UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Clustering large dimensional data via second order statistics: applications in wireless communications

Roberto Matheus Pinheiro Pereira

ADVERTIMENT La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del repositori institucional UPCommons (http://upcommons.upc.edu/tesis) i el repositori cooperatiu TDX (http://www.tdx.cat/) ha estat autoritzada pels titulars dels drets de propietat intel·lectual **únicament per a usos privats** emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei UPCommons o TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a UPCommons (*framing*). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del repositorio institucional UPCommons (http://upcommons.upc.edu/tesis) y el repositorio cooperativo TDR (http://www.tdx.cat/?locale-attribute=es) ha sido autorizada por los titulares de los derechos de propiedad intelectual **únicamente para usos privados enmarcados** en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio UPCommons No se autoriza la presentación de su contenido en una ventana o marco ajeno a UPCommons (*framing)*. Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the institutional repository UPCommons (http://upcommons.upc.edu/tesis) and the cooperative repository TDX (http://www.tdx.cat/?locale- attribute=en) has been authorized by the titular of the intellectual property rights **only for private uses** placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading nor availability from a site foreign to the UPCommons service. Introducing its content in a window or frame foreign to the UPCommons service is not authorized (framing). These rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.

Clustering Large Dimensional Data via Second Order Statistics: Applications in Wireless Communications

By

Roberto Matheus Pinheiro Pereira

Signal Theory and Communications Department
Universitat Politècnica de Catalunya

Supervisors:

PhD Supervisor 1: Xavier Mestre

PhD Supervisor 2: David Gregoratti

PhD Tutor: Antonio Pascual Iserte

A thesis submitted for the degree of

Doctor of Philosophy

Barcelona, November, 2023

À minha família.

Abstract

In many modern signal processing applications, traditional machine learning and pattern recognition methods heavily rely on the having a sufficiently large amount of data samples to correctly estimate the underlying structures within complex signals. The main idea is to understand the inherent structural information and relationships embedded within the raw data, thereby enabling a wide variety of inference tasks. Nevertheless, the definition of what constitutes a *sufficiently large* dataset remains subjective and it is often problem-dependent. In this context, traditional learning approaches often fail to learn meaningful structures in the cases where the number of features closely matches (or even exceeds) the number of observations. These scenarios emphasize the need for tailored strategies that effectively extract meaningful structured information from these high-dimensional settings. In this thesis we address fundamental challenges posed by applying traditional machine learning techniques in large dimensional settings.

Particularly, this thesis explores the comparison and clustering of symmetric positive definite matrices, such as covariance matrices, seen as objects in a Riemannian manifold. Initially, we investigate the asymptotic behavior of distances between sample covariance matrices by establishing a central limit theorem (CLT) that allows us to describe the asymptotic statistical law of these distances. We provide a general result for the class of distances that can be expressed as sums of traces of functions applied separately to each covariance matrix. This class includes conventional metrics like the Euclidean distance and Jeffreys' divergence, as well as more advanced distances found in Riemannian geometry, such as the log-Euclidean metric. Subsequently, we extend these findings to address the challenge of consistently estimating the distance between covariance matrices directly from the data. We complement this with a new statistical analysis of the asymptotic behavior of this category of distance estimators. Finally, we showcase the practical implications of these results by demonstrating how unsupervised learning algorithms can leverage them, with specific applications in wireless communications. In doing so, this thesis contributes with theoretical insights into unsu-

pervised learning mechanisms, with a practical orientation toward wireless communication systems. The overarching aim is to facilitate the integration and interpretability of unsupervised learning solutions in forthcoming wireless networks and broader signal processing challenges.

Resumen

En varias aplicaciones modernas de procesado de señales, los métodos tradicionales de aprendizaje automático y reconocimiento de patrones dependen en gran medida de la presencia de una cantidad de muestras de datos suficientemente grande para estimar correctamente las estructuras subyacentes en señales complejas. La idea principal es adquirir la información estructural inherente y las relaciones intrínsecas dentro de los datos brutos, lo que permite una amplia variedad de tareas de inferencia. Sin embargo, la definición de lo que constituye un conjunto de datos *suficientemente grande* sigue siendo subjetiva y a su vez depende del problema. En este contexto, los enfoques de aprendizaje tradicionales a menudo fallan al aprender estructuras significativas, especialmente en los casos en los que la dimensíon de los dados es muy similar (o incluso superior) al número de observaciones. Estos escenarios enfatizan la necesidad de diseñar nuevas estrategias que permitan extraer de forma eficaz información estructurada y significativa desde estos contextos de alta dimensionalidad. En esta tesis abordamos los desafíos fundamentales que plantean la aplicación de las técnicas tradicionales de aprendizaje automático en entornos de grandes dimensiones.

En concreto, esta tesis explora la comparación y el agrupamiento de matrices simétricas definidas positivas, como las matrices de covarianza, vistas como objetos en una variedad de Riemann. Inicialmente, investigamos el comportamiento asintótico de las distancias entre matrices de covarianza muestral estableciendo un teorema central del límite que nos permite describir la distribución asintótica de estas distancias. En concreto, presentamos un resultado general para la familia de distancias que pueden expresarse como sumas de trazas de funciones aplicadas por separado a cada matriz de covarianza. Esta familia incluye métricas convencionales como la distancia euclidiana y la divergencia de Jeffreys, así como distancias más avanzadas basadas en la geometría riemanniana, como la métrica log-euclidiana. Posteriormente, ampliamos estos hallazgos para abordar el reto de estimar coherentemente la distancia entre matrices de covarianza directamente a partir de los datos asociados a ellas. Complementamos este estudio con un nuevo

análisis estadístico del comportamiento asintótico de esta categoría de estimadores de distancia. Finalmente, mostramos las implicaciones prácticas de estos resultados demostrando cómo algoritmos de aprendizaje no supervisado pueden hacer uso de esas métricas y su respectivas distribuciones asintóticas, con aplicaciones específicas en la comunicación inalámbrica. De este modo, esta tesis aporta perspectivas teóricas sobre los mecanismos de aprendizaje no supervisado, con una orientación práctica hacia los sistemas de comunicación inalámbrica. El objetivo principal es facilitar la integración y la interpretabilidad de las soluciones de aprendizaje no supervisado en las redes inalámbricas de próxima generación, así como en desafíos más amplios en el procesado de señales.

Resum

En moltes aplicacions modernes de processament de senyals, els mètodes tradicionals d'aprenentatge automàtic i de reconeixement de patrons depenen de la disponibilitat d'una quantitat de mostres de dades suficientment gran per a estimar correctament les estructures subjacents en senyals complexes. La idea principal és entendre la informació estructural inherent i les relacions intrínseques a les dades en brut, permetent així una gran varietat de tasques d'inferència. No obstant això, la definició del que constitueix un conjunt de dades *suficientment gran* segueix sent subjectiva i sovint depèn del problema. En aquest context, els enfocaments d'aprenentatge tradicionals sovint no aconsegueixen aprendre estructures significatives en els casos en què el número de característiques o dimensionaliat de les dades coincideix amb (o fins i tot supera) el nombre d'observacions. Aquests escenaris emfatitzen la necessitat d'estratègies personalitzades que extreguin de manera efectiva informació estructurada significativa d'aquests entorns d'alta dimensió. En aquesta tesi abordem els reptes fonamentals que planteja l'aplicació de tècniques tradicionals d'aprenentatge automàtic en entorns de grans dimensions.

En concret, aquest treball explora la comparació i l'agrupament de matrius simètriques positives definides, com les matrius de covariança, vistes com objectes en una varietat de Riemann. Inicialment, investiguem el comportament asimptòtic de distàncies entre matrius de covariança mostrals, establint un teorema central del límit que ens permet descriure la distribució asimptòtica d'aquestes distàncies. En concret, presentem un resultat general per a la família de distàncies que poden expressar-se com a suma de traces de funcions aplicades per separat a cada matriu de covariança. Aquesta família inclou mètriques convencionals com la distància euclidiana i la divergència de Jeffrey, així com distàncies més avançades basades en la geometria riemanniana, com la mètrica log-euclidiana. Posteriorment, ampliem l'estudi per abordar el repte d'estimar coherentment la distància entre matrius de covariança directament a partir de les dades. Complementem això amb ua anàlisi estadística del comportament asimptòtic d'aquesta categoria d'estimadors de distància. Finalment, mostrem les implicacions pràctiques d'aquests resultats

provant com algoritmes d'aprenentatge no supervisat poden fer ús d'aquestes mètriques i les seves respectives distribucions asintòtiques, amb aplicacions específiques en comunicacions inalàmbriques. D'aquesta manera, aquesta tesi aporta perspectives teòriques sobre mecanismes d'aprenentatge no supervisat, amb una orientació pràctica cap a sistemes de comunicació inalàmbrica. L'objectiu principal és facilitar la integració i la interpretació de les solucions d'aprenentatge no supervisades a les xarxes inalàmbriques de la propera generació, així com a desafiaments més amplis en el processament de senyals.

Acknowledgements

This thesis is the result of a long journey, and as a consequence, there have been many people who have directly or indirectly contributed to its conclusion. I would like to use this page to express my gratitude.

First, I would like to express my deep gratitude to my PhD supervisors, Dr. Xavier Mestre and Dr. David Gregoratti, for their invaluable guidance, commitment, and patience throughout the course of my thesis research. I am absolutely sure that this thesis would have taken a much different shape without their constructive feedback on countless drafts and the ideas that have been discussed during these four years. Your support and dedication have been instrumental in the success of the initial steps of my research journey.

I would like to extend my appreciation to the entire CTTC research institution for promoting valuable research and the unique experiences it has offered me.

I am very grateful to Dr. Petar Popovski for kindly hosting me during my time at Aalborg University. I also want to express my appreciation to all the members of the Connectivity section who warmly welcomed me into their group and consistently demonstrated a willingness to collaborate.

Additionally, I want to thank the European Union for funding my research through the Marie Skłodowska-Curie Framework. My colleagues in the ITN Windmill Project, with whom I have shared countless unforgettable experiences and have extensively contributed during this four years of PhD.

I also would like to thank all the friends I have had the privilege of meeting in Spain, Denmark, Italy, Germany and many other places. While I refrain from listing names to avoid making this acknowledgment overly lengthy or inadvertently omitting any name, please know that each of you has played a significant role in making this journey much more enjoyable and memorable.

Last but certainly not least, I would like to express my deepest gratitude to my family for their unconditional support and love. Even while being an ocean away, they remained consistently present in my life. It is thanks to them that I have become the person and researcher I am today.

Contents

List of Figures

List of Tables

Notations

In general, uppercase boldface letters ($\mathbf{A}$) denote matrices, lowercase boldface letters ($\mathbf{a}$) denote (column) vectors and italics (a) denote scalars and generic non-commutative random variables. Below we provide a list of symbols and notations used throughout this thesis. The specific usage of a symbol might slightly vary depending on the context.

Matrix

$\mathbf{A}^{\mathrm{T}}, \mathbf{A}^{\mathrm{H}}$	Transpose and Hermitian (i.e., complex conjugate transpose) of a matrix $\mathbf{A}$, respectively.
$\mathrm{tr}[\mathbf{A}]$	Trace of a matrix $\mathbf{A}$.
$\det[\mathbf{A}], \mathrm{pdet}[\mathbf{A}]$	Determinant and pseudo determinant (i.e., product of non-zero eigenvalues) of a matrix $\mathbf{A}$, respectively.
$\|\mathbf{A}\|, \|\mathbf{A}\|_F$	Spectral and Frobenius norm of a matrix $\mathbf{A}$, i.e., $\sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}$ and $\sqrt{\mathrm{tr}[\mathbf{A}\mathbf{A}^{\mathrm{H}}]}$, respectively.
$\mathbf{R}_k$	True covariance matrix.
$\hat{\mathbf{R}}_k$	Sample covariance matrix obtained from $N_k \in \mathbb{N}^+$ samples.
$\mathbf{I}_M$	Identity matrix of size $M \times M$.
$\mathbf{P}_k$	Projection matrix defined as $\mathbf{A}(\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}$ for a full column-rank matrix $\mathbf{A}$ of size $M \times N_k$.
$[\mathbf{A}]_{ij}$	The entry in row i and column j of a matrix $\mathbf{A}$ (row and column indices begin at 1).
$\lambda(\mathbf{A}), \lambda_{\max}(\mathbf{A})$	Eigenvalues and largest eigenvalue of matrix $\mathbf{A}$.

$\mathbf{B}_M \asymp \mathbf{C}_M$	For two random matrices $\mathbf{B}_M$ and $\mathbf{C}_M$ of dimension $M \times M$, we write $\mathbf{B}_M \asymp \mathbf{C}_M$ if $M^{-1}\mathrm{tr}\left[\mathbf{A}_M\left(\mathbf{B}_M - \mathbf{C}_M\right)\right] \to 0$ almost surely as $M \to \infty$, where $\mathbf{A}_M$ is any sequence of deterministic $M \times M$ matrices with bounded norm.

Sets and Bounds

$\mathbb{N}, \mathbb{R}, \mathbb{C}$	The set of all natural, real and complex numbers, respectively.
$\mathbb{N}^M, \mathbb{R}^M, \mathbb{C}^M$	The set of M-dimensional vectors with entries in $\mathbb{N}, \mathbb{R}$ and $\mathbb{C}$, respectively.
$\mathrm{Re}\{\cdot\}, \mathrm{Im}\{\cdot\}$	Real and imaginary part, respectively.
$\mathcal{C}_0^-$	Negatively oriented simple closed contour enclosing zero and no other singularities.
C^-	Negatively oriented simple closed contour not crossing zero.
$\mathrm{supp}(f)$	Support of a function f, i.e., $\{x \in \Omega : f(x) \neq 0\}$ where here Ω is the function domain.
$\sup$	Supremum (least upper bound). If the set is finite, it coincides with the maximum (max).
$\inf$	Infimum (greatest lower bound). If the set is finite, it coincides with the minimum (min).

Statistical Terms

$\mathcal{N}(\mu, \sigma)$	Gaussian distribution centered at μ and with variance σ^2.
$\hat{\mathbb{E}}[\cdot]$	Empirical expectation which is equivalent to the empirical averaging of elements.
$\mathbb{E}[\cdot]$	Mathematical expectation.
$\Phi(x)$	CDF of a standard Gaussian distribution evaluated at x.

Distance Related Symbols

d_M	Distance between true covariance matrices of size $M \times M$.

$\hat{d}_M$	*Plug in* distances between sample covariance matrices of size $M \times M$.
$\tilde{d}_M$	Consistent estimator of distances between sample covariance matrices of size $M \times M$.
$\bar{d}_M, \mathfrak{m}_M, \sigma_M$	Asymptotic equivalent, (second order) mean and variance of *plug in* distances beteween covariance matrices of size $M \times M$, respectively.
$\tilde{\mathfrak{m}}_M, \tilde{\sigma}_M$	(Second order) mean and variance of consistent estimator of distances between covariance matrices of size $M \times M$, respectively.

Other Symbols

$\binom{n}{k}$	Binomial coefficient (i.e., $\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$) indexed by the pair of parameters $n \geq k \geq 0$.
$\frac{\partial f}{\partial x}$	Partial first order derivative of a function f with respect to the variable x.
$\frac{d}{dx} f(x), f'(x)$	First order derivative of a function f.
$\lim_{x \to \alpha} f(x)$	Limit for x approaching α.
$\mathbb{I}\{\cdot\}$	Indicator function, equal to 1 if event $\{\cdot\}$ is true and 0 otherwise.
i	Depending on the context used as imaginary unit $i = \sqrt{-1}$ or ith element.

Chapter 1

Introduction

Several traditional machine learning and pattern recognition methods rely on having sufficiently large amount of data samples to describe its intrinsic structure. The idea is that, by learning the inherent structural information and relationships within the original data, one can obtain sufficient knowledge to later perform different inference tasks. It is often the case that observations with latent features reside in some topological space. This implies the existence of underlying patterns within the data which can be potentially explored through learning mechanisms. For example, in face recognition, latent features are frequently employed to learn high-level facial descriptors that are used to compare and (possibly) match different faces or facial expressions. Similarly, in natural language problems, semantic and sentiment analysis tasks often leverage the use of latent features of the data to comprehend context and associate different parts of the text.

In general, when a sufficiently large amount of data is available, learning algorithms can approximate the topological space of the data, which is often beneficial for generalization purposes. However, the notion of what constitutes sufficiently large data is subjective and varies depending on the specific problem at hand and its dimensionality. For instance, in a more analytical view of the problem, classical limit theorems often assume that the data has fixed dimensionality and that by collecting more data samples one can better approximate their inherent structure. In this scenario, one observation is typically defined as a collection of samples, e.g., several readings from one sensor. Nonetheless, there exist several cases [1, 2] where the number of observations is large and comparable to the number of features. Hence, during the last decades, there has been growing interest on how to consistently learn from these *large dimensional settings*. Initial efforts have been made on trying to describe high dimensional data based only on their most important features via dimensionality reduction [3,4]. Indeed, such solutions are valid in

applications where the task is solvable in a considerably lower-dimensional space than the one where the original data lies [5]. However, by limiting the observed feature space, they fail to solve the general task of learning in large dimensional settings.

More related to this thesis, unsupervised learning aims to unveil patterns from unlabeled data. Particularly, cluster analysis aims to identify groups (or clusters) of observations that exhibit similar features and structures. This process often reveals hidden structures and relationships present within fractions of the data that may not be evident when considering the individual observations separately. Generally, by identifying distinct clusters in the data, one can make more informed decisions and improve the overall understanding of the underlying systems generating the data [6, 7]. Consequently, this makes cluster analysis particularly valuable in scenarios where obtaining labeled data proves challenging or impractical [8].

In this dissertation, our primary focus is on the analysis and synthesis of learning solutions from high-dimensional observations. Moreover, this thesis is strongly inspired by the field of wireless communications, particularly in the clustering of wireless devices in massive multiple-input multiple-output (MIMO) communications systems for a more efficient use of the spectrum resources. Therefore, in the following sections, we introduce several concepts related to clustering methods, with specific examples relevant to the field of wireless communications. Additionally, by the end of the chapter, we also present a more generic discussion of the problem, which involves the comparison (and clustering) of symmetric positive definite (SPD) matrices (e.g., covariance matrices) seen as objects lying on a Riemannian manifold.

1.1 Clustering of High Dimensional Data

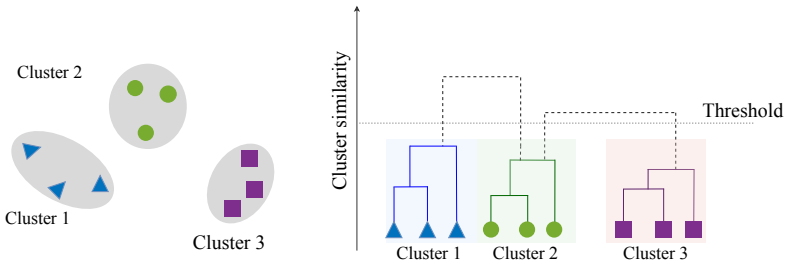
Clustering analysis usually rely on extracting meaningful and structured features of the data which are then fed into conventional clustering methods (see [4, 9, 10] for surveys applied to computer vision and data mining). One common approach is to compare the intrinsic features of different observations by directly studying the topological spaces to which they belong. A generic view of the problem considers the case where data belongs to several non-linear manifolds [11, 12]. In such scenarios, kernel methods are widely applied to account for the non-linearity in the data [13–15]. The main idea behind kernel methods is to map the input data (or features) from their original input space to a higher-dimensional space where the

relationship between the data is potentially easier to learn. However, the choice of the kernel function directly influences the results, and selecting the optimal kernel function remains a challenge in many applications [16, 17]. To address these challenges, data-driven solutions, such as deep neural networks (DNN) [12], are commonly employed to automatically learn the non-linearities and kernel functions. Nevertheless, DNN solutions often face limitations in terms of interpretability while demanding large amount of data and potentially expensive training procedures to properly learn these structures.

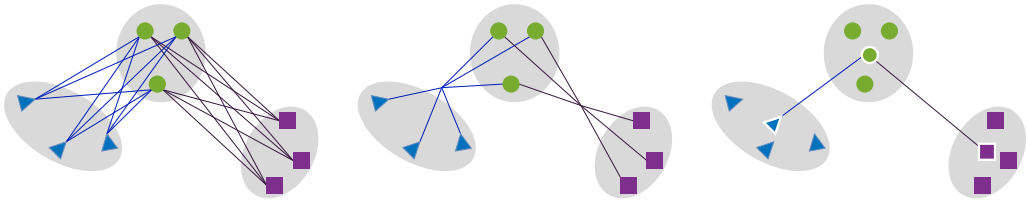
A classical clustering method, which will be extensively used during the examples presented in this thesis, is agglomerative hierarchical clustering. By consecutively combining different observations into groups, this unsupervised learning method provides an easy way to assess and interpret the relationship between these observations (see Figure 1.1a). In this bottom-up approach, initially, every observation forms a singleton, i.e., a cluster of one observation. At each merging step, the pair with the highest similarity (or, equivalently, the lowest distance) is merged to form a new cluster. The algorithm finishes when a target number of clusters is reached or when the highest similarity falls below a pre-defined threshold.

In the lower levels of the hierarchy, the comparison between singletons (individual elements) is usually done using some pre-defined metric. As elements start to be grouped together, the similarity between two different groups is often defined using some linkage method [18, 19], based on the pairwise similarity of the elements that form the two clusters. This is a common approach in agglomerative solutions as it allows for quick comparison of the non-equidimensional groups resulting from the sequential merging of elements into distinct groups. Typical choices include the average linkage (Figure 1.1b), which merges the two groups with the smallest average pairwise distance between elements within the compared clusters; and the Ward's linkage (Figure 1.1c), which merges the two groups that minimize the total within-cluster variance of the new group.

Another common solution, which requires some extra computation, is to compute and compare the centroids of the different groups (Figure 1.1d). The centroid of a group containing several elements is often defined either as the average among the observations in the cluster (e.g., the mean); or as the observation with smallest cumulative distance to the other observations of the group (e.g., the medoid). These ideas are applicable regardless of the clustering algorithm or the chosen comparison metric. For instance, the authors in [20] make a comparison among



(a) Hierarchical clustering solution)



(b) Average linkage (c) Ward's linkage (d) Centroid linkage

Figure 1.1: Visualization of a typical hierarchical clustering solution of three classes of objects (triangle, circle and square) and several linkage methods.

different metrics as well as between K-means and hierarchical clustering for the grouping and scheduling of user in frequency division duplexing (FDD) systems. As a result, they show that using hierarchical clustering outperforms K-means and K-medoids in both complexity and rate obtained after scheduling.

Naturally, there exists a vast variety of supervised and unsupervised learning methods which account for the different challenges present in high dimensional data. Several methods try to improve the learning process itself while others focus on enhancing the performance of such methods for a specific application. The objectives of this thesis are more aligned with the first case, as it proposes solutions applicable to general clustering methods. In the following subsections we focus on two concepts which are widely used throughout this thesis, namely, subspace and covariance analysis, while studying their applications into clustering methods. We further motivate the use of such analysis with examples in the field of wireless communications.

1.1.1 Subspace Clustering in Wireless Communications

In high-dimensional data analysis, a common approach to describe the underlying intrinsic features is through topological structures like subspaces. It is often the

case that the relevant features reside within certain subspaces which indicates the existence of lower-dimensional representations that capture the most descriptive characteristics of the observations. As mentioned above, this thesis is strongly motivated by wireless communications systems, hence we dedicate this section to exemplify how wireless communications systems can benefit from such subspace clustering schemes.

Let us start by noting that, in recent wireless communications systems, it is often desired to achieve simultaneous communications between one or more transmitters and multiple receivers. In these multi-user scenarios, multi-antenna radio access technologies are widely employed as means to enhance wireless communications spectral efficiency and connectivity. Particularly, space-division multiple access (SDMA) has traditionally been used to enhance spectral efficiency in the uplink. In the downlink, dirty-paper coding (DPC) achieves the channel capacity region by encoding the data at the transmitter side in order to pre-cancel interference at the receiver side. Since, in practice, DPC is difficult to implement, there has been intensive research on suboptimal solutions which combine superposition coding (SC) and spatial processing. For instance, non-orthogonal multiple access (NOMA) [21] has recently become a key mechanism to significantly enhance communication rates by allowing multiple users to superimpose their signals in the time and frequency domain. The resulting interference is then processed at the receiver side using successive interference cancellation (SIC). Similarly, joint spatial division multiplexing (JSDM) [22] and hierarchical rate splitting [23] use precoding to separate transmissions into clusters of users, and then apply the corresponding downlink processing to the resulting cluster-specific multiple-input multiple-output (MIMO) channels.

Obtaining the optimum user clustering for a particular transmission scheme is in general a very difficult problem that can only be solved by comparing all possible partitions of the different channels into groups. In order to avoid the exponentially high computational complexity of this process, when there exist sufficient degrees of freedom and a very strong line-of-sight signal, it is often assumed that groups of users with a significantly different DoA can be spatially separated, whereas those that are close together should have their interference processed using SIC. However, in the more realistic wireless scenario in which multipath is present, it is more reasonable to measure users proximity based on how well aligned the subspaces spanned by their channel matrices are. In that sense, suboptimal clustering schemes have been proposed in the literature that try to group the

multi-antenna wireless channels according to the similarity of the subspace they span (see for instance [23, 24]). The main idea is to use unsupervised clustering methods based on some similarity measure between different channel subspaces.

