$ and Fourier transform of*

f has range contained in $[0,1]$. *Let* $\mathbb{K}(x,y) = f(x-y)$ *be a kernel generated by this function.*

Then for every $\epsilon > 0$, *there exists* r_0, *so that for every* $r > r_0$ *we have the same bound (up to a constant) as in Theorem* 2.5.2. *In other words - as long as the assumptions from Theorem* 2.5.2 *are met, we have the following inequality:*

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||_F \leq \frac{(1+\epsilon)d^2(c_r\sqrt{\log n})^{d+1}}{\sqrt{n}} = (1+\epsilon)t_{n,d}.$$

In addition, a suitable value of r_0 *is given by*

$$r_0 = \sup\left\{ r \in \mathbb{R} : \int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt \geq r^{d+2}\sqrt{\epsilon}\sqrt{\frac{d}{n}}\left(\frac{c}{d}\right)^{d/2} \right\}.$$

Proof: First, we need to assure the existence of DPP with kernel $\mathbb{K}$. To do that, we will use the Theorem 1.2.5 - the spectrum of self-adjoint integral operator $\mathcal{K}$ of locally trace class determined by $\mathbb{K}$ must be contained in $[0,1]$. If Fourier transform of f has range contained in $[0,1]$, then this requirement is fulfilled and, hence, this kernel defines some DPP.

Now we will take a look at the expression $\mathbb{E}||\widehat{\Sigma} - \Sigma||_F$. As before in the proof of Theorem 2.5.2:

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||_F^2 \leq ||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2 + \mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2.$$

The element $\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2$ has the same bound as in the proof of Theorem 2.5.2, since, as one can observe, the bound does not rely on any Gaussian characteristics. We can rewrite the inequality:

$$\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2 \leq \frac{c^d r^{2d+4}}{d^{d-2}n}.$$

The element $||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2$ cannot be bounded by the same expression as in the proof of Theorem 2.5.2, since the Gaussian properties are not present. The following may be inferred from the proof:

$$\mathbb{E}v^T\widehat{\Sigma}v = \frac{|B(1)|r^{d+2}}{d+2} -$$

$$-\frac{1}{B(R-r)}\int\int_{||x-y||\leq r, ||x||\leq R-r} \langle x-y, v\rangle^2 \, g_2(x,y)dydx.$$

We can expand $g_2(x,y)$ to the following form:

$$g_2(x,y) = f(0)^2 - f(x-y)^2 = 1 - f^2(x-y), \tag{1}$$

so our expression amounts to:

$$\frac{|B(1)|r^{d+2}}{d+2}||v||^2 -$$

$$-\frac{1}{B(R-r)}\int_{||x||\leq R-r}\left(\int_{||u||\leq r}\langle u,v\rangle^2\,du-\int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du\right)dx.$$

The element:

$$\int_{||u||\leq r}\langle u,v\rangle^2\,du$$

is simply the expected value of inner product of uniform distribution and vector v on a circle $U\sim\mathcal{U}(B(1))$, so:

$$\int_{||u||\leq r}\langle u,v\rangle^2\,du=r^{d+2}|B(1)|\mathbb{E}[\langle U,v\rangle^2]=\frac{r^{d+2}|B(1)|}{d+2}.$$

Let us take a look at the other expression:

$$\int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du.$$

Since $f^2(u)$ is a density, we can also express it as the expected value:

$$\int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du=\mathbb{E}[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}],$$

where random variable Z has a density f^2. Because of that, the initial expression converges to $v^T\Sigma v$:

$$\mathbb{E}\left[v^T\widehat{\Sigma}v\right]=\mathbb{E}\left[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}\right]\xrightarrow{r\to\infty}\mathbb{E}\left[\langle Z,v\rangle^2\right]=v^T\Sigma v.$$

We got the same results as in the original proof, with the following convergence rate:

$$\mathbb{E}\left[\langle Z,v\rangle^2\right]-\mathbb{E}\left[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}\right]\leq\mathbb{E}\left[||Z||^2\mathbb{1}\{||Z||>r\}\right]=\int_{r^2}^{\infty}\mathbb{P}\left(||Z||^2>t\right)dt.$$

We can write the following convergence:

$$r^{-d-2}\int_{r^2}^{\infty}\mathbb{P}\left(||Z||^2>t\right)dt\overset{r\to\infty}{\longrightarrow}0.$$

Since this expression converges to 0, for every $\epsilon>0$, there exists r_0, so that for every $r>r_0$ we have the following:

$$\int_{r^2}^{\infty}\mathbb{P}\left(||Z||^2>t\right)dtr^{-d-2}\leq\sqrt{\epsilon}\sqrt{\frac{d}{n}}\left(\frac{c}{d}\right)^{d/2}.$$

We can utilize the following inequality from the proof of Theorem 2.5.2:

$$||\mathbb{E}\widehat{\Sigma}-\Sigma||_F^2\leq d||\mathbb{E}\widehat{\Sigma}-\Sigma||_{op}^2=d\sup_{||v||=1}[v^T\Sigma v-v^T\mathbb{E}\widehat{\Sigma}v]^2,$$

to obtain the bound on $||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2$:

$$||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2 \leq d\left(\int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt\right)^2 \leq \frac{\epsilon c^d r^{2d+4}}{d^{d-2}n}.$$

Since the bound on the expression $\mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2$ does not change, we get the following inequality:

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||_F^2 \leq ||\mathbb{E}\widehat{\Sigma} - \Sigma||_F^2 + \mathbb{E}||\widehat{\Sigma} - \mathbb{E}\widehat{\Sigma}||_F^2 \leq (1+\epsilon)\frac{c^d r^{2d+4}}{d^{d-2}n}.$$

QED

Since $c \cdot r^{d+2}$ has a polynomial growth for any positive constant c, this expression will quickly grow larger than the integral of the tail:

$$\int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt,$$

making r_0 small in magnitude. However, it is important to note that for very small values of ϵ, the parameter r_0 may still be large. We can distinguish sub-Gaussian random variables, due to their strong decay property. Since the tail distributions of these random variables converge faster than tail distribution of Gaussian random variable, we can state the following corollary.