Inspired by the multiple-input and multiple-output (MIMO) channel clustering application in wireless communications, in [25] we focus on the agglomerative clustering of complex observations that belong to many distinct non-linear complex manifolds. In the wireless communications context, this translates into having receivers with different number of antennas scattered on the environment. Specifically, we consider the case in which the subspaces that describe each observation lay within non-equidimensional Grassmannian manifolds [26, 27]. The idea is to optimize the use of the available spatial degrees of freedom by identifying groups of users that are seen from similar angles (i.e. span a similar subspace) as a single spatial entity. It is typically easier to spatially multiplex different signals among well separated groups rather than attempt individual user multiplexing. Once these groups have been spatially multiplexed, one can process the signals within each group by either orthogonal (FDMA, TDMA) or non-orthogonal (NOMA, Rate Splitting) techniques [23, 24]. Using a similar approach as above, in [28], we also show that depending on the scenario, there exist several clustering solutions that might lead to high communication rates. It is therefore important that MIMO wireless channels are clustered in a structured manner and according to their proximity in terms of the subspace they span.

1.1.2 Clustering of Covariance Matrices

A large number of applications in machine learning and signal processing rely on the analysis of multivariate data, where each observation consists of readings from multiple entities or sensors. In such cases, patterns need to be extracted based on the dependence among these multiple readings, rather than just between the unique observations themselves. In other words, the relevant information is contained in the covariance pattern of multivariate observations, rather than the actual measurements. For instance, the covariance of a time series is closely related to its spectral density. Similarly, the covariance of signals received by a number of spatially distributed antennas/sensors is directly related to the spatial distribution of the corresponding sources [29]. In both cases, the covariance pattern holds vital information about the underlying relationships within the data, making it relevant for various analyses and learning procedures.

More related to this thesis, clustering according to covariance matrices has also recently become a common approach in multivariate and functional data analysis [29–31]. The main application in this setting consists in grouping segments of data that are represented by the same correlation structure. A prominent example is the study of electroencephalography (EEG) signals, where obtaining labeled data can be challenging. EEG signals are often analyzed on short-time readings obtained using a set of sensors capable of detecting electrical signals generated by different parts of the brain. These signals are often associated to some type of motor imagery, i.e., a mental execution of a movement which not necessarily results in a muscle activation. Depending on the motor imagery, different signal correlation structures can be obtained [32]. In this context, discovering and clustering the different covariances (correlation structures) of these signals becomes essential in order to detect the different patterns generated by the brain [8, 33, 34].

Similarly as in the previous section, relevant examples can also be found in the context of multi-user MIMO communications [35, 36]. Here the objective is to group a number of channel matrices so that channels which are seen from the same spatial locations are grouped together. Since spatial distribution is directly related to the inter-antenna signal covariance, one can alternatively cluster these MIMO channels according to their receivers' covariance matrices. The main difference lies in the fact that by directly considering the signals' covariance matrices, one gains insights regarding the angular location of the sources (using the basis of these matrices) as well as to the distance between the (distinct) sources and the receiver (which can now be obtained using the spectrum of the covariance matrices).

1.2 A Family of Riemannian Distances

So far, we have primarily discussed different clustering mechanisms and their applications in the field of wireless communications. However, a crucial design choice of such algorithms consists in measuring the relationship (e.g., similarity or distance) between different observations/clusters. This often depends on the nature of the observations and their structure. For instance, as already mentioned in the section above, in multivariate analysis, one often wishes to study the relationship between the covariance matrices that represent the (multivariate) data. That is because these second order statistics provide a concise descriptor of the multiple features which are only detectable across several samples (e.g., over time or frequency). By considering a more geometric approach, one can also observe

covariance matrices as symmetric positive definite (SPD) matrices lying in a Riemann manifold [37]. Consequently, common definitions based on the Euclidean space are no longer applicable to covariance matrices. Particularly, considering the Euclidean topology to handle SPD matrices is known to lead to inadequate conclusions [38]. Hence, instead of considering traditional definitions from the Euclidean space, one should consider the topological structure of the underlying manifold. In this section, we turn our focus to the study of the distance between covariance matrices.

In general, the concept of distance between these second order statistics have their own importance beyond the clustering problem. For example, image set classification is largely based on discriminant analysis on the intra-set covariance matrices [39]. By identifying each image set with its natural second-order statistic, the classification problem can be formulated as discriminating points in the Riemmann manifold of positive semidefinite matrices. Similar approaches can be found in diffusion tensor imaging [40] context, where the main descriptor is a vortex-depending covariance matrix; or applied to radar/sonar signal processing [41,42], where the spatial covariance matrix is used to capture the spatial characteristics of the clutter. In all these settings, the conventional Euclidean metric is not appropriate for measuring proximity between the observed covariance matrices, which belong to the set of positive semidefinite matrices. Hence, a number of studies propose to rely on metrics that consider the topological structure of underlying manifolds [37,40], e.g., Riemannian based distances [42,43] instead of the classical Euclidean distance.

In this dissertation, by taking into account the Riemann geometry of SPD matrices, we are primarily interested in studying the family of (squared) distances-between two covariance matrices, namely $\mathbf{R}_1$ and $\mathbf{R}_2$ that can be mathematically expressed as

$$d_M = \sum_{l=1}^{L} \frac{1}{M} \mathrm{tr}\left[f_1^{(l)}\left(\mathbf{R}_1\right) f_2^{(l)}\left(\mathbf{R}_2\right) \right] \tag{1.1}$$

for certain functions $f_1^{(l)}, f_2^{(l)} : \mathbb{C}^{M\times M} \to \mathbb{C}^{M\times M}$, $l = 1, \ldots, L$. Typically, these functions are understood to be the result of applying scalar analytic functions to the real eigenvalues of the Hermitian matrices $\mathbf{R}_j$, $j \in \{1, 2\}$. With some abuse of notation, $f_j^{(l)}, l = 1, \ldots, L$ will also denote these scalar functions.

Note that any distance that can be expressed in the form

$$d_M = \frac{1}{M} \mathrm{tr}\left[\left(f_1(\mathbf{R}_1) - f_2(\mathbf{R}_2) \right)^2 \right]$$

for two matrix-valued functions f_1, f_2 can be seen as a particularization of the general expression in (1.6). In particular, if these two functions are chosen so that $f_j(\mathbf{R}_j) = \mathbf{R}_j$, $j \in \{1,2\}$ we will recover the conventional Euclidean distance between covariance matrices, that is

$$d_M^E = \frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_1 - \mathbf{R}_2)^2\right]. \tag{1.2}$$

Likewise, the choice $f_j(\mathbf{R}_j) = \log(\mathbf{R}_j)$ will lead to the log-Euclidean distance [44] between covariance matrices

$$d_M^{LE} = \frac{1}{M}\mathrm{tr}\left[(\log \mathbf{R}_1 - \log \mathbf{R}_2)^2\right], \tag{1.3}$$

whereas the choice $f_j(\mathbf{R}_j) = (\mathbf{R}_j)^\alpha$ for some $\alpha > 0$ will lead to the power-Euclidean distance in [45].

Similarly, after proper normalization, the symmetrized version of the Kullback-Leibler divergence between two multivariate Gaussians (usually referred to as Jeffreys' divergence [46]) can be expressed by

$$d_M^{KL} = \frac{1}{2M}\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right] + \frac{1}{2M}\mathrm{tr}\left[\mathbf{R}_2\mathbf{R}_1^{-1}\right] - 1 \tag{1.4}$$

which also conforms to the general expression in (1.6).

Remark 1.1. *Naturally, the formulations above can also be applied to sample covariance matrices. Let us consider two sets of multidimensional observations of dimensionality M, which are denoted $\mathbf{y}_1(n) \in \mathbb{C}^{M\times 1}$ and $\mathbf{y}_2(n) \in \mathbb{C}^{M\times 1}$ respectively, $n \in \mathbb{N}$. We assume that the first sample set contains N_1 observations, whereas the second one is composed of N_2 observations. We will denote by $\mathbf{Y}_1$ and $\mathbf{Y}_2$ two matrices of dimensions $M \times N_1$ and $M \times N_2$ respectively, which contain the observations associated to each of these observation sets as columns, that is*

$$\mathbf{Y}_j = \left[\begin{array}{ccc} \mathbf{y}_j(1) & \dots & \mathbf{y}_j(N_j) \end{array}\right]$$

for $j \in \{1,2\}$. When the observations are zero mean, the covariance matrix of these observations are typically estimated using the sample covariance matrices, which are defined as

$$\hat{\mathbf{R}}_j = \frac{1}{N_j}\mathbf{Y}_j\mathbf{Y}_j^H. \tag{1.5}$$

Then the family of distances (1.1) becomes

$$\hat{d}_M = \sum_{l=1}^{L} \frac{1}{M}\mathrm{tr}\left[f_1^{(l)}\left(\hat{\mathbf{R}}_1\right) f_2^{(l)}\left(\hat{\mathbf{R}}_2\right)\right], \tag{1.6}$$

hereafter also referred as the plug-in distances.

Interestingly enough, the above formulation allows us to relate the results to the proximity measures between the column spaces of the observations $\mathbf{Y}_1, \mathbf{Y}_2$ in the undersampled regime, i.e., $N_1, N_2 < M$ (so that both $\mathbf{Y}_1, \mathbf{Y}_2$ are tall matrices). In this case, it is a standard approach to consider the sum of the squared sines of the principal angles between these subspaces as a valid distance in the Grassmann manifold [47], namely

$$\hat{d}_M^{SS} = \frac{1}{M}\mathrm{tr}\left[(\mathbf{P}_1 - \mathbf{P}_2)^2\right] \tag{1.7}$$

where $\mathbf{P}_i = \mathbf{Y}_i \left(\mathbf{Y}_i^H \mathbf{Y}_i\right)^{-1} \mathbf{Y}_i^H$ is the projection matrix onto the column space[1] of $\mathbf{Y}_i$, $i \in \{1, 2\}$. Moreover, the characterization of the above quantity has an interest beyond the framework of this dissertation and can be used to characterize independence tests based on canonical correlation analysis, which typically use $\mathrm{tr}\left[\mathbf{P}_1\mathbf{P}_2\right]$ as the relevant statistic to determine whether the two sets of observations are statistically independent (see [48–51] for the problem formulation and the asymptotic characterization when the observations are spatially white; results in this dissertation extend this characterization to the general spatially colored case).

Table 1.1: Summary of distances considered in this dissertation.

Name	**Definition**
Symmetrized KL divergence	$d_M^{KL} = \frac{1}{2M}\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right] + \frac{1}{2M}\mathrm{tr}\left[\mathbf{R}_2\mathbf{R}_1^{-1}\right] - 1$
Euclidean distance	$d_M^{E} = \frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_1 - \mathbf{R}_2)^2\right].$
Log-Euclidean distance	$d_M^{LE} = \frac{1}{M}\mathrm{tr}\left[(\log\mathbf{R}_1 - \log\mathbf{R}_2)^2\right]$
Subspace Similarity	$\hat{d}_M^{SS} = \frac{1}{M}\mathrm{tr}\left[(\mathbf{P}_1 - \mathbf{P}_2)^2\right]$

One of the problems that must be faced when applying these second order learning approaches is the fact that covariance matrices are generally unknown, and consequently the inherent distances must be estimated from the corresponding data. Furthermore, one usually needs to deal with the situation where the number of available samples is not much larger than the corresponding observation dimension. In order to address these issues, multiple contributions have focused on providing good estimators when both sample size and observation dimension are large but comparable in magnitude. This was for instance the case in [30], which considered an appropriate regularization of the covariance matrix in

[1] Notice $\hat{d}_M^{SS}$ only makes sense in the undersampled regime, otherwise, one cannot possibly define the original subspaces by using this definition.

a clustering application. More recently, there have been a number of contributions that directly propose consistent estimators of the distance between covariances, namely [52–55]. In these contributions, the main target were distance measurements between covariances that could be expressed as functions of $\mathbf{R}_1^{-1}\mathbf{R}_2$. Unfortunately, there exist important distances such as the log-Euclidean metric that do not really fall into this category. Hence, one of the objectives of this dissertation is to provide an asymptotic characterization of quantities like the ones above (and summarized in Table 1.1) when the dimensions of the matrices M and the corresponding sample sizes N_1, N_2 increase to infinity at the same rate, so that their quotient converges to a fixed quantity, namely $M/N_1 \to c_1$, $M/N_2 \to c_2$ for some $0 < c_1, c_2 < \infty$. The main advantage of this characterization with respect to the more conventional one (which assumes fixed M) is the fact that here all the dimensions (M, N_1, N_2) are comparable in magnitude even in the asymptotic regime, which makes the analysis more reliable in order to analyze the behavior of $\hat{d}_M$ in the finite sample size regime.

1.3 Contributions and Thesis Outline

The primary goal of this thesis is to conduct an analysis of the distance metrics described in the previous section together with their implications in unsupervised learning methods. Moreover, we demonstrate how these results can be directly applied to the field of signal processing, with practical examples in wireless communications. By doing so, this research contributes to a deeper understanding of high dimensional data analysis and its real-world applications in signal processing and wireless communications. To that end, in Chapter 2 we will see that, under standard statistical assumptions, the above (*plug-in*) distances $\hat{d}_M$ estimated from the sample covariance matrices all asymptotically behave as deterministic quantities. Moreover, we will also prove that these distances fluctuate around these deterministic equivalents as Gaussian random variables, and characterize their asymptotic (second order) mean and variance. The results of this chapter are based on the paper:

- **R. Pereira**, X. Mestre and D. Gregoratti, "On the Asymptotic Study of Distances Between Covariance Matrices," Submitted to IEEE Transactions on Signal Processing, 2023.

In Chapter 3, we focus on the development of a consistent estimator, denoted as $\tilde{d}_M$, for these metrics which better approximate the distance between the true covariance matrices. Building upon the previous contribution, we also provide a comprehensive characterization of the estimator's asymptotic behavior, leading to a Central Limit Theorem (CLT) that effectively describes this new metric. These results are also presented in:

- **R. Pereira**, X. Mestre and D. Gregoratti, "Consistent Estimation of Distances Between Covariance Matrices," Submitted to IEEE Transactions on Information Theory.

- **R. Pereira**, X. Mestre and D. Gregoratti, "Consistent Estimators of a New Class of Covariance Matrix Distances in the Large Dimensional Regime," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096831.

Moving forward to Chapter 4, we undertake a general comparison between these two estimators, namely the traditional *plug-in* one and the consistent one. The primary goal of this chapter is to illustrate how these estimators can be utilized to assist and enhance clustering tasks. These findings are based in the publications mentioned above, along with the results presented in this thesis. The combination of these chapters offers valuable insights into the efficacy and applications of the proposed estimators in addressing large-dimensional observation problems.

The remaining chapters of this thesis are dedicated to practical examples in the field of wireless communications. Specifically, Chapter 5 extends the findings from Chapter 2 and demonstrates how these results can be applied to offer a hierarchical clustering solution for user equipments (UEs) in a MIMO communications system. Furthermore, still in Chapter 5, we also present estimators of the first moments of the *plug in* distances which can be directly obtained from the sample covariance matrices. Particularly, we will use an alternative estimator of the deterministic equivalent of the original distance (valid only for identical covariance matrices) as a correction term which accounts for the comparison non-equidimensional subspaces. This can be translated to the wireless communications domain as the comparison of clusters of UEs where the total number of antennas is different at each cluster. Similarly, in Chapter 6, we carry out a similar analysis, but this time we propose a shallow neural network based clustering technique to learn and group different UEs according to their instantaneous noisy channel to maximize the rate

achieved using a hierarchical rate splitting mechanism. By providing these practical examples and their applications in wireless communications, this thesis contributes to the understanding and advancement of large-dimensional observation problems in real-world contexts. These contributions were originally published in:

- **R. Pereira**, X. Mestre and D. Gregoratti, "Subspace Based Hierarchical Channel Clustering in Massive MIMO," 2021 IEEE Globecom Workshops, Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/GCWkshps52748.2021.9682075.

- **R. Pereira**, X. Mestre and D. Gregoratti, "Clustering Complex Subspaces in Large Dimensions," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 5712-5716, doi: 10.1109/ICASSP43922.2022.9747627.

- **R. Pereira**, A. A. Deshpande, C. J. Vaca-Rubio, X. Mestre, A. Zanella, D. Gregoratti, E. de Carvaho and P. Popovski, "User Clustering for Rate Splitting using Machine Learning," 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 2022, pp. 722-726, doi: 10.23919/EUSIPCO55093.2022.9909639.

In Chapter 7 we conclude the main body of this thesis and outline potential future directions that can be pursued based on the findings and contributions of this thesis. Appendices A-D detail several of the proofs developed throughout this dissertation. Finally, Appendix E provide a concise summary of all the methods proposed in this thesis, encompassing their general and closed form solutions alongside with corresponding assumptions and remarks on their applicability. The purpose of this appendix is to serve as a convenient reference for readers, rather than introducing new material.

Collaborations

The following publications are not directly related to the content of this dissertation but have been conducted in collaboration with other researchers and institutions during the time of this PhD.

- A. A. Deshpande, **R. Pereira**, A. Zanella, A. Pastore, X. Mestre and F. Chiariotti, "Beam Aware Stochastic Multihop Routing for Flying Ad-hoc Networks," 2022 IEEE International Conference on Communications Workshops (ICC Workshops), Seoul, Korea, Republic of, 2022, pp. 1065-1070, doi: 10.1109/ICCWorkshops53468.2022.9814607.

- C. J. Vaca-Rubio, **R. Pereira**, X. Mestre, D. Gregoratti, Z. H. Tan, E. de Carvalho and P. Popovski, "Floor Map Reconstruction Through Radio Sensing and Learning by a Large Intelligent Surface," 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 2022, pp. 1-6, doi: 10.1109/MLSP55214.2022.9943430.

- X. Mestre, **R. Pereira** and D. Gregoratti, "Asymptotic Spectral Behavior of Kernel Matrices in Complex Valued Observations," 2021 IEEE Data Science and Learning Workshop (DSLW), Toronto, ON, Canada, 2021, pp. 1-6, doi: 10.1109/DSLW51110.2021.9523410.

Chapter 2

Asymptotic Study of Distances Between Sample Covariance Matrices

In this chapter, we are interested in examining the asymptotic properties of the overall class of distances $\hat{d}_M$ as defined in Section 1.2. To begin with, we demonstrate that the set of functions $f_j^{(l)}(\hat{\mathbf{R}}_j)$, when applied to the sample covariance matrices $\hat{\mathbf{R}}_j$, accepts a deterministic equivalent that can be obtained directly from its definition, along with the covariance matrix $\mathbf{R}_j$. Using these findings, we then analyze the behavior of $\hat{d}_M$ and provide specific characterizations for some particular distances of interest. Additionally, we delve deeper into the analysis by examining the fluctuations of $\hat{d}_M$ around its deterministic equivalent.

2.1 Preliminaries

The objective of this section is to analyze the asymptotic behavior of the distance between sample covariance matrices $\hat{d}_M$ in its most general form and to describe the necessary tools for the development of this dissertation. Let us start by recalling this definition, namely

$$\hat{d}_M = \sum_{l=1}^{L} \frac{1}{M} \mathrm{tr}\left[f_1^{(l)}\left(\hat{\mathbf{R}}_1\right) f_2^{(l)}\left(\hat{\mathbf{R}}_2\right)\right]$$

for certain functions $f_1^{(l)}, f_2^{(l)} : \mathbb{C}^{M\times M} \to \mathbb{C}^{M\times M}$, $l = 1, \ldots, L$. Then, throughout this thesis, we will make the following assumptions

Assumption 1 (As1): For $j \in \{1, 2\}$ and $k = 1, \ldots, N_j$ the observations $\mathbf{y}_j(k)$ (see Remark 1.1) are all independent and can be expressed as

$$\mathbf{y}_j(k) = \mathbf{R}_j^{\frac{1}{2}} \mathbf{x}_j(k)$$

where $(\cdot)^{\frac{1}{2}}$ denotes the square root of a matrix, $\mathbf{R}_j$ is an Hermitian positive definite matrix and $\mathbf{x}_j(k)$ is a vector of i.i.d. random entries with zero mean and unit variance. We will consider a binary variable ς that will indicate whether the observations are real or complex valued. If $\varsigma = 1$ the observations are real valued, whereas $\varsigma = 0$ indicates they are complex circularly symmetric.

Assumption 2 (As2): The different eigenvalues of $\mathbf{R}_j$ are denoted $0 < \gamma_1^{(j)} < \ldots < \gamma_{\bar{M}_j}^{(j)}$ ($j \in \{1, 2\}$) and have multiplicity $K_1^{(j)}, \ldots, K_{\bar{M}_j}^{(j)}$, where $\bar{M}_j$ is the total number of distinct eigenvalues. All these quantities may vary with M but we always have $\inf_M \gamma_1^{(j)} > 0$ and $\sup_M \gamma_{\bar{M}}^{(j)} < \infty$.

Assumption 3 (As3): The quantities N_1 and N_2 depend on M, that is $N_1 = N_1(M)$ and $N_2 = N_2(M)$. Furthermore, when $M \to \infty$ we have, for $j \in \{1, 2\}$, $N_j(M) \to \infty$ in a way that $M/N_j \to c_j$ for some constant $0 < c_j < \infty$ such that $c_j \neq 1$.

Assumption **(As1)** is quite standard, except for the fact that we introduce the binary variable ς to discriminate between real-valued ($\varsigma = 1$) or complex-valued ($\varsigma = 0$) observations. Assumption **(As2)** points out that the eigenvalues of the true covariance matrices $\mathbf{R}_j$, $j \in \{1, 2\}$ may behave freely as the dimensions of the matrix grow, as long as they fluctuate in a compact interval of the positive real axis independently of M. Finally, it is worth pointing out that **(As3)** explicitly excludes the case $c_j = 1$, mainly because addressing this case requires more elaborate technical tools that are out of the scope of this thesis.

In order to analyze the behavior of $\hat{d}_M$ under the above assumptions, we will build upon random matrix theory tools. To begin with, let us consider the function of complex variable $\omega_j(z)$, given by one of the solutions to the polynomial equation

$$z = \omega_j(z) \left(1 - \frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\gamma_m^{(j)}}{\gamma_m^{(j)} - \omega_j(z)} \right). \tag{2.1}$$

More specifically, if $z \in \mathbb{C}^+$ (upper complex semiplane), $\omega_j(z)$ is the only solution of the above equation located in $\mathbb{C}^+$. If $z \in \mathbb{C}^-$(lower complex semiplane), $\omega_j(z)$ is the only solution in $\mathbb{C}^-$. Conversely, if z is real valued, $\omega_j(z)$ is defined as the only real valued solution such that

$$\frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \left(\frac{\gamma_m^{(j)}}{\gamma_m^{(j)} - \omega_j(z)} \right)^2 < 1. \tag{2.2}$$

Finally, it will also be useful to denote as $\mu_0^{(j)} < \ldots < \mu_{\bar{M}_j}^{(j)}$ the solutions to the

equation

$$\mu\left(1-\frac{1}{N_j}\sum_{m=1}^{\bar{M}_j}K_m^{(j)}\frac{\gamma_m^{(j)}}{\gamma_m^{(j)}-\mu}\right)=0. \tag{2.3}$$

It can be easily shown [56] that $\mu_0^{(j)}<0$ when $N_j<M$ (undersampled case) and $\mu_0^{(j)}=0$ when $N_j>M$ (oversampled case).

Theorem 2.1. *Let $\mathbf{A}_M$ denote a sequence of deterministic $M\times M$ matrices with bounded spectral norm*[1]*. For $z\in\mathbb{C}^+$, consider the resolvents*

$$\hat{\mathbf{Q}}_j(z)=\left(\hat{\mathbf{R}}_j-z\mathbf{I}_M\right)^{-1}$$
$$\mathbf{Q}_j(\omega)=\left(\mathbf{R}_j-\omega\mathbf{I}_M\right)^{-1}$$

*for $j\in\{1,2\}$. Under **(As1)** – **(As3)** we have*

$$\frac{1}{M}\mathrm{tr}\left[\mathbf{A}_M\hat{\mathbf{Q}}_j(z)\right]-\frac{\omega_j(z)}{z}\frac{1}{M}\mathrm{tr}\left[\mathbf{A}_M\mathbf{Q}_j\left(\omega_j(z)\right)\right]\to 0$$

almost surely.

The above result is the cornerstone of the development of this work[2], and basically states that quantities that essentially depend on the sample covariance matrix through the resolvent $\hat{\mathbf{Q}}_j(z)$ asymptotically behave as deterministic quantities. In other words, the random resolvent $\hat{\mathbf{Q}}_j(z)$ is asymptotically equivalent to a deterministic quantity, given by

$$\bar{\mathbf{Q}}_j(z)=\frac{\omega_j(z)}{z}\mathbf{Q}_j\left(\omega_j(z)\right). \tag{2.4}$$

This result can be readily used to analyze the asymptotic behavior of $\hat{d}_M$ in (1.6) under some additional assumptions on the quantities $f_j^{(l)}(\hat{\mathbf{R}}_j)$. Consider the interval $\mathcal{S}_j=[\theta_j^-,\theta_j^+]$, where

$$\theta_j^-=\inf_M\left[\gamma_1^{(j)}\times\left(1-\sqrt{M/N_j}\right)^2\right],$$
$$\theta_j^+=\sup_M\left[\gamma_{\bar{M}}^{(j)}\times\left(1+\sqrt{M/N_j}\right)^2\right]. \tag{2.5}$$

[1]Throughout this thesis, when not specified, we will also denote spectral norm of a matrix, i.e., $||\mathbf{A}||=\sup(\sqrt{\lambda_{\max}(\mathbf{A}^{\mathrm{H}}\mathbf{A})})$, just by norm. Here $\lambda_{\max}(\cdot)$ represents the largest eigenvalue of a matrix.

[2]Theorem 2.1 is well known in the literature and it can be easily obtained from [57].

Observe, in particular, that the interval $\mathcal{S}_j$ is independent of M and that $\theta_j^- > 0$ regardless of whether $c_j < 1$ (oversampled regime) or $c_j > 1$ (undersampled regime). According to [58], all the positive eigenvalues of $\hat{\mathbf{R}}_j$ belong to $\mathcal{S}_j$ with probability one for all M sufficiently large. Using this property, we will assume that the functions $f_j^{(l)}(\hat{\mathbf{R}}_j)$ accept the following representation:

Assumption 4 (As4): For $j \in \{1,2\}$ and $l = 1,\ldots,L$, the quantity $f_j^{(l)}(\hat{\mathbf{R}}_j)$ can be expressed as

$$f_j^{(l)}\left(\hat{\mathbf{R}}_j\right) = \frac{1}{2\pi \mathrm{j}} \oint_{C_j^-} f_j^{(l)}(z)\hat{\mathbf{Q}}_j(z)dz \tag{2.6}$$

with probability one for all large M, where C_j^- is a negatively oriented simple closed contour enclosing $\mathcal{S}_j$ and not crossing zero and where, with some abuse of notation, $f_j^{(l)}(z)$ denotes a complex function analytic on an open set including C_j.

Observe that the above contour C_j does not depend on M and may be chosen differently for each $j \in \{1,2\}$ and each $l = 1,\ldots,L$, even if we omit this in the notation. In particular, C_j may be chosen to enclose zero for some l and not to enclose it for some other l. This distinction is sometimes important, for example if we choose $f_j^{(l)}(z) = 1$ when $N_j < M$. In this case we have (by a direct application of the matrix inversion lemma and residue calculus)

$$\frac{1}{2\pi \mathrm{j}} \oint_{C_j^-} \hat{\mathbf{Q}}_j(z)dz = \begin{cases} \mathbf{I}_M & C_j \text{ encloses } \{0\} \\ \mathbf{P}_j & \text{otherwise.} \end{cases}$$

Remark 2.1. *In order to emphasize this distinction, from now on we will use calligraphic letters (e.g. $\mathcal{C}_j^-$) when the contour encloses $\{0\}$ and roman letters (e.g. C_j^-) when the contour does not enclose $\{0\}$. The usual notation (e.g. C_j^-), which has been used so far, will refer to either one of the two cases indistinctively.*

For simplicity, assumption **(As4)** is restricted to the common situation in which $f_j^{(l)}(z)$ can be taken to be analytic in $\mathcal{S}_j$. In this case, the quantity $f_j^{(l)}(\hat{\mathbf{R}}_j)$ can be seen as the matrix that results from the application of $f_j^{(l)}(z)$ to the positive eigenvalues of $\hat{\mathbf{R}}_j$. Similarly, from (2.4) and by using Cauchy integration, we can also express the family of functions $f_j^{(l)}(\mathbf{R}_j)$ applied to the covariance matrices $\mathbf{R}_j$ by

$$f_j^{(l)}(\mathbf{R}_j) = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} f_j^{(l)}(\omega_j(z))\mathbf{Q}_j(\omega_j(z))\,\omega_j'(z)\,dz \tag{2.7}$$

where now $\omega_j'(z)$ denotes the derivative of $\omega_j(z)$. These observations will become building blocks for the following discussions. Particularly, in the next section we

will use assumption **(As4)** to study the asymptotic behavior of $\hat{d}_M$ in (1.6), and (2.7) in Chapter 3 to build a consistent estimator of d_M in (1.1) directly from the sample covariance matrices $\hat{\mathbf{R}}_1, \hat{\mathbf{R}}_2$.

2.2 Asymptotic behavior of $\hat{d}_M$

The result presented below follows from a direct application of Theorem 2.1 together with the dominated convergence theorem.

Proposition 2.1. *Under **(As1)-(As4)** and for any given sequence of deterministic matrices $\mathbf{A}_M$ with bounded norm,*

$$\frac{1}{M}\mathrm{tr}\left[f_j^{(l)}(\hat{\mathbf{R}}_j)\mathbf{A}_M\right] - \frac{1}{2\pi\mathrm{j}}\oint_{C_j^-}\frac{f_j^{(l)}(z)}{M}\mathrm{tr}\left[\bar{\mathbf{Q}}_j(z)\mathbf{A}_M\right]dz \to 0 \tag{2.8}$$

almost surely, where $\bar{\mathbf{Q}}_j(z)$ is as defined in (2.4).

Proof. Consider M sufficiently large and the probability set for which (2.6) holds true (a set which, by **(As4)**, has probability one). We can write

$$f_j^{(l)}\left(\hat{\mathbf{R}}_j\right) - \frac{1}{2\pi\mathrm{j}}\oint_{C_j^-} f_j^{(l)}(z)\bar{\mathbf{Q}}_j(z)dz = \frac{1}{2\pi\mathrm{j}}\oint_{C_j^-} f_j^{(l)}(z)\left[\hat{\mathbf{Q}}_j(z) - \bar{\mathbf{Q}}_j(z)\right]dz.$$

Moreover, by omitting the dependence on M in $\mathbf{A}_M$, we obtain

$$\left|\oint_{C_j^-} f_j^{(l)}(z)\frac{1}{M}\mathrm{tr}\left[\mathbf{A}\left(\hat{\mathbf{Q}}_j(z) - \bar{\mathbf{Q}}_j(z)\right)\right]dz\right| \leq$$
$$\leq \sup_{z\in C_j}\left|f_j^{(l)}(z)\right|\oint_{C_j}\left|\frac{1}{M}\mathrm{tr}\left[\mathbf{A}\left(\hat{\mathbf{Q}}_j(z) - \bar{\mathbf{Q}}_j(z)\right)\right]\right||dz|$$

where, obviously, $\sup_{z\in C_j}|f_j^{(l)}(z)| < \infty$ because of its analyicity. We know from Theorem 2.1 that $M^{-1}\mathrm{tr}[\mathbf{A}(\hat{\mathbf{Q}}_j(z) - \bar{\mathbf{Q}}_j(z))] \to 0$ almost surely for all fixed $z \in C_j \cap \mathbb{C}^+$. However, $\hat{\mathbf{Q}}_j(z)$ and $\bar{\mathbf{Q}}_j(z)$ are analytic functions on an open subset including C_j, and from Lemma D.1 in Appendix D.1 we have that $\sup_M \sup_{z\in C_j}\|\hat{\mathbf{Q}}_j(z)\| < \infty$ almost surely and $\sup_M \sup_{z\in C_j}\left\|\bar{\mathbf{Q}}_j(z)\right\| < \infty$. Hence, it follows that

$$\sup_{z\in C_j}\frac{1}{M}\mathrm{tr}[\mathbf{A}(\hat{\mathbf{Q}}_j(z) - \bar{\mathbf{Q}}_j(z))] \to 0$$

as a direct application of Montel' s theorem (see [59, Chapter 7]). □

Proposition 2.1 above shows that $f_j^{(l)}(\hat{\mathbf{R}}_j)$ has a deterministic asymptotic equivalent but fails to provide a closed form expression for this quantity. In practice, one needs to particularize the integral in (2.8) to the specific choices of $f_j^{(l)}(z)$ in order to obtain a closed form expression for the corresponding asymptotic equivalent. In order to do that, it turns out to be particularly useful to use the change of variable proposed in [56] using the invertible map $z \mapsto \omega_j(z)$. Let us denote by $z_j(\omega)$ the inverse of this map, namely

$$z_j(\omega) = \omega \left(1 - \frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\gamma_m^{(j)}}{\gamma_m^{(j)} - \omega}\right).$$

Note that this is different from (2.1) as here we have $z_j(\omega)$ as a function of ω. Moreover, let $z_j'(\omega)$ denote its derivative, that is

$$z_j'(\omega) = 1 - \Gamma_j(\omega) \tag{2.9}$$

where we have defined

$$\Gamma_j(\omega) = \frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \left(\frac{\gamma_m^{(j)}}{\gamma_m^{(j)} - \omega}\right)^2 = \frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j^2(\omega)\right]. \tag{2.10}$$

All this notation allows us to write, by direct application of the change of variables,

$$\frac{1}{2\pi \mathrm{j}} \oint_{C_j^-} f_j^{(l)}(z) \bar{\mathbf{Q}}_j(z) dz = \frac{1}{2\pi \mathrm{j}} \oint_{C_{\omega_j}^-} f_j^{(l)}(z_j(\omega)) \frac{\omega}{z_j(\omega)} \mathbf{Q}_j(\omega) z_j'(\omega) \, d\omega \tag{2.11}$$

where $C_{\omega_j}^- = \omega\left(C_j^-\right)$. Now, we can directly use Proposition 2.1 to establish that $\hat{d}_M$ in (1.6) has a deterministic asymptotic equivalent.

Corollary 2.1. *Under **(As1)-(As4)** we have $\hat{d}_M - \bar{d}_M \to 0$ with probability one, where*

$$\bar{d}_M = \sum_{l=1}^{L} \frac{1}{(2\pi \mathrm{j})^2} \oint_{C_1^-} \oint_{C_2^-} f_1^{(l)}(z_1) f_2^{(l)}(z_2) \frac{1}{M} \mathrm{tr}\left[\bar{\mathbf{Q}}_1(z_1) \bar{\mathbf{Q}}_2(z_2)\right] dz_1 dz_2. \tag{2.12}$$

Proof. This will be a direct consequence of Proposition 2.1 if we are able to establish that $\sup_M \|f_j^{(l)}(\hat{\mathbf{R}}_j)\| < \infty$ almost surely for $j \in \{1, 2\}$ and $l = 1, \ldots, L$. Now, it follows from the expression in (2.6) that we have the bound

$$\left\| f_j^{(l)}\left(\hat{\mathbf{R}}_j\right) \right\| \leq \frac{1}{2\pi} \oint_{C_j} \left| f_j^{(l)}(z) \right| \left\| \hat{\mathbf{Q}}_j(z) \right\| |dz|.$$

Again $\sup_{z \in C_j} |f_j^{(l)}(z)| < \infty$ and $\sup_{M, z \in C_j} \left\| \hat{\mathbf{Q}}_j(z) \right\| < \infty$ almost surely thanks to Lemma D.1 in Appendix D.1, so that we can conclude that $\sup_M \|f_j^{(l)}(\hat{\mathbf{R}}_j)\| < \infty$ with probability one. □

In the following subsections we will illustrate how to solve the double integral in (2.12) in some specific cases of interest, namely the Euclidean, symmetric Kullback-Leibler and subspace distances. The key point will always be the use of the change of variable stemming from the invertible map $z_j \mapsto \omega_j(z_j)$. As we will see below, this change of variable will allow us to obtain a closed form expression for the asymptotic equivalent $\bar{d}_M$.

2.2.1 Euclidean distance

The Euclidean distance in (1.6) can be expressed according to **(As4)** with all the contours enclosing $\{0\}$ and

$$\sum_{l=1}^{L} f_1^{(l)}(z_1) f_2^{(l)}(z_2) = (z_1 - z_2)^2 = z_1^2 - 2z_1 z_2 + z_2^2.$$

Consequently, to evaluate $\bar{d}_M^E$ we need to solve the integral in (2.11) with $f_j^{(l)}(z) = z$ and $f_j^{(l)}(z) = z^2$. Beginning with $f_j^{(l)}(z) = z$ we see that, using the change of variables above, we can write

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_j^-} z \bar{\mathbf{Q}}_j(z) dz = \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_j}^-} \omega \mathbf{Q}_j(\omega) z_j'(\omega)\, d\omega$$

where $z_j'(\omega)$ is as in (2.9) and $\mathcal{C}_{\omega_j}^- = \omega_j(\mathcal{C}_j^-)$. Now the right hand side of the above integrand only has singularities at the eigenvalues $\gamma_m^{(j)}$, $m = 1, \ldots, \bar{M}_j$, which are all enclosed by $\mathcal{C}_{\omega_j}^-$ [56]. We can therefore enlarge the contour $\mathcal{C}_{\omega_j}^-$ and apply a second change of variables $\zeta = \omega^{-1}$ in a way that $\zeta(\mathcal{C}_{\omega_j}^-)$ encloses zero and no other singularity. The corresponding integral takes the form

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_j^-} z \bar{\mathbf{Q}}_j(z) dz = \frac{-1}{2\pi \mathrm{j}} \oint_{\zeta(\mathcal{C}_{\omega_j}^-)} z_j'\left(\zeta^{-1}\right) \zeta^2 \left(\zeta \mathbf{R}_j - \mathbf{I}_M\right) d\zeta.$$

The only singularity of the integrand corresponds to a second order pole at $\{0\}$, so that observing that $\zeta(\mathcal{C}_{\omega_j}^-)$ is positively oriented and computing the residue at $\zeta = 0$ we have

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_j^-} z \bar{\mathbf{Q}}_j(z) dz = \mathbf{R}_j.$$

Following exactly the same approach one finds that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_j^-} z^2 \bar{\mathbf{Q}}_j(z) dz = \mathbf{R}_j^2 + \left(\frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j\right]\right) \mathbf{R}_j \tag{2.13}$$

and consequently

$$\bar{d}_M^E = \frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_1 - \mathbf{R}_2)^2\right] + \frac{1}{MN_1}\mathrm{tr}^2\left[\mathbf{R}_1\right] + \frac{1}{MN_2}\mathrm{tr}^2\left[\mathbf{R}_2\right].$$

Observe that, as expected, the asymptotic equivalent is different from the Euclidean distance between the true covariance matrices. This illustrates the fact that distances between sample covariance matrices are generally inconsistent estimators of the corresponding distance between the true covariance matrices. In some cases, such as in the case of the Euclidean distance, it is possible to modify $\hat{d}_M$ so that it converges to the Euclidean distance between the true covariances. However, this is not always possible for all distances in the form of (1.6), particularly in the undersampled regime.

2.2.2 Symmetrized Kullback-Leibler distance

The symmetrized KL distance in (1.4) and its generalization[3] can both be expressed as in **(As4)** with

$$\sum_{l=1}^{L} f_1^{(l)}(z_1) f_2^{(l)}(z_2) = \frac{1}{2}\frac{z_2}{z_1} + \frac{1}{2}\frac{z_1}{z_2} - 1$$

where in the first two terms the contours do not enclose zero, whereas they do in the last term. Therefore, to find the asymptotic equivalent $\bar{d}_M^{KL}$ we need to evaluate the integral in (2.11) with $f_j^{(l)}(z) = z$ and $f_j^{(l)}(z) = z^{-1}$, in both cases assuming that the corresponding contour does not contain zero. As the first case has already been considered before, let us study the integral for $f_j^{(l)}(z) = z^{-1}$ and observe that we can particularize (2.11) to

$$\frac{1}{2\pi \mathrm{j}}\oint_{\mathrm{C}_j^-} z^{-1}\bar{\mathbf{Q}}_j(z)dz = \frac{1}{2\pi \mathrm{j}}\oint_{\mathrm{C}_{\omega_j}^-} \frac{1-\Gamma_j(\omega)}{\omega^2\left(1-\frac{1}{N_j}\sum_{m=1}^{\bar{M}_j} K_m^{(j)}\frac{\gamma_m^{(j)}}{\gamma_m^{(j)}-\omega}\right)^2}\omega\mathbf{Q}_j(\omega)d\omega$$

where $\mathrm{C}_{\omega_j}^- = \omega_j(\mathrm{C}_j^-)$.

Now, recalling the definition of $\mu_0^{(j)} < \ldots < \mu_{\bar{M}_j}^{(j)}$ from (2.3), it turns out that $\mu_0^{(j)}$ is the only root in the above set that is not enclosed by $\mathrm{C}_{\omega_j}^-$, and therefore all the singularities of the above integrand fall inside $\mathrm{C}_{\omega_j}^-$ except for a potential singularity at $\mu_0^{(j)}$. Hence, we can enlarge $\mathrm{C}_{\omega_j}^-$ in the above integral so that it encloses $\mu_0^{(j)}$ if we then add the corresponding residue, which turns out to be equal to

$$\frac{\mathbf{R}_j\mathbf{Q}_j^2(\mu_0^{(j)})}{1-\Gamma_j\left(\mu_0^{(j)}\right)}$$

[3] In the undersampled regime, one can replace the inverse of the sample covariance $(\mathbf{R}_j)^{-1}$ by its respective the Moore-Penrose pseudoinverse $(\mathbf{R}_j)^{\#}$.

(notice that $\mathrm{C}_{\omega_j}^-$ is negatively oriented). Once this has been evaluated, we can apply the change of variables $\zeta = \omega^{-1}$, leading to

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_j^-} z^{-1} \bar{\mathbf{Q}}_j(z) dz = \frac{\mathbf{R}_j \mathbf{Q}_j^2(\mu_0^{(j)})}{1 - \Gamma_j\left(\mu_0^{(j)}\right)} + \\ + \frac{1}{2\pi \mathrm{j}} \oint_{C_0^-} \frac{1 - \Gamma_j\left(\zeta^{-1}\right)}{\left(1 - \frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\gamma_m^{(j)} \zeta}{\gamma_m^{(j)} \zeta - 1}\right)^2} \zeta^{-1} \mathbf{Q}_j(\zeta^{-1}) d\zeta$$

where now C_0^- is a negatively oriented contour enclosing zero and no other singularity. Noting that the second term of the above expression is zero (the integrand presents a removable singularity at zero), we can conclude that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_j^-} z^{-1} \bar{\mathbf{Q}}_j(z) dz = \frac{\mathbf{R}_j \mathbf{Q}_j^2(\mu_0^{(j)})}{1 - \Gamma_j\left(\mu_0^{(j)}\right)}.$$

With all the above intermediate results, it follows directly that

$$\bar{d}_M^{KL} = \frac{\mathrm{tr}\left[\mathbf{R}_1 \mathbf{Q}_1^2(\mu_0^{(1)}) \mathbf{R}_2\right]}{2M\left(1 - \Gamma_1\left(\mu_0^{(1)}\right)\right)} + \frac{\mathrm{tr}\left[\mathbf{R}_2 \mathbf{Q}_2^2(\mu_0^{(2)}) \mathbf{R}_1\right]}{2M\left(1 - \Gamma_2\left(\mu_0^{(2)}\right)\right)} - 1.$$

It is particularly interesting to note that in the oversampled situation, that is when $N_1, N_2 > M$, we have $\mu_0^{(1)} = \mu_0^{(2)} = 0$ and therefore

$$\bar{d}_M^{KL} = \frac{1}{2M}\left(\frac{N_1 \mathrm{tr}\left[\mathbf{R}_1^{-1} \mathbf{R}_2\right]}{N_1 - M} + \frac{N_2 \mathrm{tr}\left[\mathbf{R}_2^{-1} \mathbf{R}_1\right]}{N_2 - M}\right) - 1.$$

2.2.3 Subspace distance

The subspace distance also responds to the form in (1.6)-(2.6) with all the functions $f_j^{(l)}(z) = 1$ and where none of the contours encloses $\{0\}$, that is

$$\hat{d}_M^{SS} = \frac{N_1 + N_2}{M} + \frac{1}{2\pi^2} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \frac{\mathrm{tr}[\hat{\mathbf{Q}}_1(z_1) \hat{\mathbf{Q}}_2(z_2)]}{M} dz_1 dz_2.$$

Using the above integration technique we directly see that

$$\begin{aligned} \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_j^-} \bar{\mathbf{Q}}_j(z) dz &= \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_j}^-} \omega \mathbf{Q}_j(\omega) \frac{1 - \Gamma_j(\omega)}{z_j(\omega)} d\omega \\ &= \mu_0^{(j)} \mathbf{Q}_j(\mu_0^{(j)}) + \frac{1}{2\pi \mathrm{j}} \oint_{C_{\omega_j}^-} \omega \mathbf{Q}_j(\omega) \frac{1 - \Gamma_j(\omega)}{z_j(\omega)} d\omega \\ &= \mathbf{R}_j \mathbf{Q}_j(\mu_0^{(j)}) \end{aligned}$$

and consequently

$$\bar{d}_M^{SS} = \frac{N_1}{M} + \frac{N_2}{M} - \frac{2}{M}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\mu_0^{(1)})\mathbf{R}_2\mathbf{Q}_2(\mu_0^{(2)})\right]. \tag{2.14}$$

Obviously, this distance only makes sense in the undersampled regime (otherwise, one cannot possibly define the original subspaces).

2.3 Asymptotic fluctuations

In this section, we analyze the fluctuations of the above distances around their asymptotic equivalents. The main idea is to show that the distance $\hat{d}_M$ between sample covariance matrices asymptotically behaves as a Gaussian random variable around its asymptotic equivalent $\bar{d}_M$. More specifically, let us consider the following normalized random variable

$$\hat{\zeta}_M = M\left(\hat{d}_M - \bar{d}_M\right) = \frac{1}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-} g_M(z_1,z_2)\times \\ \times\,\mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2) - \bar{\mathbf{Q}}_1(z_1)\bar{\mathbf{Q}}_2(z_2)\right]dz_1dz_2$$

where $g_M(z_1,z_2) = \sum_{l=1}^{L} f_1^{(l)}(z_1)f_2^{(l)}(z_2)$. We will establish a central limit theorem (CLT) on $\hat{\zeta}_M$ that will basically state that it asymptotically behaves as a Gaussian random variable with a certain mean and variance. In order to introduce the relevant quantities that will describe the asymptotic mean and variance, we need some notation. For a given $M \times M$ deterministic matrix $\mathbf{A}$, we denote

$$\Omega_j(\omega;\mathbf{A}) = \mathbf{A} + \phi_j(\omega;\mathbf{A})\,\mathbf{I}_M \tag{2.15}$$

where $\phi_j(\omega;\mathbf{A})$ is the scalar function

$$\phi_j(\omega;\mathbf{A}) = \frac{\omega}{1-\Gamma_j(\omega)}\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j^2(\omega)\,\mathbf{A}\right]. \tag{2.16}$$

We define the asymptotic (second order) mean of $\hat{\zeta}_M$ as

$$\mathfrak{m}_M = \frac{\varsigma}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-}\frac{\omega_1}{z_1}\frac{\omega_2}{z_2}g_M(z_1,z_2)\mathfrak{m}(\omega_1,\omega_2)\,dz_1dz_2 \tag{2.17}$$

where we have introduced the function

$$\mathfrak{m}(\omega_1,\omega_2) = \mathfrak{m}_1(\omega_1,\mathbf{Q}_2(\omega_2)) + \mathfrak{m}_2(\omega_2,\mathbf{Q}_1(\omega_1))$$

with

$$\mathfrak{m}_j\left(\omega_j, \mathbf{A}\right) = \frac{1}{N_j} \frac{\mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j^3\left(\omega_j\right) \Omega_j\left(\omega_j; \mathbf{A}\right)\right]}{1 - \Gamma_j(\omega_j)} \tag{2.18}$$

for $j \in \{1, 2\}$, and where we have used the shorthand notation $\omega_j = \omega_j(z_j)$. At this point, it is also worth recalling the definition of the binary variable ς from **(As3)** which discriminates between real-valued ($\varsigma = 1$) and complex-valued ($\varsigma = 0$) observations. Then, interestingly enough, we observe that the second-order mean, $\mathfrak{m}_M$, consistently becomes 0 when dealing with complex-valued observations (see Appendix A.1). This will happen regardless of the specific distance metric.

The asymptotic variance is defined as

$$\sigma_M^2 = \frac{1+\varsigma}{(2\pi \mathrm{j})^4} \oint_{C_1^-} \oint_{C_1^-} \oint_{C_2^-} \oint_{C_2^-} g_M(z_1, z_2) g_M(z_1', z_2') \times \\ \times \frac{\omega_1 \omega_2}{z_1 z_2} \frac{\omega_1' \omega_2'}{z_1' z_2'} \Sigma^2\left(\omega_1, \omega_2, \omega_1', \omega_2'\right) dz_1 dz_2 dz_1' dz_2' \tag{2.19}$$

where (writing again $\omega_j = \omega_j(z_j)$ and $\omega_j' = \omega_j(z_j')$)

$$\begin{aligned} \Sigma^2\left(\omega_1, \omega_2, \omega_1', \omega_2'\right) = & \sigma_1^2\left(\omega_1, \omega_1'; \mathbf{Q}_2\left(\omega_2\right), \mathbf{Q}_2(\omega_2')\right) \\ & + \sigma_2^2\left(\omega_2, \omega_2'; \mathbf{Q}_1\left(\omega_1\right), \mathbf{Q}_1(\omega_1')\right) \\ & + \varrho(\omega_1, \omega_1', \omega_2, \omega_2') \end{aligned} \tag{2.20}$$

and where we have introduced the quantities (with some abuse of notation with respect to (2.10))

$$\Gamma_j(\omega, \omega') = \frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j\left(\omega\right) \mathbf{Q}_j\left(\omega'\right)\right]$$

together with

$$\varrho(\omega_1, \omega_1', \omega_2, \omega_2') = \frac{\mathrm{tr}^2\left[\mathbf{R}_1 \mathbf{Q}_1\left(\omega_1\right) \mathbf{Q}_1\left(\omega_1'\right) \mathbf{R}_2 \mathbf{Q}_2\left(\omega_2\right) \mathbf{Q}_2\left(\omega_2'\right)\right]}{N_1 N_2 \left(1 - \Gamma_1(\omega_1, \omega_1')\right)\left(1 - \Gamma_2(\omega_2, \omega_2')\right)} \tag{2.21}$$

and

$$\begin{aligned} \sigma_j^2\left(\omega, \omega'; \mathbf{A}, \mathbf{B}\right) = & \frac{1}{1 - \Gamma_j(\omega, \omega')} \frac{1}{N_j} \times \\ & \times \mathrm{tr}\left[\mathbf{R}_j \mathbf{Q}_j\left(\omega\right) \mathbf{Q}_j\left(\omega'\right) \Omega_j\left(\omega; \mathbf{A}\right) \mathbf{R}_j \mathbf{Q}_j\left(\omega\right) \mathbf{Q}_j\left(\omega'\right) \Omega_j\left(\omega'; \mathbf{B}\right)\right] \\ & + \frac{1}{\left(1 - \Gamma_j(\omega, \omega')\right)^2} \frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j^2\left(\omega\right) \mathbf{Q}_j\left(\omega'\right) \Omega_j\left(\omega; \mathbf{A}\right)\right] \times \\ & \times \frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j\left(\omega\right) \mathbf{Q}_j^2\left(\omega'\right) \Omega_j\left(\omega'; \mathbf{B}\right)\right]. \end{aligned} \tag{2.22}$$

We have now all the necessary notation to introduce the main result of this section.