Corollary 3.3.3 *Let function f be a symmetric function, such that f^2 is a sub-Gaussian density function, with covariance matrix Σ, $f(0) = 1$ and Fourier transform of f has range contained in $[0,1]$. Let $\mathbb{K}(x,y) = f(x-y)$ be a kernel generated by this function. Then as long as the assumptions from Theorem 2.5.2 are met, we have the following inequality:*

$$\mathbb{E}||\widehat{\Sigma} - \Sigma||_F \leq \frac{d^2(c_r\sqrt{\log n})^{d+1}}{\sqrt{n}} = t_{n,d}.$$

Proof: In the proof of Theorem 3.3.2 we have obtained the following inequality:

$$\mathbb{E}\left[\langle Z, v\rangle^2 \mathbb{1}\{||Z|| \leq r\}\right] - \mathbb{E}\left[\langle Z, v\rangle^2\right] \leq \mathbb{E}\left[||Z||^2 \mathbb{1}\{||Z|| > r\}\right] = \int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt,$$

for random variable Z with density f^2. If Z is sub-Gaussian, we can utilize the fact, that sub-Gaussian random variable decays faster than a Gaussian random variable. We obtain the following bound:

$$\int_{r^2}^{\infty} \mathbb{P}\left(||Z||^2 > t\right) dt \leq \int_{r^2}^{\infty} \mathbb{P}\left(||X||^2 > t\right) dt,$$

where $X \sim \mathcal{N}(0, \Sigma)$. The remaining part of the proof follows from the proof of the Theorem 2.5.2. QED

Remark 3.3.4 *In the Theorem* 3.3.2 *the assumption on Fourier transform is necessary, but what about other assumptions on f? Let f be a function from the theorem, with the exception $f(0) = a \neq 1$. As f^2 remains a density function, $f^2(0) \geq 0$. We get the dissimilarity:*

$$g_2(x,y) = f(0)^2 - f(x-y)^2 = a^2 - f^2(x-y).$$

We get the following:

$$\mathbb{E}v^T\widehat{\Sigma}v = \frac{|B(1)|r^{d+2}}{d+2}||v||^2 -$$

$$-\frac{1}{B(R-r)}\int_{||x||\leq R-r}\left(a^2\int_{||u||\leq r}\langle u,v\rangle^2\,du - \int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du\right)dx =$$

$$= (1-a^2)\frac{|B(1)|r^{d+2}}{d+2}||v||^2 + \mathbb{E}\,\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}],$$

for random variable Z with density f^2. Instead of converging to $v^T\Sigma v$, this expression diverges as $r\to\infty$, so this assumption is necessary.

Moving forward, let us examine the density assumption. For the following equation to hold, we stated, that function f^2 is a density function of some random variable Z:

$$\int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du = \mathbb{E}[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}],$$

However, it is enough for f^2 to be integrable and bounded by some density function k at each point. Indeed, then:

$$\int_{||u||\leq r}\langle u,v\rangle^2 f^2(u)du \leq \int_{||u||\leq r}\langle u,v\rangle^2 k(u)du \leq \mathbb{E}[\langle Z,v\rangle^2\,\mathbb{1}\{||Z||\leq r\}],$$

where Z is a random variable generated by k. In other words, it is enough to assume:

$$\int_{\mathbb{R}^n} f^2(x)dx \leq 1.$$

Bibliography

[1] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[2] D. Child. *The Essentials of Factor Analysis*. Bloomsbury Academic, 2006.

[3] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[4] S. Ghosh and P. Rigollet. Gaussian determinantal processes: a new model for directionality in data. *Proc. Natl. Acad. Sci. USA*, 117(24):13207–13213, 2020.

[5] R. A. Horn and Ch. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[6] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of Gaussian analytic functions and determinantal point processes*, volume 51 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2009.

[7] https://www.numbeo.com/cost-of living/.

[8] K. Johansson. Non-intersecting paths, random tilings and random matrices. *Probab. Theory Related Fields*, 123(2):225–280, 2002.

[9] I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, 1986.

[10] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[11] A. Lenard. States of classical statistical mechanical systems of infinitely many particles. I. *Arch. Rational Mech. Anal.*, 59(3):219–239, 1975.

[12] O Macchi. The coincidence approach to stochastic point processes. *Advances in Appl. Probability*, 7:83–122, 1975.

[13] C. Meyer. *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[14] J. Pitman and D.S. Rosenberg. Lecture 10 : Setup for the central limit theorem, from the course statistics 205a at university of california, berkeley, fall 2002.

[15] A. Soshnikov. Determinantal random point fields. *Uspekhi Mat. Nauk*, 55(5(335)):107–160, 2000.

[16] M. Stevens. Equivalent symmetric kernels of determinantal point processes, 2019.