Theorem 2.2. *Assume that* ***(As1)-(As4)*** *hold and that the observations are Gaussian distributed. If* $\liminf_{M\to\infty} \sigma_M^2 > 0$ *we have*

$$\frac{\hat{\zeta}_M - \mathfrak{m}_M}{\sigma_M} \to \mathcal{N}(0,1).$$

Proof. See Appendix A.1. □

The above result can be used to approximate the behavior of the distance between sample covariance matrices $\hat{d}_M$ for finite values of M, N_1, N_2. Indeed, one can approximate $\hat{d}_M$ as a Gaussian random variable with mean value $\bar{d}_M + \mathfrak{m}_M/M$ and variance σ_M^2/M^2. This will be a fundamental help in order to establish the performance of the different distance measures in specific problems. Let us now see how this general result particularizes to some of the distances that have been introduced before. Detailed proofs are provided in Appendix A.

2.3.1 Euclidean distance

By direct evaluation of the integrals in (2.17)-(2.19) when $g(z_1, z_2) = (z_1 - z_2)^2$ one can establish that $\mathfrak{m}_M$ takes the form

$$\mathfrak{m}_M^E = \varsigma \left(\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1^2\right] + \frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right] \right) \tag{2.23}$$

whereas $(\sigma_M)^2$ particularizes to

$$\begin{aligned} \frac{\left(\sigma_M^E\right)^2}{1+\varsigma} = & 2\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\right]\right)^2 + 2\left(\frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2^2\right]\right)^2 + \frac{4}{N_1 N_2}\mathrm{tr}^2[\mathbf{R}_1\mathbf{R}_2] \\ & + \frac{4}{N_1}\mathrm{tr}\left[(\mathbf{R}_1\Delta_1)^2\right] + \frac{4}{N_2}\mathrm{tr}\left[(\mathbf{R}_2\Delta_2)^2\right] \end{aligned} \tag{2.24}$$

where we have introduced $\Delta_j = (\mathbf{R}_1 - \mathbf{R}_2) + (M/N_j)\mathbf{I}_M$, for $j \in \{1, 2\}$. All the terms are obviously positive, so that one can easily see that the variance is uniformly bounded away from zero. This shows that the CLT holds for the Euclidean distance between sample covariance matrices.

Of particular interest is the expression of the variance when $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$. In this specific case, it simplifies to

$$\frac{\left(\sigma_M^E\right)^2}{1+\varsigma} = 2\left(\frac{N_1+N_2}{N_1 N_2}\right)^2 \mathrm{tr}^2\left[\mathbf{R}^2\right] + 4\left(\frac{1}{N_1^3} + \frac{1}{N_2^3}\right)\mathrm{tr}\left[\mathbf{R}^2\right]\mathrm{tr}^2\left[\mathbf{R}\right].$$

2.3.2 Symmetrized Kullback-Leibler distance

In this case, we need to consider the integrals in (2.17)-(2.19) for $g(z_1, z_2) = z_1/(2z_2) + z_2/(2z_1) - 1$ where only the contours of the last integral enclose $\{0\}$. One can show that the second order mean takes the form

$$\mathfrak{m}_M^{KL} = \varsigma \sum_{\substack{i,j\in\{1,2\} \\ i\neq j}} \frac{d[\omega_i \mathfrak{m}_i(\omega_i, \mathbf{R}_j)]/d\omega_i|_{\omega_i=\mu_0^{(i)}}}{2\left(1-\Gamma_i(\mu_0^{(i)})\right)} \tag{2.25}$$

where we recall that $\mathfrak{m}_j(\omega, \mathbf{A})$ is defined in (2.18).

Regarding the variance, one can proceed in a similar way to show that

$$\begin{aligned}\frac{\left(\sigma_M^{KL}\right)^2}{1+\varsigma} = &\frac{\partial^2 \left[\omega_1\omega_1' \Upsilon_{11}\left(\omega_1, \omega_1'\right)\right]/\partial\omega_1\partial\omega_1'|_{\omega_1=\omega_1'=\mu_0^{(1)}}}{4\left(1-\Gamma_1(\mu_0^{(1)})\right)^2} \\ &+ \frac{\partial^2 \left[\omega_2\omega_2' \Upsilon_{22}\left(\omega_2, \omega_2'\right)\right]/\partial\omega_2\partial\omega_2'|_{\omega_2=\omega_2'=\mu_0^{(2)}}}{4\left(1-\Gamma_2(\mu_0^{(2)})\right)^2} \\ &+ \frac{\partial^2 \left[\omega_1\omega_2 \Upsilon_{12}\left(\omega_1, \omega_2\right)\right]/\partial\omega_1\partial\omega_2|_{\omega_1=\mu_0^{(1)},\omega_2=\mu_0^{(2)}}}{2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}\end{aligned} \tag{2.26}$$

where we have defined

$$\begin{aligned}\Upsilon_{11}\left(\omega_1, \omega_1'\right) = &\frac{\mathrm{tr}^2\left[\mathbf{R}_2\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{Q}_1\left(\omega_1'\right)\right]}{N_1N_2\left(1-\Gamma_1(\omega_1, \omega_1')\right)} + \\ &+ \sigma_1^2\left(\omega_1, \omega_1'; \mathbf{R}_2, \mathbf{R}_2\right) + \frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_1\left(\omega_1'\right)\right]\end{aligned} \tag{2.27}$$

where $\Upsilon_{22}\left(\omega_2, \omega_2'\right)$ is defined equivalently but interchanging the two indexes $(1 \leftrightarrow 2)$ and where

$$\begin{aligned}\Upsilon_{12}\left(\omega_1, \omega_2\right) = &\frac{1}{N_1N_2}\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\right] \\ &- \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{Q}_2\left(\omega_2\right)\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\Omega_1\left(\omega_1; \mathbf{R}_2\right)\right] \\ &- \frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\Omega_2\left(\omega_2; \mathbf{R}_1\right)\right].\end{aligned} \tag{2.28}$$

The expression of the second order mean and variance can be significantly simplified in the oversampled case, where we will always have $\mu_0^{(1)} = \mu_0^{(2)} = 0$. In this situation, the second order mean particularizes to

$$\mathfrak{m}_M^{KL} = \frac{\varsigma}{2}\left[\frac{N_1\mathrm{tr}[\mathbf{R}_2\mathbf{R}_1^{-1}]}{(N_1-M)^2} + \frac{N_2\mathrm{tr}[\mathbf{R}_1\mathbf{R}_2^{-1}]}{(N_2-M)^2}\right]$$

whereas the variance takes the simple form

$$\frac{(\sigma_M^{KL})^2}{1+\varsigma} = \frac{N_1^2\Upsilon_{11}(0,0)}{4(N_1-M)^2} + \frac{N_2^2\Upsilon_{22}(0,0)}{4(N_2-M)^2} + \frac{N_1N_2\Upsilon_{12}(0,0)}{2(N_1-M)(N_2-M)}$$

where

$$\Upsilon_{11}(0,0) = \frac{N_1+N_2-M}{N_2(N_1-M)}\left[\mathrm{tr}[(\mathbf{R}_1^{-1}\mathbf{R}_2)^2] + \frac{\mathrm{tr}^2[\mathbf{R}_1^{-1}\mathbf{R}_2]}{N_1-M}\right]$$

with $\Upsilon_{22}(0,0)$ equivalently defined by swapping indexes $(1 \leftrightarrow 2)$, and where

$$\Upsilon_{12}(0,0) = \frac{M^2}{N_1N_2} - \frac{M}{N_1} - \frac{M}{N_2}.$$

The fact that $\liminf_M(\sigma_M^{KL})^2 > 0$ is easy to see in the oversampled case, by simply using the fact that $\alpha\mathrm{tr}(\mathbf{A}) + \beta\mathrm{tr}(\mathbf{A}^{-1}) > 2M\sqrt{\alpha\beta}$ for $\alpha, \beta > 0$ and positive definite $\mathbf{A}$. Indeed, by using this inequality with $\mathbf{A} = (\mathbf{R}_1^{-1}\mathbf{R}_2)^2$ and noting that the terms of the form $\mathrm{tr}^2(\cdot)$ are positive, we see that

$$\frac{(\sigma_M^{KL})^2}{1+\varsigma} > \frac{N_1N_2}{2(N_1-M)(N_2-M)} \times \\ \times \left(\frac{M(N_1+N_2-M)}{\sqrt{(N_1-M)(N_2-M)N_1N_2}} + \frac{M^2}{N_1N_2} - \frac{M}{N_1} - \frac{M}{N_2}\right).$$

Next, we observe that $(N_1+N_2-M)/(N_1-M) > 1+N_2/N_1$, so that

$$\frac{M^2(N_1+N_2-M)^2}{(N_1-M)(N_2-M)N_1N_2} > \left(\frac{M}{N_1}+\frac{M}{N_2}\right)^2$$

and we can conclude that

$$(\sigma_M^{KL})^2 > \frac{M^2}{2(N_1-M)(N_2-M)}$$

which is bounded away from zero.

2.3.3 Subspace distance

In the case of the subspace distance, it is shown in Appendix A that the second order mean takes the form

$$\mathfrak{m}_M^{SS} = -2\mu_0^{(1)}\mathfrak{m}_1(\mu_0^{(1)}, \mathbf{R}_2\mathbf{Q}_2(\mu_0^{(2)})) - 2\mu_0^{(2)}\mathfrak{m}_2(\mu_0^{(2)}, \mathbf{R}_1\mathbf{Q}_1(\mu_0^{(1)})). \tag{2.29}$$

whereas the asymptotic variance can be written as

$$\begin{aligned}\frac{\left(\sigma_M^{SS}\right)^2}{1+\varsigma} &= 4\left(\mu_0^{(1)}\right)^2 \sigma_1^2\left(\mu_0^{(1)}, \mu_0^{(1)}; \mathbf{R}_2\mathbf{Q}_2\left(\mu_0^{(2)}\right), \mathbf{R}_2\mathbf{Q}_2\left(\mu_0^{(2)}\right)\right) \\ &+ 4\left(\mu_0^{(2)}\right)^2 \sigma_2^2\left(\mu_0^{(2)}, \mu_0^{(2)}; \mathbf{R}_1\mathbf{Q}_1\left(\mu_0^{(1)}\right), \mathbf{R}_1\mathbf{Q}_1\left(\mu_0^{(1)}\right)\right) \\ &+ 4\left(\mu_0^{(1)}\mu_0^{(2)}\right)^2 \frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1^2(\mu_0^{(1)})\mathbf{R}_2\mathbf{Q}_2^2(\mu_0^{(2)})\right]}{N_1N_2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}. \end{aligned} \tag{2.30}$$

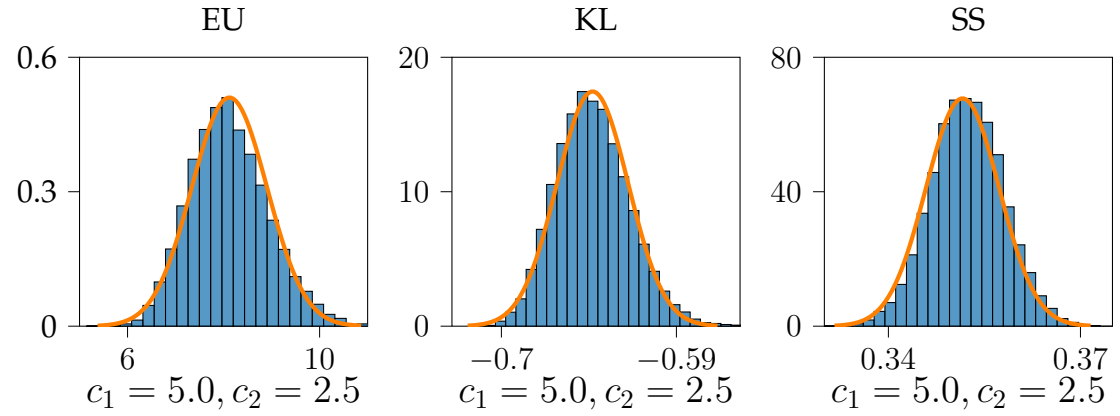
Close examination of the expression of the variance reveals that, since $\mathbf{Q}_j(\mu_0^{(j)})$ is positive definite, the first two terms of (2.30) are non-negative. Moreover, it is easy to see that $|\mu_0^{(j)}| \geq (M/N_j - 1)\gamma_{\bar{M}_j}^{(j)}$, which is bounded away from zero. Finally, a direct application of Lemma D.1 in Appendix D.1 shows that the third term in (2.30) is bounded away from zero, and hence the CLT holds.

2.4 Numerical Consistency of Asymptotic Descriptors

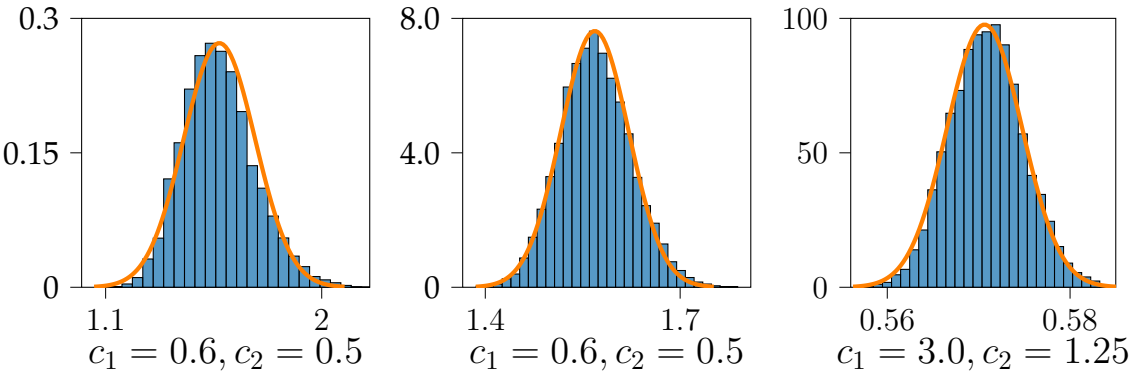
In order to validate the results presented above, we consider two multidimensional observation sets $\mathbf{Y}_1 \in \mathbb{C}^{M\times N_1}$ and $\mathbf{Y}_2 \in \mathbb{C}^{M\times N_2}$ associated to two (possibly distinct) Toeplitz covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ with first rows $[\rho_j^0, \ldots, \rho_j^{M-1}]$, for $j \in \{1, 2\}$. In this section, through numerical evaluation, we assess the consistency of the asymptotic descriptors ($\bar{d}_M, \mathfrak{m}_M$ and σ_M) of the family of metrics $\hat{d}_M$ defined in Chapter 1.6. Specifically, we are interested in the particularizations $\hat{d}_M^E$, $\hat{d}_M^{KL}$, $\hat{d}_M^{SS}$. We compare the asymptotic descriptors from the theorems above against the empirical distribution of their respective metrics, namely, Euclidean (EU), Symmetrized Kullback Leibler (KL) and subspace (SS) distances. Figure 2.1 compares the histogram (in blue) and the asymptotic values (in orange) of these metrics for $M = 100, \rho_1 = 0.7, \rho_2 = 0.8$ and some specific choices of c_1, c_2. Observe that there is a very good match between the asymptotic and the empirical distribution regardless of the considered metric or of whether c_1, c_2 are large or small. In other words, we have that the random distribution of the distance $\hat{d}_M$ seems to be correctly approximated by

$$\hat{d}_M \sim \mathcal{N}\left(\bar{d}_M + \frac{\mathfrak{m}_M}{M}, \frac{\sigma_M^2}{M^2}\right). \tag{2.31}$$

Similar results are also observed by comparing the Normalized Mean Squared Error (NMSE) between the empirical and asymptotic first- and second-order mo-



(a) Undersampling



(b) Large sampling

Figure 2.1: Histogram of empirical distribution (in blue) and asymptotic descriptors (in orange) of the different metrics EU, KL and SS.

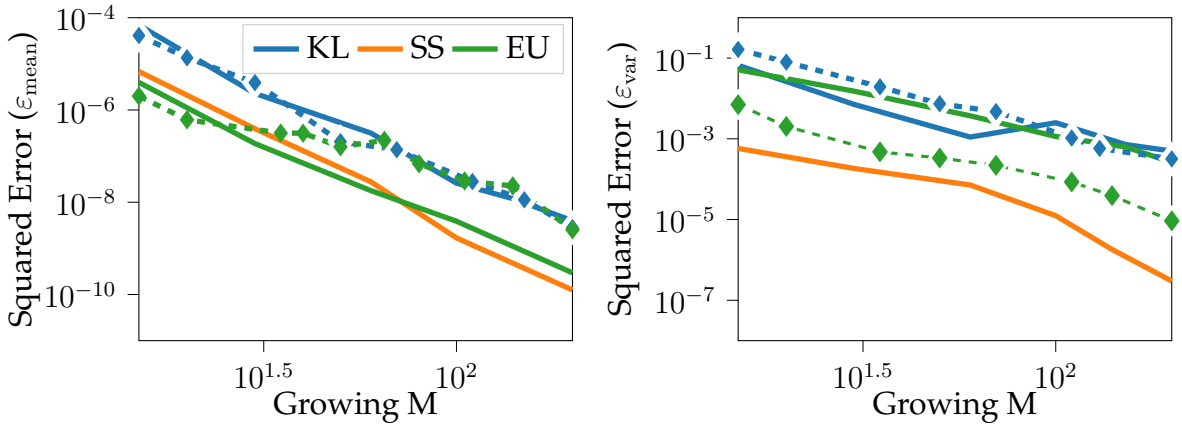
ments of $\hat{d}_M$, given by

$$\varepsilon_{\text{mean}} = \frac{\left(\hat{\mathbb{E}}[\hat{d}_M] - (\bar{d}_M + M^{-1}\mathfrak{m}_M)\right)^2}{\left(\bar{d}_M + M^{-1}\mathfrak{m}_M\right)^2} \tag{2.32}$$

and

$$\varepsilon_{\text{var}} = \frac{\left(\hat{\text{var}}[\hat{d}_M] - M^{-2}\sigma^2\right)^2}{M^{-4}\sigma^4},$$

respectively, where the empirical expectation ($\hat{\mathbb{E}}[\hat{d}_M]$) and variance ($\hat{\text{var}}[\hat{d}_M]$) are obtained from the observations by replacing the expectation with empirical averages. Figure 2.2 presents these results for growing M, N_1, N_2 and fixed c_1, c_2. The empirical quantities (expectation and variance) are estimated over 10^5 samples. The solid lines portray the results for the undersampled regime ($c_1 = 5.0$ and $c_2 = 2.5$) while the dashed lines represent the results for the oversampled regime ($c_1 = 0.5$ and



(a) SE of Mean

(b) SE of Variance

Figure 2.2: Normalized mean squared error (y-axis) between asymptotic descriptors (mean and variance) and their empirical values, obtained from multiple realizations of $\hat{d}_M$, for growing M (x-axis) in the undersampled (solid lines) and oversampled (dashed lines) scenarios.

$c_2 = 0.7$). For all the three distances we observe that, as the system grows (x-axis), the NMSE (y-axis) of the compared quantities (empirical and asymptotic descriptors) decay (tend to zero). Moreover, we notice that for a specific moment (mean or variance) and fixed asymptotic regime (under or oversampled), the NMSE for the different metrics are usually in the same order of magnitude. This comes as a consequence of the fact that all the asymptotic regimes studied in this work follow as a particularization of the generic results presented in (2.12), (2.17) and (2.19).

These results support the idea that our descriptors are consistent both when $\mathbf{R}_1 = \mathbf{R}_2$ and $\mathbf{R}_1 \neq \mathbf{R}_2$ for the metrics studied in this chapter, namely the Euclidean, symmetric Kullback-Leibler and subspace distances. In Chapter 4, we further develop these results to the task of clustering observations according to their respective distributions. Particularly, motivated by need of such mechanism in wireless communications, we will discuss how to cluster user equipments based on the alignment between the subspaces that span their channel matrices.

Chapter 3

Consistent Estimators of Distances Between True Covariance Matrices

As mentioned in the previous chapter, one of the problems that must be faced when applying second order learning approaches is the fact that covariance matrices are generally unknown and consequently the inherent distances must be estimated from the corresponding data. As discussed throughout Chapters 2, one way to do so is by plugging the two sample covariance matrices $\hat{\mathbf{R}}_j, j \in \{1,2\}$, as defined in (1.5), into (1.1) to obtain $\hat{d}_M$ as in (1.6). Unfortunately, this naive approach (hereafter denoted by *plug-in* distance) only approximates the original distance d_M between the associated covariance matrices $\mathbf{R}_j$ up to a certain bias, i.e., $|\hat{d}_M - d_M|$ converges to a constant different from zero as M, N_j grow large (see results in Section 3.4 for a detailed comparison). In this chapter, we circumvent this issue by proposing a consistent estimator for d_M. More formally, for a certain collection of functions $f_1^{(l)}, f_2^{(l)} : \mathbb{C}^{M\times M} \to \mathbb{C}^{M\times M}, l = 1,\dots,L$, applied to the covariances matrices $\mathbf{R}_1, \mathbf{R}_2$, we propose an estimator of d_M, denoted $\tilde{d}_M$ that is consistent (i.e. $d_M - \tilde{d}_M \to 0$) when the observation dimension (M) and the number of samples ($N_j, j = 1,2$) grow to infinity at the same rate.

To do so, we will assume that the functions[1] $f_j^{(l)}$ ($j \in \{1,2\}$) are analytic on a subset including all the eigenvalues of $\mathbf{R}_j$. So that we can express (by Cauchy integration of each element of the matrix)

$$f_j^{(l)}(\mathbf{R}_j) = \frac{1}{2\pi \mathrm{j}} \oint_{C_\omega^-} f_j^{(l)}(\omega)\mathbf{Q}_j(\omega)\, d\omega \tag{3.1}$$

[1]We recall that with some abuse of notation, we also defined $f_j^{(l)} : \mathbb{C}^M \to \mathbb{C}^M$ to be the scalar function applied to the eigenvalues of the covariance matrix $\mathbf{R}_j$.

where C_ω^- is a negatively oriented simple closed contour enclosing all the eigenvalues of $\mathbf{R}_j$ and

$$\mathbf{Q}_j(\omega) = (\mathbf{R}_j - \omega \mathbf{I}_M)^{-1}$$

is the resolvent defined in Section 2.1. Moreover, by considering the change of variable $\omega \mapsto z$ (introduced in (2.1)) we can reformulate the above expression and write

$$f_j^{(l)}(\mathbf{R}_j) = \frac{1}{2\pi \mathrm{j}} \oint_{C^-} f_j^{(l)}(\omega_j(z)) \mathbf{Q}_j(\omega_j(z)) \omega_j'(z)\, dz \tag{3.2}$$

where now $C^- = \omega_j^{-1}(C_\omega^-)$ (see also Chapter 2 and Remark 2.1). Then, since the contour enclosing zero C^- can be chosen independently of M, we can find asymptotic equivalents of $f_j^{(l)}(\mathbf{R}_j)$ by essentially finding asymptotic equivalents of the quantities inside the argument of the integral above (3.2). We formulate the result in the following section, which basically states the consistent estimator takes the form

$$\tilde{d}_M = \sum_{l=1}^{L} \frac{1}{M} \mathrm{tr}\left[\hat{h}_1^{(l)}\left(\hat{\mathbf{R}}_1\right) \hat{h}_2^{(l)}\left(\hat{\mathbf{R}}_2\right)\right]$$

for some matrix-valued functions $\hat{h}_j^{(l)}(\hat{\mathbf{R}}_j)$ of the sample covariance matrix which are asymptotically equivalent to $f_j^{(l)}(\mathbf{R}_j)$.

3.1 Improved Estimation of Riemannian Distances

Let us start by assuming that the functions $f_j^{(l)}(\omega)$ in (3.1) are sufficiently regular, in the sense that they are analytic in a sufficiently large region of the complex plane. One possibility would be to assume analycity in $\mathbb{C}\backslash\mathbb{R}^-$ (that is the whole complex plane except for the negative real axis). However, in practice this would rule out a number of situations in which we can achieve a consistent estimator even if the number of available samples is lower than the observation dimension, that is $N_j < M$ (undersampled case). So, instead, we will consider analytical functions on the whole complex plane except for only a subset of the negative real axis. In addition, it is worth pointing out that the analycity of the function $f_j^{(l)}(\omega)$ can easily be replaced by considering a wider class of continuously differentiable functions on the plane (except for a portion of the negative real axis). However, we choose to keep it here in order to simplify the proofs and because it clearly holds for all the covariance distance measures discussed in this thesis, namely the Euclidean, symmetrized Kullback-Leibler and log-Euclidean distances (see Section 1.2).

Assumption 5 (As5): For $j \in \{1,2\}$ and $l = 1,\ldots,L$, the functions $f_j^{(l)}(\omega)$ are analytic on the set $\mathbb{C}\backslash(-\infty,\mu_{\text{inf}}^{(j)}]$, where $\mu_{\text{inf}}^{(j)} = \inf_M \mu_0^{(j)}$ and $\mu_0^{(j)}$ is the smallest solution to the equation

$$0 = \mu\left(1 - \frac{1}{N_j}\text{tr}\left[\mathbf{R}_j\mathbf{Q}_j\left(\mu\right)\right]\right) \tag{3.3}$$

which is a restatement of (2.3). In particular, if $\inf_M N_j/M > 1$ (oversampled regime[2]) we have $\mu_{\text{inf}}^{(j)} = 0$ whereas $\mu_{\text{inf}}^{(j)} < 0$ if $\sup_M N_j/M < 1$ (undersampled regime).

Remark 3.1. *For two random matrices $\mathbf{B}_M$ and $\mathbf{C}_M$ and two analytic functions $f,h : \mathbb{C}\to\mathbb{C}$ applied to the eigenvalues of these matrices, we write $f(\mathbf{B}_M) \asymp h(\mathbf{C}_M)$ if*

$$\frac{1}{M}\text{tr}\left[\mathbf{A}_M\left(f(\mathbf{B}_M) - h(\mathbf{C}_M)\right)\right] \to 0$$

almost surely as $M\to 0$, where $\mathbf{A}_M$ is any sequence of deterministic $M\times M$ matrices with bounded norm. Moreover, we will frequently use the equivalence

$$f(\mathbf{A}_M) \equiv \oint_C f(z)(z\mathbf{I}_M - \mathbf{A}_M)^{-1}dz$$

to represent the contour integral

$$[f(\mathbf{A_M})]_{ij} = \oint_C f(z)(z\mathbf{I}_M - \mathbf{A}_M))^{-1}_{ij}dz$$

applied element wise for $i,j = 1,\ldots,M$.

Proposition 3.1. *Under **(As1)-(As3)**, from Section 2.1, and **(As5)**, defined above, we have*

$$f_j^{(l)}\left(\mathbf{R}_j\right) \asymp \hat{h}_j^{(l)}\left(\hat{\mathbf{R}}_j\right) \equiv \frac{1}{2\pi\text{j}}\oint_{\text{C}^-}\hat{h}_j^{(l)}(z)\hat{\mathbf{Q}}_j(z)dz, \tag{3.4}$$

where, with some abuse of notation, $\hat{h}_j^{(l)}(z)$ denotes both a $\mathbb{C}^{M\times M}\to\mathbb{C}^{M\times M}$ function and the scalar (random) function

$$\hat{h}_j^{(l)}(z) = f_j^{(l)}(\hat{\omega}_j\left(z\right))\frac{z}{\hat{\omega}_j\left(z\right)}\hat{\omega}_j'\left(z\right) \tag{3.5}$$

where $\hat{\omega}_j\left(z\right)$ denotes the consistent estimator of $\omega_j\left(z\right)$ given by

$$\hat{\omega}_j\left(z\right) = \frac{z}{1 - \frac{1}{N_j}\text{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]} \tag{3.6}$$

[2]Note that in the oversampled regime ($\mu_{\text{inf}}^{(j)} = 0$) the function $f_j^{(l)}(\omega)$ is not required to be analytic at the origin, which is the case in some important distance metrics such as the KL metric (for which $f_j^{(l)}(\omega) = \omega^{-1}$) or the log-Euclidean metric (for which $f_j^{(l)}(\omega) = \log\omega$).

and $\hat{\omega}_j'(z)$ *represents its derivative, namely*

$$\hat{\omega}_j'(z) = \frac{1 - \frac{M}{N_j} + z^2 \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{Q}}_j^2(z)\right]}{\left(1 - \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j \hat{\mathbf{Q}}_j(z)\right]\right)^2}. \tag{3.7}$$

Furthermore, we have $\sup_{M \geq M_0} \|\hat{h}_j^{(l)}(\hat{\mathbf{R}}_j)\| < +\infty$ *almost surely for sufficiently large* M_0.

Proof. See Appendix B.1. □

Proposition 3.1 provides in (3.4) the basic piece that can be used to build consistent estimators of the general distances in the form of (1.1). Indeed, it is a direct consequence of Proposition 3.1 that $d_M - \tilde{d}_M \to 0$ almost surely, where

$$\begin{aligned}
\tilde{d}_M &= \sum_{l=1}^{L} \frac{1}{M}\mathrm{tr}\left[\hat{h}_1^{(l)}\left(\hat{\mathbf{R}}_1\right)\hat{h}_2^{(l)}\left(\hat{\mathbf{R}}_2\right)\right] \\
&= \frac{-1}{4\pi^2}\oint_{\mathrm{C}^-}\oint_{\mathrm{C}^-}\sum_{l=1}^{L}\hat{h}_1^{(l)}(z_1)\hat{h}_2^{(l)}(z_2)\frac{1}{M}\mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2))\right]dz_1dz_2 \\
&= \frac{-1}{4\pi^2}\oint_{\mathrm{C}^-}\oint_{\mathrm{C}^-}\left(\sum_{l=1}^{L} f_1^{(l)}(\hat{\omega}_1(z_1))f_2^{(l)}(\hat{\omega}_2(z_2))\right)\times \\
&\qquad \times \frac{1}{M}\mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2))\right]\frac{z_1 z_2 \hat{\omega}_1'(z_1)\hat{\omega}_2'(z_2)}{\hat{\omega}_1(z_1)\hat{\omega}_2(z_2)}dz_1dz_2.
\end{aligned} \tag{3.8}$$

Hence, $\tilde{d}_M$ provides a general expression for a consistent estimator of a distance between covariance matrices taking the form in (1.1). Interestingly enough, this general expression can be particularized to well-known distances commonly found in the literature, exhibiting straightforward closed-form analytical representations. These specific cases will be examined in more detail in the subsequent subsections. We begin by considering the conventional Euclidean and symmetrized KL distances. In these two cases, the solution to (3.8) will be directly derived from conventional residual calculus procedures. Obtaining the consistent estimator for the log-Euclidean distance necessitates additional efforts, and its derivation will be presented towards the conclusion of this section.

Remark 3.2. *At this point, it is also interesting to contextualize the above results in relation to prior works that have directly proposed consistent estimators for quantifying the distance between covariance matrices. Let us start by pointing out that several of the contributions in this domain, e.g. [52–55] to name a few, are primarily designed for distances between covariances expressed as functions of the product* $\hat{\mathbf{R}}_1^{-1}\hat{\mathbf{R}}_2$. *It is also worth noticing*

that, in many of this works, $\hat{\mathbf{R}}_1$ is considered full rank, which imposes strong restrictions when considering the undersampled regime.

More specifically, in [52,54], the authors strongly rely on the results of [57] to associate the eigenvalue distributions of $\hat{\mathbf{R}}_1^{-1}\hat{\mathbf{R}}_2$ and $\mathbf{R}_1^{-1}\mathbf{R}_2$. To alleviate the restrictions in $c_1 < 1$ the authors also consider the case where the covariance matrix $\mathbf{R}_1$ is known and inevitable. This, however, might not be a realistic assumption in many real-world applications. More recently, in [60], the same authors expanded upon their initial findings to consistently estimate distances between covariance matrices in the undersampled regime. However, in practice, the consistency is only achieved for polynomial functions. Finally, it is important to note that there exist important distances such as the log-Euclidean metric that do not fall into this category. In this context, the results presented in this section and later in this chapter differs from previous works already presented in the literature in two main aspects: (i) by considering both the undersampled and oversampled regime; and (ii) by also introducing a new general CLT on the consistent estimators (see Section 3.2).

3.1.1 Estimation of the Euclidean distance

We have seen in the previous chapter that the Euclidean distance takes the form in (1.1) with

$$\sum_{l=1}^{L} f_1^{(l)}(\omega_1) f_2^{(l)}(\omega_2) = (\omega_1 - \omega_2)^2 = \omega_1^2 + \omega_2^2 - 2\omega_1\omega_2$$

so that we have to solve the integral in (3.8) for $f_j^{(1)}(\omega_j) = \omega_j$ and $f_j^{(2)}(\omega_j) = \omega_j^2$. We start by noticing that the function

$$\hat{\mathbf{Q}}_j(z) z \hat{\omega}_j'(z) = z\hat{\mathbf{Q}}_j(z) \frac{1 - \frac{M}{N_j} + z^2 \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{Q}}_j^2(z)\right]}{\left(1 - \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right)^2}$$

presents all its poles inside the contour C. Indeed, by definition C encloses all the eigenvalues of $\hat{\mathbf{R}}_j$ almost surely for all large M (see [58]), as well as the solutions to the equation $1 = \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]$ (see [61]). Furthermore, in the undersampled case the function presents a removable singularity at zero, so that effectively all the singularities are located inside the contour C. We can therefore enlarge the contour C without changing the value of the integral, and then apply the change of variables $z \mapsto \zeta = z^{-1}$, which only has a single singularity at $\zeta = 0$. This can be

done to show that, when $f_j^{(l)}(\omega) = \omega$, we have

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} f_j^{(l)}(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \hat{\mathbf{Q}}_j(z) dz = \\ = \frac{1}{2\pi \mathrm{j}} \oint_{C_0^-} \frac{1 - \frac{M}{N_j} + \frac{1}{\zeta^2}\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{Q}}_j^2(\zeta^{-1})\right]}{\left(1 - \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j \hat{\mathbf{Q}}_j(\zeta^{-1})\right]\right)^2} \hat{\mathbf{Q}}_j(\zeta^{-1}) \frac{1}{\zeta^3} d\zeta$$

where C_0^- encloses only $\zeta = 0$ and no other singularity. Noting that

$$\lim_{\zeta \to 0} \zeta^{-1} \hat{\mathbf{Q}}_j(\zeta^{-1}) = -\mathbf{I}_M$$

we see that the above integrand presents a second order pole at $\zeta = 0$ and therefore one easily computes (for the case $f_j^{(l)}(\omega) = \omega$)

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} f_j^{(l)}(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \hat{\mathbf{Q}}_j(z) dz = \hat{\mathbf{R}}_j.$$

Proceeding in exactly the same way for $f_j^{(l)}(\omega) = \omega^2$ we see that the pole at $\zeta = 0$ has now order three, and therefore (after some algebra)

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} f_j^{(l)}(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \hat{\mathbf{Q}}_j(z) dz = \frac{1}{2\pi \mathrm{j}} \oint_{C_0^-} \frac{1 - \frac{M}{N_j} + \frac{1}{\zeta^2}\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{Q}}_j^2(\zeta^{-1})\right]}{\left(1 - \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j \hat{\mathbf{Q}}_j(\zeta^{-1})\right]\right)^3} \hat{\mathbf{Q}}_j(\zeta^{-1}) \frac{1}{\zeta^4} d\zeta \\ = \hat{\mathbf{R}}_j^2 - \left(\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\right]\right) \hat{\mathbf{R}}_j.$$

As a consequence of the above two integrals, we can conclude that the estimator in (3.8) particularizes to the Euclidean distance as

$$\tilde{d}_M^E = \frac{1}{M}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2\right)^2\right] - \frac{1}{MN_1}\mathrm{tr}^2\left[\hat{\mathbf{R}}_j\right] - \frac{1}{MN_2}\mathrm{tr}^2\left[\hat{\mathbf{R}}_2\right]$$

which corresponds to the conventional estimator corrected by the square of the normalized trace of the two sample covariance matrices. Obviously, the estimator becomes the conventional *plug-in* one if N_1, N_2 increase but M remains fixed.

3.1.2 Estimation of the symmetrized KL distance

We recall that the symmetrized KL distance corresponds to the definition in (1.1) with

$$\sum_{l=1}^{L} f_1^{(l)}(\omega_1) f_2^{(l)}(\omega_2) = \frac{1}{2}\frac{\omega_2}{\omega_1} + \frac{1}{2}\frac{\omega_1}{\omega_2} - 1.$$

Note that in this case the function ω^{-1} is not holomorphic at the origin, which implies that we can only tolerate $\mu_{\text{inf}} = 0$ in **(As5)**. In particular, this implies that we can only obtain a consistent estimator for the oversampled case (namely $N_1 > M$ and $N_2 > M$).

Here again, we need to solve the different integrals in (3.8) for the functions $f_1^{(l)}(\omega) = \omega$ (already done in the previous section), $f_j^{(l)}(\omega) = 1$ and $f_j^{(l)}(\omega) = \omega^{-1}$. Let us first consider the simpler case where $f_j^{(l)}(\omega) = 1$. To solve the corresponding integral, we can simply notice that the function

$$\frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\hat{\mathbf{Q}}_j(z) = \hat{\mathbf{Q}}_j(z)\frac{1 - \frac{M}{N_j} + z^2\frac{1}{N_j}\text{tr}\left[\hat{\mathbf{Q}}_j^2(z)\right]}{1 - \frac{1}{N_j}\text{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]}$$

presents all its poles inside the contour C (note that there is no singularity at zero because, by definition (3.3), $\hat{\mathbf{Q}}_j(0) = \hat{\mathbf{R}}_j^{-1}$ which is well-defined since we are in the oversampled situation so that $\hat{\mathbf{R}}_j$ is invertible with probability one). Hence, we can enlarge the contour C as much as we want and consider again the change of variable $\zeta = z^{-1}$, after which the integrand will only have a singularity at $\zeta = 0$. Consequently, we can find

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}^-}\frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\hat{\mathbf{Q}}_j(z)dz = \frac{1}{2\pi\mathrm{j}}\oint_{C_0^-}\hat{\mathbf{Q}}_j(\zeta^{-1})\frac{1 - \frac{M}{N_j} + \zeta^{-2}\frac{1}{N_j}\text{tr}\left[\hat{\mathbf{Q}}_j^2(\zeta^{-1})\right]}{1 - \frac{1}{N_j}\text{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(\zeta^{-1})\right]}\frac{d\zeta}{\zeta^2} = \mathbf{I}_M.$$

Regarding the integral (3.8) for $f_j^{(l)}(\omega) = \omega^{-1}$, we consider the function

$$\frac{\hat{\mathbf{Q}}_j(z)z\hat{\omega}_j'(z)}{\hat{\omega}_j^2(z)} = \frac{\hat{\mathbf{Q}}_j(z)}{z}\left(1 - \frac{M}{N_j} + z^2\frac{1}{N_j}\text{tr}\left[\hat{\mathbf{Q}}_j^2(z)\right]\right) \tag{3.9}$$

and observe again that all its singularities are inside the contour C except for a simple pole at $z = 0$. Therefore, we can deform the contour C into a larger one $\mathcal{C}$ (see Remark 2.1) that now encloses $\mathcal{S}_j \cup \{0\}$ and write (for the case $f_j^{(l)}(\omega) = \omega^{-1}$)

$$\begin{aligned}\frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}^-} f_j^{(l)}(\hat{\omega}_j(z))\frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\hat{\mathbf{Q}}_j(z)dz &= \left(1 - \frac{M}{N_j}\right)\hat{\mathbf{R}}_j^{-1} + \\ &+ \frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}^-}\frac{\hat{\mathbf{Q}}_j(z)}{z}\left(1 - \frac{M}{N_j} + z^2\frac{1}{N_j}\text{tr}\left[\hat{\mathbf{Q}}_j^2(z)\right]\right)dz\end{aligned}$$

where the first term is the residue of (3.9) at zero. We can now see that the second integral is zero by enlarging the contour and applying the change of variable $\zeta =$

z^{-1}, after which the corresponding integrand becomes analytic at zero. We can therefore conclude that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} f_j^{(l)}(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \hat{\mathbf{Q}}_j(z) dz = \left(1 - \frac{M}{N_j}\right) \hat{\mathbf{R}}_j^{-1}.$$

With this, we have now all the ingredients to evaluate the integral at (3.8), which provides a consistent estimator for the symmetrized KL distance between covariance matrices, namely

$$\tilde{d}_M^{KL} = \left(1 - \frac{M}{N_1}\right) \frac{1}{2M} \mathrm{tr}\left[\hat{\mathbf{R}}_1^{-1} \hat{\mathbf{R}}_2\right] + \left(1 - \frac{M}{N_2}\right) \frac{1}{2M} \mathrm{tr}\left[\hat{\mathbf{R}}_2^{-1} \hat{\mathbf{R}}_1\right] - 1.$$

3.1.3 Estimation of the Log-Euclidean distance

The log-Euclidean distance takes the form in (1.1) with

$$\sum_{l=1}^{L} f_1^{(l)}(\omega_1) f_2^{(l)}(\omega_2) = (\log \omega_1 - \log \omega_2)^2$$

and therefore to evaluate the integral in (3.8) one must consider the two functions $f_j^{(l)}(\omega) = \log \omega$ and $f_j^{(l)}(\omega) = (\log \omega)^2$. Observe now that these two functions are analytic everywhere except for the negative real axis (including zero). Hence, similarly as in the symmetric KL distance, in **(As5)** we must have $\mu_{\mathrm{inf}}^{(j)} = 0$, implying that $\mu_0^{(j)} = 0$ for all M and hence $N_j > M$ (oversampled regime).

The first integral (with respect to $f_j^{(l)}(\omega) = \log \omega$) was already solved by [62], and we will recall it in the context of this dissertation. In order to present the closed form solution for this integral, let us denote by $\hat{\lambda}_1^{(j)} < \ldots < \hat{\lambda}_M^{(j)}$ and $\hat{\mathbf{e}}_1^{(j)}, \ldots, \hat{\mathbf{e}}_M^{(j)}$ the eigenvalues and associated eigenvectors of the sample covariance matrix $\hat{\mathbf{R}}_j$. It was shown in [62] that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \hat{\mathbf{Q}}_j(z) dz = \sum_{k=1}^{M} \beta_k^{(j)} \hat{\mathbf{e}}_k^{(j)} \left(\hat{\mathbf{e}}_k^{(j)}\right)^H$$

where the coefficients $\beta_k^{(j)}$, $k = 1, \ldots, M$, are defined as

$$\beta_k^{(j)} = 1 + \left(1 + \sum_{\substack{m=1 \\ m \neq k}}^{M} \frac{\hat{\lambda}_k^{(j)}}{\hat{\lambda}_m^{(j)} - \hat{\lambda}_k^{(j)}} - \sum_{m=1}^{M} \frac{\hat{\mu}_k^{(j)}}{\hat{\lambda}_m^{(j)} - \hat{\mu}_k^{(j)}}\right) \log \hat{\lambda}_k^{(j)}$$
$$+ \left(\sum_{\substack{r=1 \\ r \neq k}}^{M} \frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_r^{(j)} - \hat{\lambda}_k^{(j)}} \log \hat{\lambda}_r^{(j)} - \sum_{r=1}^{M} \frac{\hat{\mu}_r^{(j)}}{\hat{\mu}_r^{(j)} - \hat{\lambda}_k^{(j)}} \log \hat{\mu}_r^{(j)}\right)$$

where we have denoted by $\hat{\mu}_0^{(j)} < \ldots < \hat{\mu}_M^{(j)}$ the M solutions to the polynomial equation in (3.3) by interchanging the covariance matrix $\mathbf{R}_j$ by its estimator $\hat{\mathbf{R}}_j$.

Let us denote by $\alpha^{(j)}$ the integral of the term in $f_j^{(l)}(\omega) = \log^2(\hat{\omega}_j(z))$, that is

$$\alpha^{(j)} = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log^2(\hat{\omega}_j(z)) \frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)} \frac{1}{M} \mathrm{tr}\left[\hat{\mathbf{Q}}_j(z)\right] dz \tag{3.10}$$

Using very similar arguments as above, it is shown in Appendix B.2 that $\alpha^{(j)}$ takes the form

$$\begin{aligned}
\alpha^{(j)} = & \left(\frac{N_j}{M} - 1\right) \sum_{r=1}^{M} \left(1 + \log \hat{\mu}_r^{(j)}\right)^2 - \left(1 + \log \hat{\lambda}_k^{(j)}\right)^2 \\
& + \frac{1}{M} \sum_{k=1}^{M} \left(1 + \log \hat{\lambda}_k^{(j)}\right)^2 - \left(\frac{N_j}{M} - 1\right) \log^2\left(1 - \frac{M}{N_j}\right) \\
& + 1 + \frac{2}{M} \sum_{k=1}^{M} \sum_{r=1}^{M} \left[\Phi_2\left(\frac{\hat{\mu}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\right) - \Phi_2\left(\frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\right)\right] \\
& + \frac{2}{M} \sum_{k=1}^{M} \left(\sum_{\substack{r=1 \\ r \neq k}}^{M} \log \frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_k^{(j)}} \log \frac{\hat{\lambda}_k^{(j)}}{\left|\hat{\lambda}_k^{(j)} - \hat{\lambda}_r^{(j)}\right|} \right. \\
& \left. - \sum_{r=1}^{M} \log \frac{\hat{\mu}_r^{(j)}}{\hat{\lambda}_k^{(j)}} \log \frac{\hat{\lambda}_k^{(j)}}{\left|\hat{\lambda}_k^{(j)} - \hat{\mu}_r^{(j)}\right|} \right)
\end{aligned} \tag{3.11}$$

where we have defined

$$\Phi_2(x) = \begin{cases} \mathrm{Li}_2(x) & x < 1 \\ \frac{\pi^2}{3} - \frac{1}{2}\log^2 x - \mathrm{Li}_2(x^{-1}) & x \geq 1 \end{cases} \tag{3.12}$$

and where $\mathrm{Li}_2(x) = -\int_0^x y^{-1}\log(1-y)dy$ is the dilogarithm function.

Combining the two expressions above, we obtain a closed form expression for the consistent estimator of the log-Euclidean distance between covariance matrices, namely

$$\tilde{d}_M^{LE} = \alpha^{(1)} + \alpha^{(2)} - \frac{2}{M} \sum_{k=1}^{M} \sum_{m=1}^{M} \beta_k^{(1)} \beta_m^{(2)} \left| \left(\hat{\mathbf{e}}_k^{(1)}\right)^H \hat{\mathbf{e}}_m^{(2)} \right|^2$$

where $\alpha^{(j)}$ and $\beta_k^{(j)}$, $j = 1, \ldots, M$ are defined as above.

3.2 A general CLT of Consistent Estimators

In this section, we follow a similar approach as the one conducted in the previous chapter to study the fluctuations of $\tilde{d}_M$ around the true values d_M. To that effect,

we will basically show that the random variable $M(\tilde{d}_M - d_M)$ asymptotically behaves as a Gaussian distribution with a certain mean and variance that will be characterized below in its most generic form and particularized to the case of the Euclidean, symmetrized KL and log-Euclidean distances.

Let us start by recalling, from (3.8), that the estimators above can all be expressed as

$$\tilde{d}_M = \frac{1}{(2\pi j)^2} \oint_{\mathrm{C}^-} \oint_{\mathrm{C}^-} \sum_{l=1}^{L} \hat{h}_1^{(l)}(z_1)\hat{h}_2^{(l)}(z_2) \frac{1}{M}\mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2))\right] dz_1 dz_2$$

and that, from the definitions (1.1) and (3.2), the distance between covariance matrices can be re-written as

$$\begin{aligned} d_M &= \frac{1}{(2\pi j)^2} \oint_{\mathrm{C}^-} \oint_{\mathrm{C}^-} \sum_{l=1}^{L} f_1^{(l)}\left(\omega_1(z_1)\right) \omega_1'(z_1) f_2^{(l)}\left(\omega_2(z_2)\right) \omega_2'(z_2) \times \\ &\qquad \times \frac{1}{M}\mathrm{tr}\left[\mathbf{Q}_1(z_1)\mathbf{Q}_2(z_2))\right] dz_1 dz_2 \\ &= \frac{1}{(2\pi j)^2} \oint_{\mathrm{C}^-} \oint_{\mathrm{C}^-} \sum_{l=1}^{L} h_1^{(l)}\left(z_1\right) h_2^{(l)}\left(z_2\right) \frac{1}{M}\mathrm{tr}\left[\bar{\mathbf{Q}}_1(z_1)\bar{\mathbf{Q}}_2(z_2))\right] dz_1 dz_2 \end{aligned}$$

where in the last step we have used the definition

$$h_j^{(l)}(z) = f_j^{(l)}(\omega_j\left(z\right)) \frac{z}{\omega_j\left(z\right)} \omega_j'\left(z\right)$$

to represent the asymptotic equivalent of the random quantity $\hat{h}_j^{(l)}(z)$.

Then, similar to Section 2.3, we have that

$$\begin{aligned} \tilde{\zeta}_M = M\left(\tilde{d}_M - d_M\right) = \frac{1}{(2\pi \mathrm{j})^2} \oint_{C_1^-} \oint_{C_2^-} \tilde{g}_M(z_1, z_2) \times \\ \times \mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2) - \bar{\mathbf{Q}}_1(z_1)\bar{\mathbf{Q}}_2(z_2)\right] dz_1 dz_2 \end{aligned} \tag{3.13}$$

with the only difference in the integrand being that now we have

$$\tilde{g}_M(z_1, z_2) = \left(\sum_{l=1}^{L} h_1^{(l)}\left(z_1\right) h_2^{(l)}\left(z_2\right)\right) - \tilde{g}^{(1)}(z_1) - \tilde{g}^{(2)}(z_2) \tag{3.14}$$

where, for $j, k = 1, 2$ and $j \neq k$, we have defined

$$\tilde{g}^{(j)}(z_j) = \sum_{l=1}^{L} f_j^{(l)}(\omega_j)\phi\left(\omega_j; f_k^{(l)}(\mathbf{R}_k)\right) \tag{3.15}$$

using the definition of $\phi\left(\omega; \mathbf{A}\right)$ in (2.16).

Theorem 3.1. *In addition to* ***(As1)-(As3)*** *and* ***(As5)****, assume that the observations are Gaussian distributed. If* $\liminf_{M\to\infty}\sigma_M^2 > 0$*, we have*

$$\frac{M(\tilde{d}_M - d_M) - \tilde{\mathfrak{m}}_M}{\tilde{\sigma}_M} \to \mathcal{N}(0,1).$$

where we have now defined

$$\begin{aligned}\tilde{\mathfrak{m}}_M = &\sum_{l=1}^{L}\frac{\varsigma}{2\pi\mathrm{j}}\oint_{\mathrm{C}^-_{\omega_1}} f_1^{(l)}(\omega_1)\frac{\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3(\omega_1)\,f_2^{(l)}(\mathbf{R}_2)\right]}{N_1(1-\Gamma_1(\omega_1))}d\omega_1\\ &+\sum_{l=1}^{L}\frac{\varsigma}{2\pi\mathrm{j}}\oint_{\mathrm{C}^-_{\omega_2}} f_2^{(l)}(\omega_2)\frac{\mathrm{tr}\left[\mathbf{R}_2^2\mathbf{Q}_2^3(\omega_2)\,f_1^{(l)}(\mathbf{R}_1)\right]}{N_2(1-\Gamma_2(\omega_2))}d\omega_2\end{aligned} \tag{3.16}$$

and where we have also introduced the function (for $j\in\{1,2\}$*)*

$$\Gamma_j(\omega) = \frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j^2(\omega)\right]. \tag{3.17}$$

Likewise, we have also defined (denoting again $\omega_j=\omega_j(z_j)$ *and* $\tilde{\omega}_j=\omega_j(\tilde{z}_j)$*)*

$$\begin{aligned}\tilde{\sigma}_M^2 = \frac{1+\varsigma}{(2\pi\mathrm{j})^4}\oint_{\mathrm{C}^-_{\omega_1}}\oint_{\mathrm{C}^-_{\omega_1}}\oint_{\mathrm{C}^-_{\omega_2}}\oint_{\mathrm{C}^-_{\omega_2}} f(\omega_1,\omega_2)f(\tilde{\omega}_1,\tilde{\omega}_2)\times\\ \times\,\overline{\Sigma}^2(\omega_1,\omega_2,\tilde{\omega}_1,\tilde{\omega}_2)\,d\omega_1 d\omega_2 d\tilde{\omega}_1 d\tilde{\omega}_2\end{aligned} \tag{3.18}$$

where the last term consists of three terms

$$\begin{aligned}\overline{\Sigma}^2(\omega_1,\omega_2,\tilde{\omega}_1,\tilde{\omega}_2) = &\,\overline{\sigma}_1^2(\omega_1,\tilde{\omega}_1;\mathbf{Q}_2(\omega_2),\mathbf{Q}_2(\tilde{\omega}_2))\\ &+\overline{\sigma}_2^2(\omega_2,\tilde{\omega}_2;\mathbf{Q}_1(\omega_1),\mathbf{Q}_1(\tilde{\omega}_1))\\ &+\varrho(\omega_1,\tilde{\omega}_1,\omega_2,\tilde{\omega}_2)\end{aligned} \tag{3.19}$$

with the following definitions. The first two terms are specific instances of the function

$$\begin{aligned}\overline{\sigma}_j^2(\omega,\tilde{\omega};\mathbf{A},\mathbf{B}) = &\frac{\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega)\,\mathbf{Q}_j(\tilde{\omega})\,\mathbf{A}\mathbf{R}_j\mathbf{Q}_j(\omega)\,\mathbf{Q}_j(\tilde{\omega})\,\mathbf{B}\right]}{N_j(1-\Gamma_j(\omega,\tilde{\omega}))}\\ &+\frac{\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j^2(\omega)\,\mathbf{Q}_j(\tilde{\omega})\,\mathbf{A}\right]\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j(\omega)\,\mathbf{Q}_j^2(\tilde{\omega})\,\mathbf{B}\right]}{N_j^2\left(1-\Gamma_j(\omega,\tilde{\omega})\right)^2}\end{aligned} \tag{3.20}$$

where we recall the bivariate function from (2.10) defined as

$$\Gamma_j(\omega,\tilde{\omega}) = \frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j(\omega)\,\mathbf{Q}_j(\tilde{\omega})\right].$$

Finally, the third term in (3.19) takes the form

$$\varrho(\omega_1,\tilde{\omega}_1,\omega_2,\tilde{\omega}_2) = \frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1(\omega_1)\,\mathbf{Q}_1(\tilde{\omega}_1)\,\mathbf{R}_2\mathbf{Q}_2(\omega_2)\,\mathbf{Q}_2(\tilde{\omega}_2)\right]}{N_1N_2\left(1-\Gamma_1(\omega_1,\tilde{\omega}_1)\right)\left(1-\Gamma_2(\omega_2,\tilde{\omega}_2)\right)}. \tag{3.21}$$

Proof. See Appendix B.3. □

3.3 Simplified Expressions in the Oversampled Regime

Even if the expression obtained above appears to be difficult to evaluate due to the presence of the contour integrals, one can typically simplify these expressions using conventional residue calculus. This is illustrated next for the three distances that have been considered in this section, namely the Euclidean distance, the symmetrized Kullback Leibler divergence and the log-Euclidean distance.

3.3.1 Particularization to the Euclidean distance

For the conventional Euclidean norm we have $f(\omega_1, \omega_2) = (\omega_1 - \omega_2)^2$ and both the integrands in (3.16) and (3.18) have all the singularities inside the corresponding contours. The strategy to solve these integrals is therefore to apply the change of variable $\omega_j \mapsto \zeta_j = \omega_j^{-1}$ after enlarging the contours, so that the resulting integrands after the change of variable have only a singularity at $\zeta_j = 0$. Using this technique, one can show that the asymptotic (second order) mean takes the form

$$\tilde{\mathfrak{m}}_M^E = \varsigma \left(\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1^2\right] + \frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right] \right).$$

Regarding the asymptotic variance, one can use exactly the same integration technique to show that

$$\begin{aligned} \frac{\tilde{\sigma}_M^2}{1+\varsigma} = & \, 2\left(\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1^2\right] \right)^2 + 4\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1 \boldsymbol{\Delta} \mathbf{R}_1 \boldsymbol{\Delta}\right] \\ & + 2\left(\frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right] \right)^2 + 4\frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2 \boldsymbol{\Delta} \mathbf{R}_2 \boldsymbol{\Delta}\right] \\ & + 4\frac{1}{N_1 N_2} \mathrm{tr}^2\left[\mathbf{R}_1 \mathbf{R}_2\right] \end{aligned}$$

where, now, $\boldsymbol{\Delta} = \mathbf{R}_1 - \mathbf{R}_2$. Obviously, the three terms are positive, so that in order to show that $\liminf_{M\to\infty} \sigma_M^2 > 0$ it is sufficient to see that any of these are bounded away from zero. In particular, using the fact that the eigenvalues of $\mathbf{R}_1$ are located inside a compact of $\mathbb{R}^+$ independent of M one trivially sees that the first term is bounded away from zero.

An interesting consequence from the above is that it becomes fairly easy to obtain the estimators of $\tilde{\mathfrak{m}}_M^E$ and $\tilde{\sigma}_M^E$ directly from the data. We will denote these by $\hat{\tilde{\mathfrak{m}}}_M^E$ and $\hat{\tilde{\sigma}}_M^E$, respectively. We recall from (2.13), that

$$\frac{1}{N_j} \mathrm{tr}[\mathbf{R}_j^2] \asymp \frac{1}{N_j} \mathrm{tr}[\hat{\mathbf{R}}_j^2] - \frac{1}{N_j^2} \mathrm{tr}^2[\hat{\mathbf{R}}_j]. \tag{3.22}$$

Hence, for $\mathbf{R}_1 = \mathbf{R}_2$, these estimators take the form

$$\hat{\tilde{\mathfrak{m}}}_M^E = \varsigma \left(\frac{1}{N_1} \mathrm{tr}\left[\hat{\mathbf{R}}_1^2\right] - \frac{1}{N_1^2} \mathrm{tr}^2\left[\hat{\mathbf{R}}_1\right] + \frac{1}{N_2} \mathrm{tr}\left[\hat{\mathbf{R}}_2^2\right] - \frac{1}{N_2^2} \mathrm{tr}^2\left[\hat{\mathbf{R}}_2\right] \right). \tag{3.23}$$

and

$$\begin{aligned} \frac{\left(\hat{\tilde{\sigma}}_M^E\right)^2}{1+\varsigma} &= 2\left(\frac{1}{N_1} \mathrm{tr}\left[\hat{\mathbf{R}}_1^2\right] - \frac{1}{N_1^2} \mathrm{tr}^2\left[\hat{\mathbf{R}}_1\right] \right)^2 \\ &\quad + 2\left(\frac{1}{N_2} \mathrm{tr}\left[\hat{\mathbf{R}}_2^2\right] - \frac{1}{N_2^2} \mathrm{tr}^2\left[\hat{\mathbf{R}}_2\right] \right)^2 \\ &\quad + 4\frac{1}{N_1 N_2} \mathrm{tr}^2\left[\hat{\mathbf{R}}_1 \hat{\mathbf{R}}_2\right]. \end{aligned} \tag{3.24}$$

3.3.2 Particularization to the symmetrized KL divergence

For the symmetrized KL divergence we need to particularize the above expressions to the case

$$f(\omega_1, \omega_2) = \frac{1}{2}\left(\frac{\omega_1}{\omega_2} + \frac{\omega_2}{\omega_1} \right) - 1.$$

We recall that the estimator for this particular distance has only been defined in the oversampled case ($N_j > M$), and in this case the contour C_{ω_j} does not enclose $\{0\}$ [56]. Hence, the integrands in (3.16) and (3.18) have all the singularities inside the contour, except for a potential singularity at zero. The integration strategy therefore consists in enlarging the contour C_{ω_j} so that it also encloses this singularity, compensating the result by adding the corresponding residue at zero (note that the original contour is always negatively oriented). The resulting integral can therefore be solved applying the change of variables $\omega_j \mapsto \zeta_j = \omega_j^{-1}$.

Using this integration technique, one can easily show that the asymptotic (second order) mean takes the form

$$\tilde{\mathfrak{m}}_M^{KL} = \frac{\varsigma}{2}\left(\frac{1}{N_1 - M} \mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\right] + \frac{1}{N_2 - M} \mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right] \right).$$

whereas the asymptotic variance can be expressed as

$$\begin{aligned} \frac{\tilde{\sigma}_M^2}{(1+\varsigma)(N_1+N_2-M)} &= -\frac{M}{2N_1N_2} + \frac{\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{R}_1\mathbf{R}_2^{-1}\right]}{4\left(N_2-M\right)N_1} + \frac{\mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\mathbf{R}_1^{-1}\mathbf{R}_2\right]}{4\left(N_1-M\right)N_2} \\ &\quad + \frac{1}{4N_1}\left(\frac{\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right]}{N_2-M} \right)^2 + \frac{1}{4N_2}\left(\frac{\mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\right]}{N_1-M} \right)^2. \end{aligned}$$

One can easily show that this is positive by applying the inequality $\alpha \mathrm{tr}\left[\mathbf{A}\right] + \beta \mathrm{tr}\left[\mathbf{A}^{-1}\right] \geq 2M\sqrt{\alpha\beta}$ (valid for any positive $M \times M$ matrix $\mathbf{A}$). Indeed, using this inequality we see that the sum of the first three terms of the above expression is non-negative. The fact that the other two terms are bounded away from zero follows easily from the fact that the eigenvalues of the covariance matrices are located in a compact interval of $\mathbb{R}^+$ independent of M. This shows that $\liminf_{M\to\infty} \sigma_M^2 > 0$ and the CLT holds.

Finally, when we have equal covariance matrices $\mathbf{R}_1^{-1}\mathbf{R}_2 = \mathbf{I}_M$, the asymptotic estimators of these quantities become

$$\hat{\mathfrak{m}}_M^{KL} = \frac{\varsigma}{2}\left(\frac{M}{N_1 - M} + \frac{M}{N_2 - M}\right) \tag{3.25}$$

and

$$\begin{aligned}\frac{\left(\hat{\tilde{\sigma}}_M^{KL}\right)^2}{(1+\varsigma)(N_1+N_2-M)} &= \frac{1}{4N_1}\left(\frac{M}{N_2-M}\right)\left(1+\frac{M}{N_2-M}\right) \\ &\quad + \frac{1}{4N_2}\left(\frac{M}{N_1-M}\right)\left(1+\frac{M}{N_1-M}\right) - \frac{M}{2N_1N_2}\end{aligned}$$

which are both independent of $\mathbf{R}_1$ and $\mathbf{R}_2$. These results become particularly useful in practical scenarios where one does not need to have access to $\mathbf{R}_1$ nor $\mathbf{R}_2$.

3.3.3 Particularization to the Log-Euclidean distance

Finally, the log-Euclidean distance we have $f(\omega_1, \omega_2) = (\log\omega_1 - \log\omega_2)^2$ and the above integral tricks are not useful anymore due to the fact that the integrands are not holomorphic on the whole negative axis. We can still find the solution to the asymptotic (second-order) mean by evaluating the residues at the poles given by the eigenvalues of $\mathbf{R}_j$ and the solutions to $\Gamma_j(\omega_j) = 1$, which are denoted $\theta_m^{(j)}$, $m = 1, \ldots, 2\bar{M}_j$. One can readily see that (assuming that the roots $\theta_m^{(j)}$ are of multiplicity

one) the asymptotic (second order) mean takes the form

$$\begin{aligned}\tilde{\mathfrak{m}}_M^{LE} = &-\varsigma\sum_{m=1}^{\bar{M}_1}\frac{1}{K_m^{(1)}}\mathrm{tr}\left[\mathbf{E}_m^{(1)}(\mathbf{E}_m^{(1)})^H\left(\log\gamma_m^{(1)}\mathbf{I}_M-\log\mathbf{R}_2\right)^2\right]\\ &-\varsigma\sum_{m=1}^{\bar{M}_2}\frac{1}{K_m^{(2)}}\mathrm{tr}\left[\mathbf{E}_m^{(2)}(\mathbf{E}_m^{(2)})^H\left(\log\mathbf{R}_1-\log\gamma_m^{(2)}\mathbf{I}_M\right)^2\right]\\ &+\frac{\varsigma}{2}\sum_{m=1}^{2\bar{M}_1}\frac{\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3(\theta_m^{(1)})\left(\log\theta_m^{(1)}\mathbf{I}_M-\log\mathbf{R}_2\right)^2\right]}{\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3(\theta_m^{(1)})\right]}\\ &+\frac{\varsigma}{2}\sum_{m=1}^{2\bar{M}_2}\frac{\mathrm{tr}\left[\mathbf{R}_2^2\mathbf{Q}_2^3(\theta_m^{(2)})\left(\log\mathbf{R}_1-\log\theta_m^{(2)}\mathbf{I}_M\right)^2\right]}{\mathrm{tr}\left[\mathbf{R}_2^2\mathbf{Q}_2^3(\theta_m^{(2)})\right]}\end{aligned}$$

where $\mathbf{E}_m^{(j)}$ is an $M\times K_m^{(j)}$ matrix that contains the eigenvectors associated with the eigenvalue $\gamma_m^{(j)}$, assumed to have multiplicity $K_m^{(j)}$. In the particular case where $\mathbf{R}_1=\mathbf{R}_2=\mathbf{R}$ this simplifies to

$$\tilde{\mathfrak{m}}_M^{LE}=\varsigma\sum_{m=1}^{2\bar{M}}\frac{\mathrm{tr}\left[\mathbf{R}^2\mathbf{Q}^3\left(\theta_m\right)\left(\log\theta_m\mathbf{I}_M-\log\mathbf{R}\right)^2\right]}{\mathrm{tr}\left[\mathbf{R}^2\mathbf{Q}^3\left(\theta_m\right)\right]}.$$

We have not been able to come up with a manageable expression for asymptotic variance. We therefore will only evaluate it using numerical integration.

3.4 Numerical Consistency of the Estimators

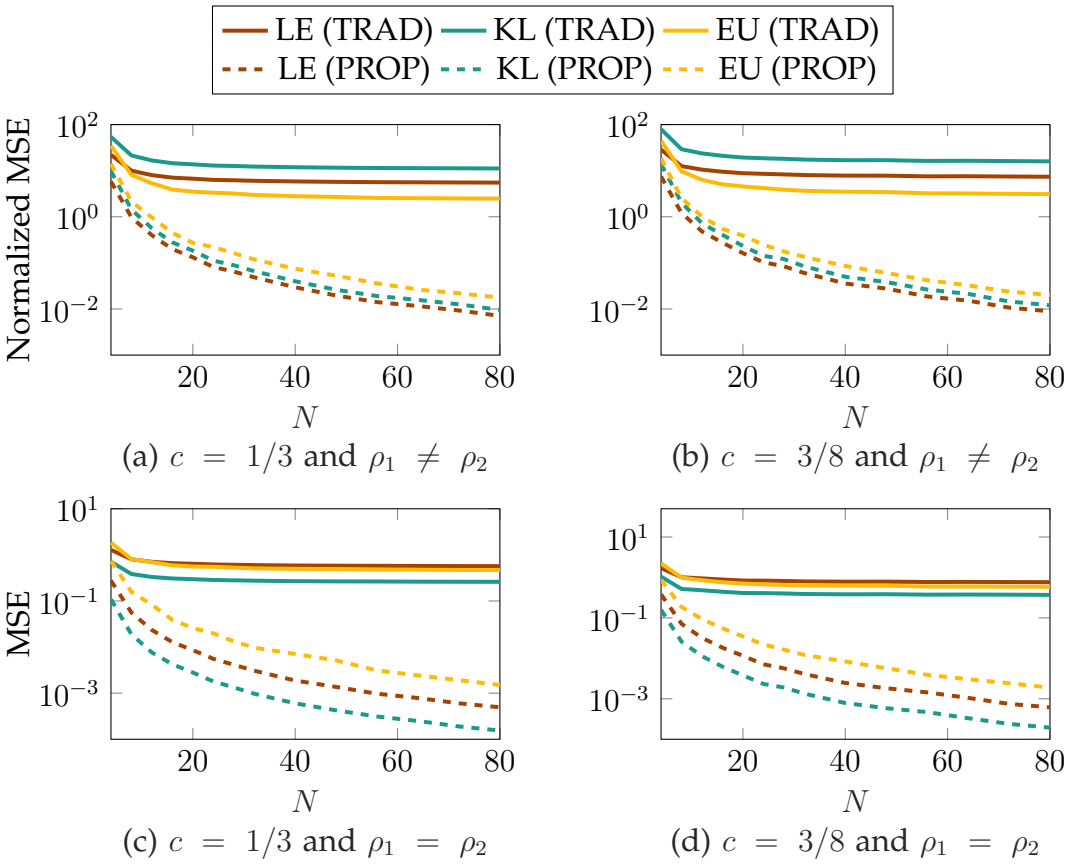
We start by comparing the traditional *plug-in* estimators $\hat{d}_M$ to the proposed consistent estimators $\tilde{d}_M$. Specifically, we are interested in the particularizations already discussed throughout this work, namely, Euclidean distance (EU), Symmetrized Kullback-Leibler (KL) and the log-Euclidean norm (LE) distances. We recall that both the traditional and our proposed methods rely on the information available in the samples' covariance matrices $\hat{\mathbf{R}}_1$, $\hat{\mathbf{R}}_2$ and try to approximate the true distance between the covariance matrices $\mathbf{R}_1$, $\mathbf{R}_2$. Here we use the same definition of the previous chapters where two multidimensional observation sets $\mathbf{Y}_1\in\mathbb{C}^{M\times N_1}$ and $\mathbf{Y}_2\in\mathbb{C}^{M\times N_2}$ are associated to two (possibly distinct) Toeplitz covariance matrices $\mathbf{R}_1$ and $\mathbf{R}_2$ with first rows $[\rho_j^0,\ldots,\rho_j^{M-1}]$, for $j\in\{1,2\}$. Figure 3.1 illustrates the relative Mean Square Error (MSE) of our proposed consistent estimators compared to the true distance over 10^3 samples for different choices of the coefficients ρ_1,ρ_2

and $c = M/N_1 = M/N_2$. For the proposed estimators (dashed lines in the figures) the MSE between these quantities is given by

$$\varepsilon_{\text{PROP}} = \hat{\mathbb{E}}\left[\left(\frac{\tilde{d}_M - d_M}{d_M}\right)^2\right]$$

where the empirical expectation ($\hat{\mathbb{E}}[\cdot]$) is equivalent to empirical averaging. Following the same approach, one can also define $\varepsilon_{\text{TRAD}}$ (solid lines in the figures) by interchanging the proposed estimator $\tilde{d}_M$ with the traditional *plug-in* distance $\hat{d}_M$.

Specifically, Figures 3.1(a)-(b) portrays the case where observations are drawn from distinct processes ($\rho_1 = 0.3$ and $\rho_2 = 0.6$). In this scenario, the traditional estimators fail to converge to the actual distance between the two covariance matrices, whereas the MSE of the proposed estimators continuously decay by growing M, N. The lack of consistency of the conventional estimators is more apparent for the LE



(a) $c = 1/3$ and $\rho_1 \neq \rho_2$ (b) $c = 3/8$ and $\rho_1 \neq \rho_2$

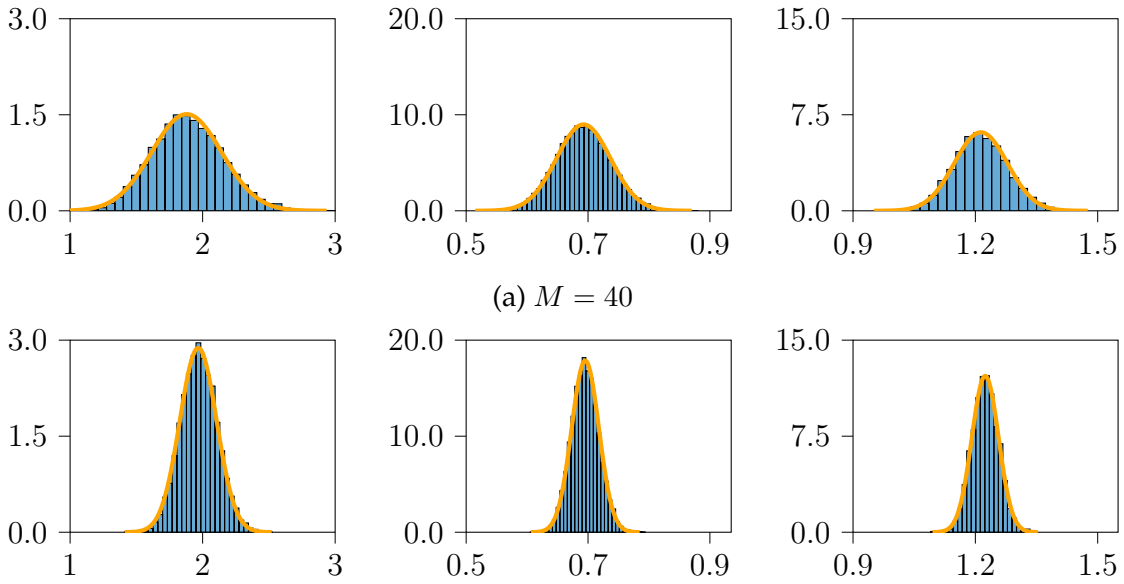
(c) $c = 1/3$ and $\rho_1 = \rho_2$ (d) $c = 3/8$ and $\rho_1 = \rho_2$

Figure 3.1: Relative MSE related to different metrics in different scenarios (a)-(d) with respect to the growth of $N = N_1 = N_2$ (x–axis). In all these curves, the system dimension M is scaled proportionally so that $M/N = c$ is constant.

and the KL norms, although all of them are inconsistent. A similar behavior is displayed in Figures 3.1(c)-(d), for the case where $\rho_1 = \rho_2 = 0.6$. In this case, we have that $d_M = 0$ and hence we consider only the MSE of the distances with zero, i.e., the norm of the empirical distances ($\hat{\mathbb{E}}[\hat{d}_M^2]$ and $\hat{\mathbb{E}}[\tilde{d}_M^2]$). Hence, these quantities should all converge to zero. We notice, however, that all three traditional methods converge to another quantity away from zero while our proposed estimators continuously decay as the system grows. Both these results corroborate the accuracy of the estimators proposed in this work, illustrating the advantage of the new estimators for relatively low values of M, N.

3.4.1 Consistency of Asymptotic Descriptors

We also compare the asymptotic descriptors from the theorems above against the empirical distribution of their respective metrics. Figure 3.2 (in the next page) compares the histogram (in blue) and the asymptotic values (in orange) of these metrics for $c_1 = M/N_1 = 1/10$, $c_2 = M/N_2 = 1/2$ and some specific choices of M and ρ_1, ρ_2. Observe that there is a very good match between the asymptotic and the empirical distributions regardless of the considered metric. This fact, together with the results from the previous section, gives strong indications that for large M, N the random quantity $\tilde{d}_M$ approximates its associated deterministic d_M value



(a) $M = 40$

(b) $M = 80$

Figure 3.2: Histogram of empirical distribution (in blue) and asymptotic descriptors (in orange) of the different metrics EU, KL and LE for fixed $\rho_1 = 0.8, \rho_2 = 0.4$.

better than the traditional *plug-in* method. These results support the idea that our proposed asymptotic descriptors are consistent in several scenarios.

Chapter 4

Applications to Clustering

Up until now, in this dissertation, we have primarily focused on how to estimate and asymptotically describe several distances between covariance matrices. Naturally, this analysis alone already provides valuable results for the research community. Nonetheless, in this chapter we further exploit these results and their applications to unsupervised clustering, namely, to the clustering of random observations according to covariance matrices. In general, clustering solutions compare different pairs of elements and, at each step, decide which pair of elements should be merged together. After the clustering process is over, one often wishes to gain further insights on how good the clustering solution is. In this chapter, we propose studying the quality of a clustering solution based on the pairwise comparisons that generate this solution.

To illustrate this idea in a practical scenario, consider the situation where a clustering solution is being evaluated by an external controller. Typically, to ensure privacy, this controller has no access to the elements that were used to generate the clustering solution. Instead, it is possible that this external controller has access to general information regarding the (several) underlying processes that generate data or, alternatively, to some labeled dataset which describes these processes[1]. Hence, one can also assume the controller to have access to the covariance matrices that generate the processes, which can be obtained from the controller's labeled data; and to the clustering solutions, which can be sent to the controller. In this scenario, the controller wants to evaluate how good this clustering solution is, but has no way to directly compare the covariance matrices to the elements that were

[1]This assumption is not completely unsupported and can be found, for instance, in teacher-student knowledge distillation learning mechanisms [63], where the goal is to transfer the knowledge from a large and complex learning model (teacher) to a smaller one (student). The teacher often has access to a labeled dataset while the student only has access to the local data.

clustered. Therefore, it has to rely on some statistical mechanism to evaluate how good or how bad this clustering solution is.

In the remainder of this chapter, we will employ the asymptotic descriptors derived in Chapters 2 and 3 to study each merge performed by a hierarchical clustering algorithm as a binary hypothesis test (see Section 4.1), with the null hypothesis being that elements belong to the same group (i.e., have the same covariance). Specifically, we will use the fact that the distance estimators (consistent and *plug-in* ones) can be seen as random variables and, in Section 4.2, use their CLTs to study the probability of having correctly merged two elements that should be merged together. Through numerical experiments, we will demonstrate that the empirical rate of correctly detecting merges can be closely approximated by the cumulative density function of the standard normal distribution. In Section 4.3, we follow a similar intuition and also consider the probability of having wrongly merged two elements that belong to different groups. We will conclude this chapter by showcasing how to use such analysis to gain insights on the performance of binary predictors using the different metrics discussed throughout this work.

4.1 Statistical Analysis of Clustering Evidence

Let us consider a set $\{\hat{\mathbf{Y}}_k\}_{k=1}^{K}$ containing K elements, where each element is a realization of either one of the two processes, denoted by $g = 1$ or $g = 2$ and described by the covariance matrices $\mathbf{R}_g$. We do not know which element is associated to which process (covariance matrix). In learning mechanisms, one often wishes to study whether two elements $\hat{\mathbf{Y}}_i, \hat{\mathbf{Y}}_j$ are generated by the same underlying processes. That is, whether one of the covariance matrices $\mathbf{R}_g$ is (simultaneously) associated with the two elements $\hat{\mathbf{Y}}_i, \hat{\mathbf{Y}}_j$. This analysis helps to determine whether the two elements being compared should be merged into a single cluster or different ones. More formally, one can define each clustering decision as a binary hypothesis test

$$\begin{aligned} \mathrm{H}_0(i,j) &: \mathbf{R}_{(i)} = \mathbf{R}_{(j)} \\ \mathrm{H}_1(i,j) &: \mathbf{R}_{(i)} \neq \mathbf{R}_{(j)} \end{aligned} \tag{4.1}$$

where here we have denoted the covariance matrix associated to the kth element by $\mathbf{R}_{(k)} \in \{\mathbf{R}_g : g = 1, 2\}$. Statistically, if the null hypothesis is accepted, then there exists sufficient evidence that the ith and jth elements should be clustered together. Otherwise, they should be clustered in different groups. Naturally, this

becomes fairly easy if one has access to the covariance matrices of the processes ($\mathbf{R}_g, g = 1, 2$) and to their associations with the covariance matrices of the elements ($\mathbf{R}_{(k)}, k = 1, \ldots, K$). Indeed, there is no need to perform clustering as the solution is trivially obtained by directly comparing these second order moment matrices. Typically, for the kth element, the clustering algorithm does not know if $\mathbf{R}_{(k)} = \mathbf{R}_1$ or $\mathbf{R}_{(k)} = \mathbf{R}_2$. It does not know $\mathbf{R}_g, g = 1, 2$ either, otherwise, the task could be converted into classification instead of clustering. In fact, in a more realistic scenario, it is often the case that the algorithm does not have access to any of the elements' covariance matrices $\mathbf{R}_{(k)}, k = 1, \ldots, K$. Instead, it only has access to the elements $\hat{\mathbf{Y}}_k, k = 1, \ldots, K$ and needs to perform clustering decisions based on these.

Coming back to our example described at the beginning of this chapter, the controller would have access to the covariance matrices $\mathbf{R}_g, g = 1, 2$, but no access to the elements $\hat{\mathbf{Y}}_k, k = 1, \ldots, K$. Hence, it may rely on studying the hypothesis test in (4.1) to evaluate how good or bad is the pairwise merge of two elements. In the remainder of this chapter, we will describe how the statistical study of this hypothesis test can assist in gaining insights on the quality of a clustering solution. In particular, we will detail how the asymptotic characterization of various metrics (described in Chapters 2 and 3) allows us to describe the behavior of this hypothesis test for a given number of samples N_i, N_j. This analysis is later used to study the probability of correctly clustering two elements.

4.2 Probability of Detection

In this section, we will focus on the probability that a clustering solution has detected the merging of two elements, hereafter also denoted as the detection (or merging) rate. Particularly, when the null hypothesis in (4.1) holds true, this translates into correctly deciding for the merge of two elements that should be merged together (correct merging rate). Throughout this section we will demonstrate that, under the null hypothesis, this correct merging rate is closely related to the cumulative distribution function (CDF) of the standard normal distribution. This behavior will happen regardless of the chosen metric. Conversely, the rate at which two elements are wrongly merged is strongly influenced by the chosen metric and, as the system grows, tends towards zero. To carry out this analysis, we will strongly rely on the asymptotic behavior of the family of metrics described in Section 1.2.

Particularly, in this section, we will mainly consider the consistent estimator distance described in Chapter 3. One of the main advantages of doing so is that, by definition (see Section 1.2), when the null hypothesis $\mathrm{H}_0(i,j)$ holds true (i.e., $\mathbf{R}_{(i)} = \mathbf{R}_{(j)}$) we have that $d_M = 0$ which will ease the mathematical notation and the explanation of the following discussion. Nonetheless, the conclusions discussed in this section can be readily applicable to the *plug-in* distances (described in Chapter 2) by considering their respective definitions.

Let us start by recalling some results from Chapter 3. The consistent estimated distance $\tilde{d}_M$ between two sample covariance matrices $\hat{\mathbf{R}}_i$ and $\hat{\mathbf{R}}_j$ asymptotically behaves as a Gaussian random variable around its asymptotic mean $d_M + M^{-1}\mathfrak{m}_M$ with asymptotic variance $M^{-2}\tilde{\sigma}_M^2$. Specifically, when the observations associated with these sample covariance matrices are generated by the same process (i.e., $\mathbf{R}_{(i)} = \mathbf{R}_{(j)}$), by definition, we have that $d_M = 0$. Hence, when comparing two elements, $\hat{\mathbf{Y}}_i$ and $\hat{\mathbf{Y}}_j$, one can study the null hypothesis in (4.1) by using their associated z-score (also known as z-statistic), which we define as

$$z_{ij}^{(g)} = \frac{\tilde{d}_M(\hat{\mathbf{R}}_i, \hat{\mathbf{R}}_j) - M^{-1}\tilde{\mathfrak{m}}_M^{(g)}}{M^{-1}\tilde{\sigma}_M^{(g)}}$$

where $\tilde{\mathfrak{m}}_M^{(g)}$, $\tilde{\sigma}_M^{(g)}$ denote the second order mean and standard deviation of the distance $\tilde{d}_M$, which are obtained by considering $\mathbf{R}_{(i)} = \mathbf{R}_{(j)} = \mathbf{R}_g$ into their respective definitions (see Theorem 3.1).

In a practical scenario, after formulating the hypothesis test, one often uses the z-score to assess the plausibility of the hypothesis being tested. In our scenario, this translates into studying whether two elements are generated by the same covariance matrix or not. This plausibility is usually measured by comparing the z-score against some pre-defined threshold. For instance, the test's significance is a typical threshold that allows one to specify the maximum type-I error rate that they are willing to accept. Our objective is to show that as the dimension of the system increases, we will be able to correctly decide between the null and the alternative hypothesis in (4.1). To do so, let us denote by $\bar{\alpha}_M = \alpha M$ a threshold that scales proportionally to the dimension of the system M. We will assume that, if $z_{ij}^{(g)} \leq \bar{\alpha}_M$, then there exists significant evidence that the two elements $\hat{\mathbf{Y}}_i, \hat{\mathbf{Y}}_j$ should be merged together. Otherwise, they should not be merged together.

One way to evaluate the above is to compare the empirical rate of merges which, for the gth group, is given by

$$\mathbb{P}_{ij}^{(g)}(\bar{\alpha}_M) = \hat{\mathbb{E}}\left[\mathbb{I}\left\{z_{ij}^{(g)} \leq \bar{\alpha}_M\right\}\right]$$

with $\mathbb{I}\{\cdot\}$ being the indicator function, against its asymptotic value, namely the CDF of a standard Gaussian distribution

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{t} e^{\frac{-t^2}{2}} dt$$

which, in turn, is evaluated at $t = \bar{\alpha}_M$. In other words, when the null hypothesis holds true, $z_{ij}^{(g)}$ should asymptotically behave as a standard normal random variable and $\mathbb{P}_{ij}^{(g)}(\bar{\alpha}_M)$ should be approximated by $\Phi(\bar{\alpha}_M)$.

In order to numerically present this idea, we consider the ith multidimensional element $\hat{\mathbf{Y}}_g^{(i)} \in \mathbb{R}^{M \times N_g}$ to be associated to the gth group by the Toeplitz covariance matrix $\mathbf{R}_g$ with first row defined by $[\rho_g^0, \ldots, \rho_g^{M-1}]$. Additionally, let us consider three elements, $\hat{\mathbf{Y}}_1^{(i)}$, $\hat{\mathbf{Y}}_1^{(j)}$ and $\hat{\mathbf{Y}}_2^{(m)}$, where the first two are considered to be associated to the same process ($g = 1$) and the last one to another process ($g = 2$). Then, for some given threshold $\bar{\alpha}_M$, we can estimate the empirical rate $\mathbb{P}_{kl}^{(g)}(\bar{\alpha}_M)$ that two distinct elements (i.e., $k, l \in \{i, j, m\}$ and $k \neq l$) are merged at the gth group by comparing different realizations of the elements $\hat{\mathbf{Y}}_1^{(i)}$, $\hat{\mathbf{Y}}_1^{(j)}$ and $\hat{\mathbf{Y}}_2^{(m)}$. Particularly, in this thesis, this estimation is based on 10^3 realizations of these elements.

To perform this comparison, we start by noticing that the only correct merge is between the pair (i, j). The others combinations (i, m) and (j, m) would be considered as incorrect merges (type-II errors). Moreover, by construction, the optimal merge should occur when using the covariance of the $g = 1$ group. This is because the pair of elements (i, j) are associated to the covariance matrix $\mathbf{R}_1$. Generally, in the remainder of this chapter, when there is no confusion, we will denote by (k, l) any of the three possible merging pairs (i, j), (i, m) or (j, m). Figure 4.1 compares the empirical rates $\mathbb{P}_{ij}^{(g)}(\bar{\alpha}_M), \mathbb{P}_{im}^{(g)}(\bar{\alpha}_M), \mathbb{P}_{jm}^{(g)}(\bar{\alpha}_M)$ against their expected theoretical value $\Phi(\bar{\alpha}_M)$ under the null hypothesis. For each metric (EU, KL, LE), $\mathbb{P}_{ij}^{(g)}(\bar{\alpha}_M)$ (blue circles in the plot) is associated to the correct merge of two elements while $\mathbb{P}_{im}^{(g)}(\bar{\alpha}_M), \mathbb{P}_{jm}^{(g)}(\bar{\alpha}_M)$ (orange plus sign and green cross, respectively) are both associated to the merging of elements that should not be merged together, i.e., to the wrong merge of two elements. The comparison of these empirical rates is done for growing[2] $M, N_i = N_j = N_m$, fixed $\rho_1 = 0.6, \rho_2 = 0.5$, initial $\alpha = 1/8$, and for the different consistent estimator distances considered in this thesis, namely the Euclidean, symmetrized Kullback-Leibler and the log-Euclidean distances (see

[2]Other settings, e.g. $N_i \neq N_j$, under the null hypothesis, result in similar behaviors hence are not displayed here. This is also a consequence of the fact that, asymptotically, the normalized quantity $(\tilde{d}_M - M^{-1}\tilde{\mathfrak{m}}_M^{(g)})/(M^{-1}\tilde{\sigma}_M^{(g)})$ will behave (under $\mathrm{H}_0(i, j)$ in $g = 1$) as a Gaussian random variable with zero mean and unit variance.

Chapter 3). A reference diagonal line is also provided to indicate a perfect match between these values.

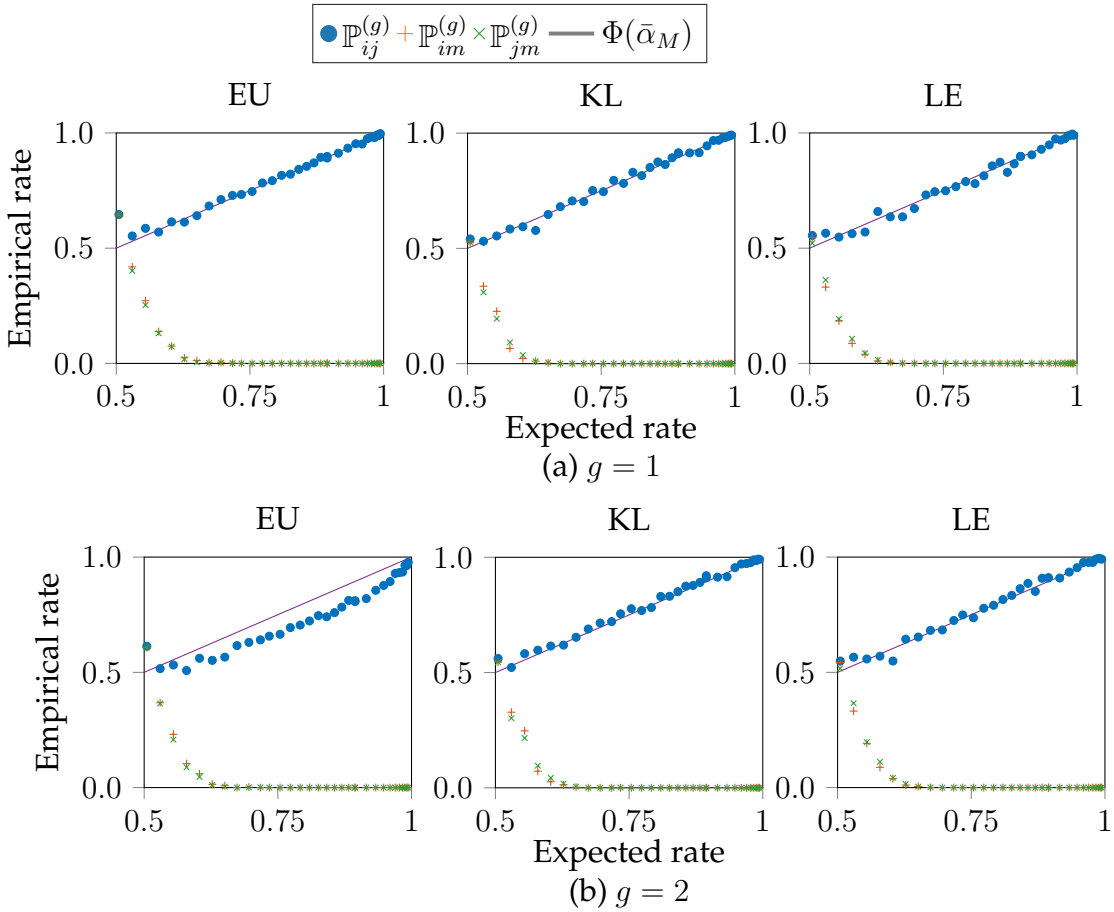


Figure 4.1: Rate of merging of two elements compared against the theoretical expected result when the null hypothesis hold. (a) Merging of two elements in the correct cluster $g = 1$; (b) and in the alternative one $g = 2$, i.e., the other possible null hypothesis.

In general, when the null hypothesis is true, as the system grows larger, the empirical rate of correctly detecting merges $\mathbb{P}_{ij}^{(g=1)}(\bar{\alpha}_M)$ (represented by the blue solid circles) tends to its theoretical value $\Phi(\bar{\alpha}_M)$ (represented by the solid lines crossing the plots). This tendency is visually represented by a close alignment between the blue circles and the solid lines. Moreover, this tendency is a natural consequence of the construction of the z-score. As described above, under the null hypothesis, $z_{ij}^{(g=1)}$ asymptotically follows a standard normal distribution ($z_{ij}^{(g=1)} \sim \mathcal{N}(0,1)$), then, as the system grows large, $\mathbb{P}_{ij}^{(g=1)}(\bar{\alpha}_M)$ asymptotically approaches $\Phi(\bar{\alpha}_M)$. This property holds true for all the considered metrics. Indeed, after closer

analysis, a strong correspondence is also observed in the correct (under the null hypothesis and $g = 1$) rate of detection across all metrics.

Particularly, for the consistent estimator of the KL distance, we recall the results from Section 3.3.2 and note that, whenever under the null hypothesis (i.e., the element's covariance matrices are equal $\mathbf{R}_{(1)} = \mathbf{R}_{(2)}$), the asymptotic quantities $\tilde{\mathfrak{m}}_M^{KL}$ and $\tilde{\sigma}_M^{KL}$ do not depend on $\mathbf{R}_{(1)}$ nor $\mathbf{R}_{(2)}$. Consequently, under the null hypothesis, $\mathbb{P}_{kl}^{(1)}(\bar{\alpha}_M)$ and $\mathbb{P}_{kl}^{(2)}(\bar{\alpha}_M)$ are asymptotically equivalent. This is visually depicted in the figures by having a similar behavior for of the KL distance in Figure 4.1(a) and Figure 4.1(b). In a clustering application, this implies that, when employing the symmetrized KL distance, if two observations belong to the same cluster, it does not matter to which cluster they are associated to ($g = 1$ or $g = 2$), the distribution of $\tilde{d}_M(\hat{\mathbf{R}}_i, \hat{\mathbf{R}}_j)$ is fully described based on the quantities M, N_i, N_j. We also note a similar behavior when evaluating the LE distance and comparing the different figures. However, this time around, the similarity between the results is due to the proximity of the distributions for the first ($g = 1$) and second group ($g = 2$). Nonetheless, this is not a general rule, for the EU distance, for instance, $\mathbb{P}_{ij}^{(g=2)}(\bar{\alpha}_M)$ (wrong cluster assignment) is below the reference line $\Phi(\bar{\alpha}_M)$ while $\mathbb{P}_{ij}^{(g=1)}(\bar{\alpha}_M)$ (correct cluster assignment) is above it.

Finally, we emphasize that the behaviors observed when applying the KL and LE distances do not necessarily pose any detrimental impact on the clustering analysis itself, as the purpose of clustering is to merge elements based on the hypothesis test described in (4.1), which remains unbiased toward the group affiliation of the elements ($g = 1$ or $g = 2$). In this scenario, the primary advantage is that, when considering the alternative hypothesis, the rate of wrongly merging two elements ($\mathbb{P}_{im}^{(g)}(\bar{\alpha}_M)$ and $\mathbb{P}_{jm}^{(g)}(\bar{\alpha}_M)$ in our setting) converges to zero regardless of the true cluster assignment. In the next section, we further investigate the impacts of the alternative hypothesis into the general clustering task with more than two elements.

4.3 Assessing Clustering Performance

In the previous section, we investigated the probability of correctly detecting a merge. However, more generally, one wishes to choose a metric that properly maximizes the probability of detection while minimizing the probability of false alarm (type-II error). Hence, in this section, we shift our focus to this more general case and evaluate how the choice of different metrics impacts both probability of

detection and type-II error. A common way to visualize the trade-off between these quantities is in terms of the Receiver Operating Characteristic (ROC) curve that describes the probability of detection as a function of the probability of false alarm, i.e., fixed probability of keeping the elements in different clusters when they should be clustered together.

Similarly as above, in order to numerically present this idea, we consider the ith multidimensional element $\hat{\mathbf{Y}}_g^{(i)} \in \mathbb{R}^{M\times N_g}$ to be associated to the gth group by the Toeplitz covariance matrix $\mathbf{R}_g$ with first row defined by $[\rho_g^0, \ldots, \rho_g^{M-1}]$. The class assignment is unknown to the algorithm, then, for each binary comparison, the ROC curve depicts the probability of correctly merging two elements that belong to the same cluster (e.g., $\hat{\mathbf{Y}}_1^{(i)}$ and $\hat{\mathbf{Y}}_1^{(j)}$, for $g = 1$), for some fixed false alarm rate. Figure 4.2 presents the ROC curves for these distances in various scenarios. A random binary classifier is also represented by a dashed line crossing the origin and upper right corner of these plots. An optimal solution aims to get closer to the upper left corner of the plot, this would represent a maximum probability of detection and for a given probability of false alarm. During our experiments, we noticed that this is possible for the trivial case of large $|\rho_1 - \rho_2|$ and increasing M, N_1, N_2. Nonetheless, in the more general case, we notice that the choice of the best metric is highly dependent on the specific choice of scenario and its parameters. In this context, regardless of the used estimator, different settings (rows of Figure 4.2(a)) will lead to different distributions of the ROC curves even for small changes in ρ_1, ρ_2 which describe the true covariance matrices.

A similar behavior is also noticed by simply swapping $\rho_1 \leftrightarrow \rho_2$ (rows of Figure 4.2(b)-(c)). These analyses become particularly useful in clustering scenarios where more than two elements are being compared. Consider, for example, two groups $g = 1, 2$, each defined by the set of elements $\hat{\mathcal{Z}}_g = \{\hat{\mathbf{Y}}_g^{(l)}\}_{l=1}^{L_g}$ of size $L_g > 1$. Moreover, assume that all the elements associated to gth group are also generated by to the covariance matrix $\mathbf{R}_g$ with N_g observations each. Then, given the joint set $\hat{\mathcal{Z}}_2 \cup \hat{\mathcal{Z}}_2$, the clustering goal is to form two distinct clusters, one containing all the L_1 elements generated by $\mathbf{R}_1$, namely $\hat{\mathcal{Z}}_1$, and another cluster containing all the L_2 elements associated to $\mathbf{R}_2$, namely $\hat{\mathcal{Z}}_2$. In this scenario, separately analyzing the merges of the elements in the two groups $\hat{\mathcal{Z}}_1$ or in $\hat{\mathcal{Z}}_2$ might be misleading as both merges need to occur for an accurate clustering to happen. In other words, one has to ensure that the distance between $\mathbf{Y}_1^{(k)}$ and $\mathbf{Y}_2^{(l)}$ (elements from different clusters) is smaller than the distance between $\mathbf{Y}_g^{(i)}$ and $\mathbf{Y}_g^{(i)}$ (elements from the

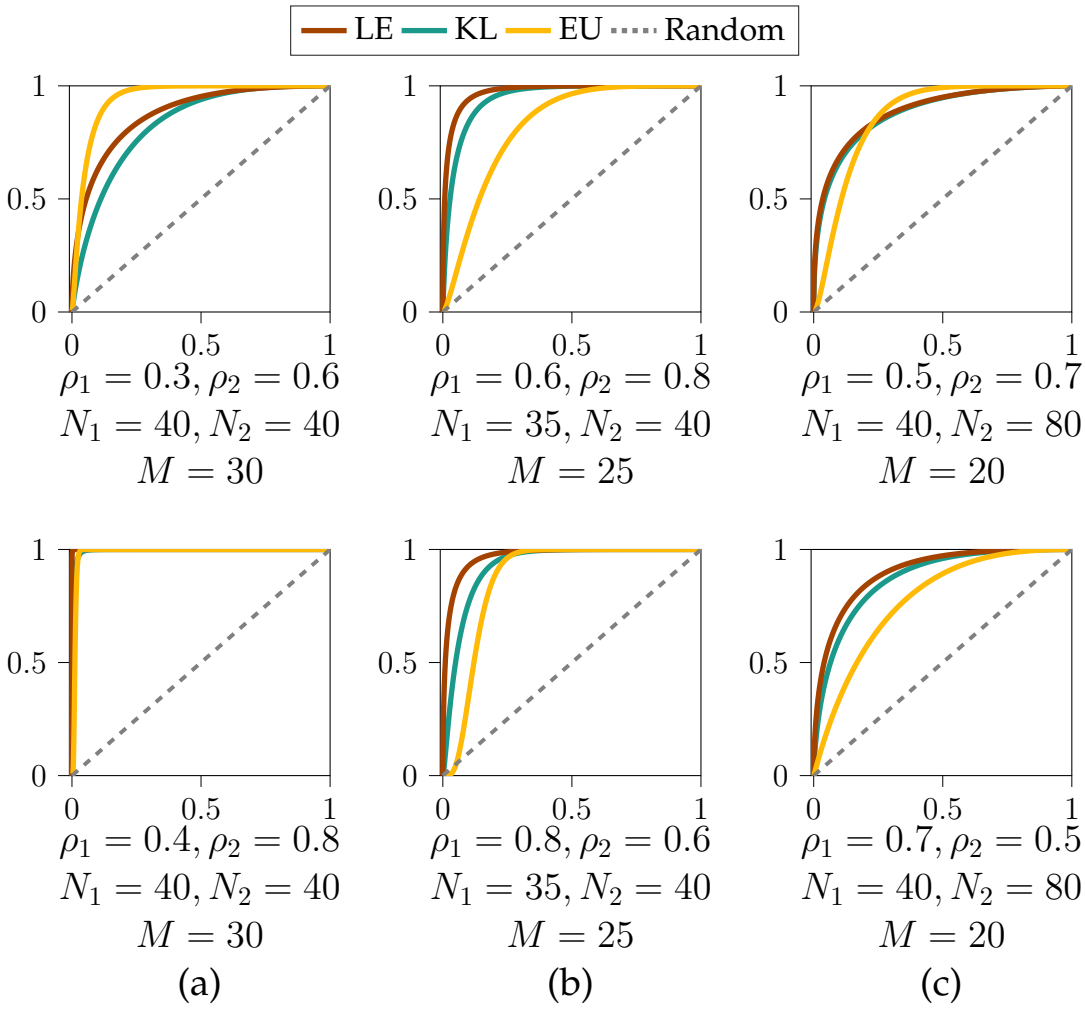


Figure 4.2: ROC curves for binary hypothesis test in various scenarios using consistent estimator. The choice of the best metric depends on each scenario.

same cluster), simultaneously for both $g = 1$ and $g = 2$. Otherwise, it might happen that $\tilde{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)}) > \tilde{d}_M(\hat{\mathbf{Y}}_1^{(i)}, \hat{\mathbf{Y}}_1^{(j)})$ happens, indicating a correct merge, but it is also possible that $\tilde{d}_M(\hat{\mathbf{Y}}_2^{(i)}, \hat{\mathbf{Y}}_2^{(j)}) > \tilde{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$ happens which would indicate wrongly merging two elements. Hence, in this case, to ensure the correct clustering, we need to assess scenarios simultaneously.

To help better illustrate this concept, Figure 4.3 compares the (asymptotic) probability density functions of $\tilde{d}_M(\hat{\mathbf{Y}}_1^{(i)}, \hat{\mathbf{Y}}_1^{(j)})$ (blue) and $\tilde{d}_M(\hat{\mathbf{Y}}_2^{(i)}, \hat{\mathbf{Y}}_2^{(j)})$ (green) against the PDF of $\tilde{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$ (magenta) for the same setting as in the Figure 4.2(c). Notice that these PDFs do not depend on the particular realization of any observations because, by construction, in our experiment, all elements associated with the gth covariance matrix have the same size $M \times N_g$. The first two PDFs (blue and green) represent the distance between elements that belong to the same clus-

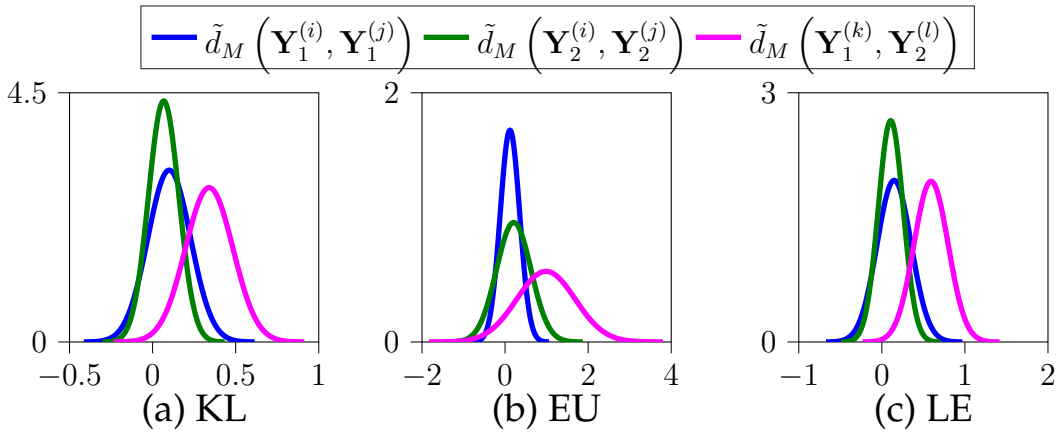


Figure 4.3: Comparison of the PDF of the (consistent) distances between elements of the same class (blue and green) and elements of different classes (magenta) for $M = 20$, $N_1 = 40, N_2 = 6.0$ and $\rho_1 = 0.5, \rho_2 = 0.7$.

ter ($g = 1$ or $g = 2$, respectively). The final PDF (magenta) represents the distance between elements that belong to different groups and should not be merged together. Optimally, the first two PDFs should be to the left of the latter one, i.e., distances of elements within the same group (inner-group) should be smaller than the ones between elements of different groups (inter-group). We recall from Theorem 3.1 in Chapter 3 that, under the null hypothesis, the first-order asymptotic mean of the consistent estimator is zero (the true covariance matrices are equal hence $d_m = 0$). As a result, it (usually) becomes fairly easy to correctly distinguish $\tilde{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$ from the other two curves, $\tilde{d}_M(\hat{\mathbf{Y}}_g^{(i)}, \hat{\mathbf{Y}}_g^{(j)}), g \in \{1, 2\}$, with a rather small false alarm rate (refer to Section 4.3.2 for a similar analysis using the *plug-in* distance). Nonetheless, as there might exist overlaps between these three distributions, to reliably analyze the effectiveness of a metric in clustering different elements, one needs to perform a multiple hypothesis test to assert that the events $\tilde{d}_M(\hat{\mathbf{Y}}_g^{(i)}, \hat{\mathbf{Y}}_g^{(j)}) < \tilde{d}_M(\hat{\mathbf{Y}}_g^{(k)}, \hat{\mathbf{Y}}_{g'}^{(l)})$ hold true with high probability, for $g \neq g'$ and $g, g' \in \{1, 2\}$.

As we need to account for the probability of multiple events, it becomes natural to use a multiclass ROC curve to analyze this joint probability for the different configurations of M, N_1, N_2 and covariances $\mathbf{R}_1, \mathbf{R}_2$. Moreover, the multiclass ROC curve simultaneously associated to both of these events becomes the average between their singular ROC curves. Particularly, Figure 4.4(a)-(b) presents these curves for the same settings as in Figure 4.2(b)-(c), respectively. An interesting consequence of directly analyzing these curves is that they allow us to compare the suitability of distance metrics for a specific problem without the need of acquiring

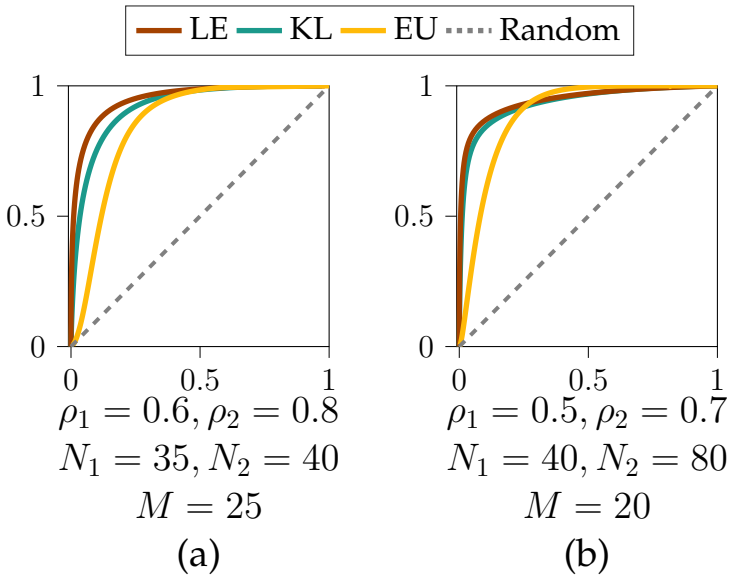


Figure 4.4: Average ROC curves for binary hypothesis test in two different scenarios using *plug-in* estimator.

any data. In other words, by using the proposed asymptotic descriptors, one is capable of assessing the behavior of a metric in a specific scenario based solely on its statistics. This becomes particularly useful in agglomerative clustering tasks where a pair of elements are combined based on how distant they are from one another. Finally, in what follows, we will use the area under the (multiclass) ROC curve (AUC) to summarize into a single quantity the suitability of a metric applied in a specific setting. For two distinct random processes, the AUC is associated to the probability that the model ranks two random positive examples (samples that belong to the same groups) more highly than a random negative example (samples that belong to different groups). In a clustering scenario, this represents correctly merging two elements that belong to the same cluster.

4.3.1 Impacts on Clustering using Consistent Estimators

To analyze the impacts of the results presented above to the task of selecting an optimal metric to cluster random elements, we will consider two random processes and their associated groups $\hat{\mathcal{Z}}_1$ and $\hat{\mathcal{Z}}_2$, each containing L_1, L_2 elements, respectively, that need to be clustered. We first evaluate the performance of a hierarchical clustering with the average linkage to cluster these elements considering the different consistent estimators $\tilde{d}_M$ of the metrics EU, KL and LE (see Section 4.3.2 for a similar analysis using the *plug-in* estimators). At the beginning, each element

forms a cluster on its own and the final goal is to iteratively combine the most similar groups until we are left with only two disjoint sets. At the end of each simulation, we evaluate the clustering results by considering the accuracy (ACC) and the adjusted rand index (ARI) [64]. These are obtained by comparing the empirical clustering solution (result of the hierarchical clustering) to the expected (ground truth) one. The accuracy of the clustering result is calculated as the percentage of elements that are associated to the correct cluster and is given by

$$\mathrm{ACC} = \frac{|\mathcal{G}_1 \cap \hat{\mathcal{Z}}_1| + |\mathcal{G}_2 \cap \hat{\mathcal{Z}}_2|}{L_1 + L_2},$$

where $\mathcal{G}_1$ and $\mathcal{G}_2$ are the empirical clustering results associated to the true ones, $\hat{\mathcal{Z}}_1$ and $\hat{\mathcal{Z}}_2$, respectively. This association is not directly provided by the clustering algorithm. Instead, here we associate the solutions to $\mathcal{G}_1$ and $\mathcal{G}_2$ such that the ACC above is maximized.

The ARI is another common way to assess the quality of a clustering solution and basically consists in counting the number of pairs of elements that are grouped together (or separately) in both the true ($\hat{\mathcal{Z}}_1, \hat{\mathcal{Z}}_2$) and the empirical clustering ($\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2$) solutions, i.e., the number of true positives (respectively, true negative) pairs of elements. Hereafter also denoted by TP and TN, respectively. More specifically, the rand index (RI) coefficient is a measure which accounts for both these quantities and it is given by

$$\mathrm{RI} = \frac{TP + TN}{\binom{L}{2}}$$

where the denominator is the number of all possible pairwise combination among the L elements being clustered. The RI coefficient ranges from 0 to 1 and, if all the pairwise combinations, (TP and FP ones) are correct, it means that both the predicted and ground truth clustering solutions are equal, so that we obtain $\mathrm{RI} = 1$. However, it is well-known this traditional definition of the RI coefficient is (often) sensitive to the number of clusters [64, 65] and that the above definition can (by chance) yield high RI values even when the clustering solution has only a few agreements to the ground truth one (e.g., high TP or TN). A typical improvement on the traditional RI is the adjusted rand index (ARI)

$$\mathrm{ARI} = \frac{\mathrm{RI} - \left[\sum_i \binom{L_i}{2} \sum_j \binom{|\hat{\mathcal{G}}_j|}{2}\right] / \binom{L}{2}}{\frac{1}{2}\left[\sum_i \binom{L_i}{2} + \sum_j \binom{|\hat{\mathcal{G}}_j|}{2}\right] - \left[\sum_i \binom{L_i}{2} \sum_j \binom{|\hat{\mathcal{G}}_j|}{2}\right] / \binom{L}{2}},$$

where L_i (respectively $|\hat{\mathcal{G}}_i|$) represents the total number of elements in the ith cluster according to the true (respectively predicted) clustering solution. As a result, the ARI definition above describes a normalized metric that now takes into account both the probability of randomly combining two elements together and the probability of this happening due to the clustering algorithm (see [64]). ARI values close to (or below) zero represent solutions equal (respectively, worse) than a random clustering algorithm. Alternatively, high ARI values indicate a good agreement between the clustering solution and the ground truth.

Moreover, to validate our proposed metric selection mechanism, for each metric, we compare their AUC (obtained from their asymptotic descriptors) against their average ACC and average ARI, each obtained over 10^3 simulations. We consider each true group to contain $L_1 = L_2 = 8$ elements that need to be clustered, each of these elements being real-valued observation sets of size $M \times N_g$, for $g = 1, 2$. Table 4.1 presents this comparison for different scenarios. Notice that there seems to be a strong correlation between the average ARI and the AUC/ARI. Specifically, it seems that the metric with the highest AUC also yields the best clustering assignment (highest ACC and ARI). However, it is important to note that it is not possible to directly predict the overall ACC nor the ARI of the hierarchical solution solely based on the AUC. The ACC and the ARI of the clustering solution may vary depending on several factors such as the number of elements and the chosen linkage method. Therefore, while the AUC provides valuable information about the performance of a metric, it should not be used as the sole indicator of clustering accuracy. Other factors should be considered, such as the specific characteristics of the system and of the clustering algorithm being used.

Table 4.1: Comparison of AUC, ACC and ARI for different consistent estimators.

Scenario					**KL**			**EU**			**LE**		
M	N_1	N_2	ρ_1	ρ_2	**AUC**	**ACC**	**ARI**	**AUC**	**ACC**	**ARI**	**AUC**	**ACC**	**ARI**
30	40	40	0.3	0.6	0.830	0.678	0.342	**0.905**	**0.958**	**0.893**	0.875	0.893	0.788
30	40	40	0.4	0.8	0.967	0.997	0.993	0.841	0.994	0.984	**0.998**	**1.000**	**1.000**
25	40	35	0.8	0.6	0.915	0.874	0.722	0.800	0.820	0.619	**0.954**	**0.984**	**0.963**
20	40	60	0.5	0.7	0.918	0.880	0.787	0.831	0.886	0.683	**0.934**	**0.951**	**0.913**

4.3.2 Impacts on Clustering using *Plug-in* Distance

A similar analysis as the one above can also be performed by using the traditional *plug-in* estimators $\hat{d}_M$, where now the distances are either KL, EU or SS. We recall that, in this case, we are primarily concerned with the undersampled scenario where the number of observations N_g is lower than the dimensionality of the underlying system M. Figure 4.5 presents the ROC curves for these distance estimators for varying N_1, N_2, ρ_1, ρ_2 and fixed $M = 30$. Again, in order to achieve an optimal solution, the goal is for the ROC curve to approach the upper part of the plot. This position corresponds to a maximum probability of detection and a given probability of false alarm. Conversely, classifiers that lie below the random classifier (dashed line) represent the case where it is not possible to correctly distinguish between the two groups. Specifically, for the *plug-in* distance, this is a consequence of the undesired scenarios where we observe larger distances between elements from the same group than between elements from different groups, e.g., $\hat{d}(\hat{\mathbf{Y}}_1^{(i)}, \hat{\mathbf{Y}}_1^{(j)}) > \hat{d}(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$, for $i \neq j, k \neq l$. To help better illustrate this concept, Figure 4.6 now compares the probability density functions of $\hat{d}_M(\hat{\mathbf{Y}}_1^{(i)}, \hat{\mathbf{Y}}_1^{(j)})$ (blue) and $\hat{d}_M(\hat{\mathbf{Y}}_2^{(i)}, \hat{\mathbf{Y}}_2^{(j)})$ (green) against the PDF of $\hat{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$ (magenta) considering the *plug-in* estimators and for the same setting as in the Figure 4.5(c). Particularly, Figure 4.6(b) shows that, when applying the SS distance, even for large false alarm rates, it is very unlikely to obtain (in the scenario considered here) $\hat{d}_M(\hat{\mathbf{Y}}_1^{(i)}, \hat{\mathbf{Y}}_1^{(j)}) < \hat{d}_M(\hat{\mathbf{Y}}_1^{(k)}, \hat{\mathbf{Y}}_2^{(l)})$. As mentioned before, a direct consequence of this behavior is that its respective ROC curve (orange line in the left hand side plot of Figure 4.5(c)) is pushed below the random classifier. Notice that this could happen when using any of the *plug-in* distances depending on the scenario considered.

Similarly as in the previous section, in the subsequent discussion we will use the area under the (averaged multi-class) ROC curves as a method to summarize, into a single quantity, the effectiveness of a metric applied in a specific context. We will consider two random processes and their associated group of elements $\mathcal{G}_1, \mathcal{G}_2$, each containing $L_g = 8$ elements that need to be clustered. This time around, we evaluate the performance of a hierarchical clustering to cluster these elements considering the different consistent estimators of the metrics KL, EU and SS. Table 4.2 presents the comparison of the AUC, ACC and ARI for different settings and metrics using the *plug-in* estimators. Results are similar to the ones achieved when employing the consistent estimator, i.e., higher AUC is strongly correlated with the metric that leads to higher ACC/ARI values. Specifically, for the cases where

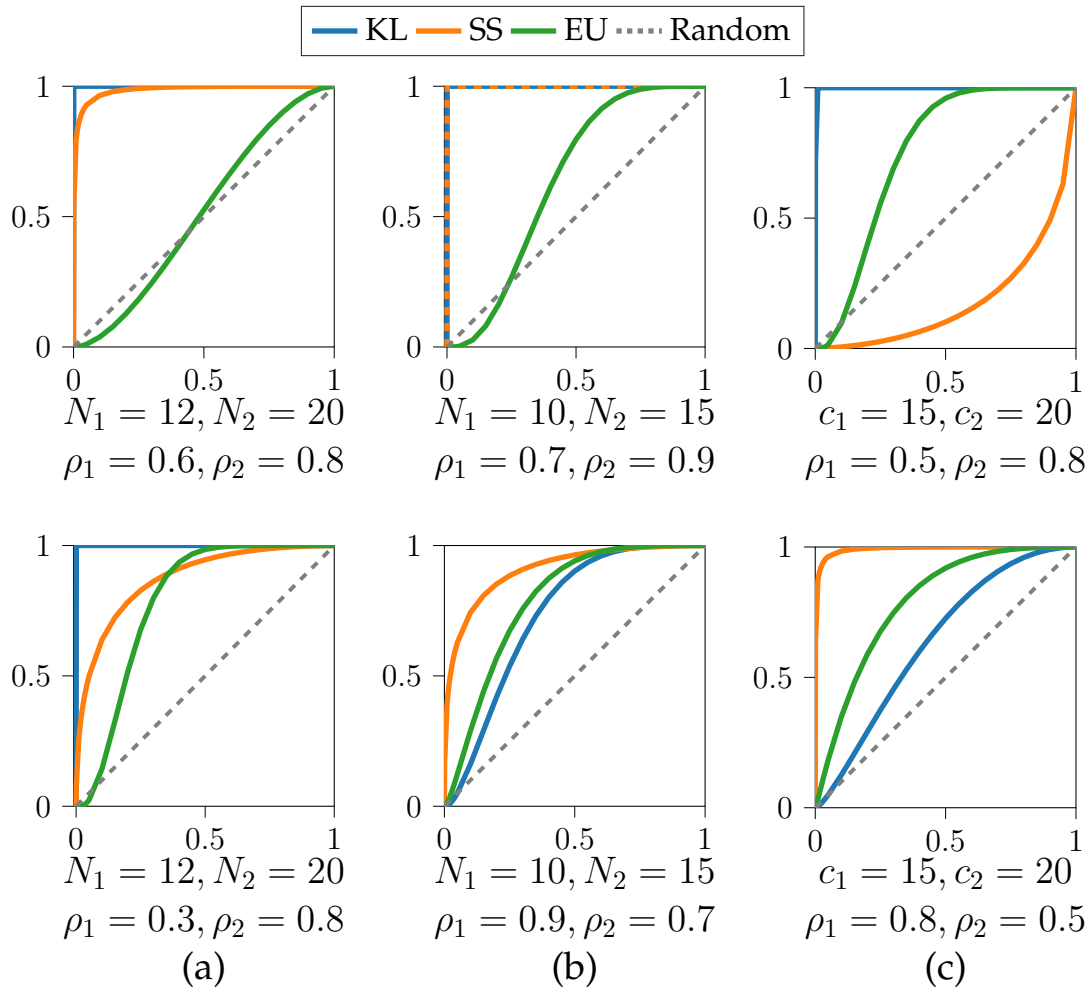


Figure 4.5: ROC curves for binary hypothesis test of in various scenarios using *plug-in* estimator and for $M = 30$.

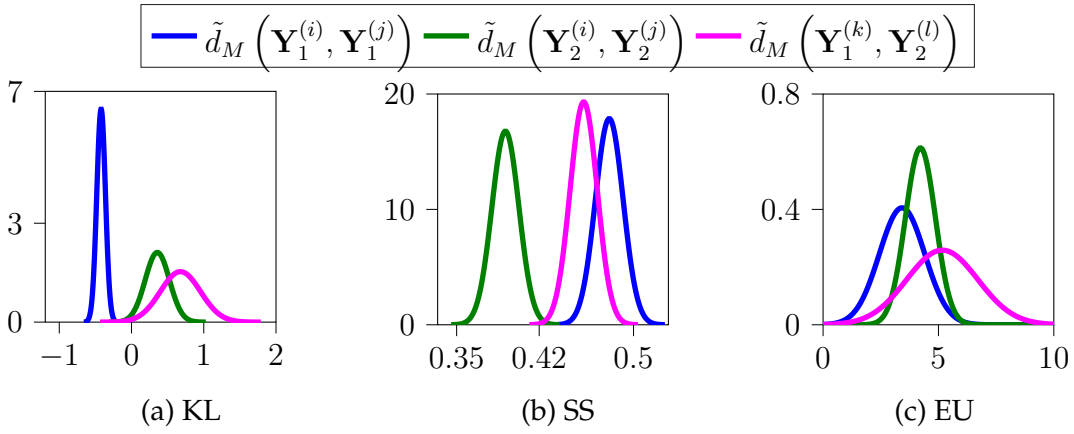


Figure 4.6: Comparison of the PDF of the distances between elements of the same class (blue and green) and elements of different classes (magenta) for $M = 30$, $c_1 = 1.5, c_2 = 2.0$ and $\rho_1 = 0.8, \rho_2 = 0.5$.

Table 4.2: Comparison of AUC, ACC and ARI for different *plug-in* estimators.

Scenario					KL			EU			SS		
M	N_1	N_2	ρ_1	ρ_2	**AUC**	**ACC**	**ARI**	**AUC**	**ACC**	**ARI**	**AUC**	**ACC**	**ARI**
30	15	20	0.5	0.9	**0.983**	**0.999**	**0.997**	0.927	0.927	0.811	0.682	0.417	0.035
30	20	25	0.4	0.8	0.818	0.759	0.438	**0.952**	**0.950**	**0.877**	0.513	0.431	0.004
30	12	12	0.3	0.8	0.737	0.615	0.067	**0.886**	**0.747**	**0.383**	0.592	0.411	0.017
30	10	25	0.7	0.9	0.818	0.787	0.499	0.782	0.640	0.274	**1.000**	**1.000**	**1.000**
30	15	25	0.5	0.7	0.645	0.564	0.001	0.628	0.451	0.004	**0.883**	**0.917**	**0.846**
30	15	20	0.3	0.8	0.797	0.635	0.129	**0.847**	**0.723**	**0.400**	0.698	0.417	0.044

AUC is close to 0.5 (random classifier) the ARI becomes very close to zero while the ACC becomes close to 0.5 which might be misleading in some scenarios. In general, the similarity between the results obtained in this section (using the *plug-in*) and the ones obtained in the previous section (using the consistent estimators) further suggests that it is possible to favor one metric over another based solely on the AUC of their respective ROC curves.

4.4 Conclusions

The asymptotic characterization of the general class of distances between sample covariance matrices presented in Section 1.2 has been applied to assert the quality of clustering solutions. Specifically, we have described this process as a binary hypothesis test and evaluated its probability of detection and type-II error. Together, these quantities provide valuable insights into determining if two elements are generated by the same or different processes (i.e., whether they should be clustered together or not). We showcase these results in several scenarios, employing both our proposed consistent and *plug-in* estimators. Particularly, this illustrates that the assertion mechanism described in this chapter is applicable regardless of the metric system used and both in the over- and undersampled case. Finally, we stress that the results described here are also applicable to other clustering mechanisms that are based on pairwise similarity comparisons and to other distances that follow the same structure as the class of distances discussed throughout this work.

Chapter 5

Subspace Similarity Applied to Wireless Communications

Multi-antenna radio access technologies are extensively utilized to improve the spectral efficiency and connectivity of wireless communications systems. By employing multiple antennas at both the transmitter and the receiver sides (MIMO systems), these technologies can make use of advanced signal processing techniques to exploit the spatial dimension of the wireless channel. For instance, by exploiting the spatial degrees of freedom provided by multiple antennas, MIMO systems can overcome the adverse effects of fading while improving the overall quality and reliability of wireless connections. Particularly, in the uplink (MAC channel), space-division multiple access (SDMA) is a traditional method that has been extensively used to enhance spectral efficiency by exploiting the spatial separation of transmitters to optimize the use of different channelization protocols (enhancing time and frequency multiplexing using spatial multiplexing). Similarly, in the downlink (broadcast channel), dirty–paper coding (DPC) has shown to achieve the channel capacity region. While DPC has theoretical advantages by exploiting the presence of self-induced interference, the practical complications associated with the need for accurate channel state estimation and the inherent computational complexity make its implementation challenging. In this context, several suboptimal alternatives have been proposed in the literature, such as non-orthogonal multiple access (NOMA) [21], joint spatial division multiplexing (JSDM) [22], and hierarchical rate splitting (H-RSMA) [23] to name a few. These downlink methods share a common approach used to improve communication efficiency, which involves partitioning receivers into clusters based on some pre-defined criteria. In such scenarios, one can rely on clustering for leveraging the spatial relationships

among signals and mitigating multi-transmitter interference, thereby enhancing overall communication performance.

This chapter focuses on the general task of clustering of users based on the scattering at the base station side, regardless of whether the application is in the uplink or downlink. Finding the optimal user clustering for a specific transmission scheme is generally a challenging problem that involves comparing all possible partitions of different channels into groups. This approach incurs exponentially high computational complexity, making it infeasible in practice. Hence, when there are sufficient degrees of freedom, it is reasonable to form groups based on spatial proximity such that receivers/transmitters with similar angles of arrival (AoA)/angles of departure (AoD) are grouped together. Unfortunately, when multipath is present, geometric/spatial proximity is almost meaningless and one should instead measure users' proximity based on how well aligned the subspaces spanned by their channel matrices are [23, 24]. This is closely related to what we have discussed in the previous chapters of this dissertation: the similarity between two subspaces can be measured as

$$\hat{d}_M^{SS} = \frac{1}{M}\mathrm{tr}\left[(\mathbf{P}_1 - \mathbf{P}_2)^2\right].$$

where $\mathbf{P}_k$ (see below) represents the projection matrix onto the column space of the kth user channel. A similar line of reasoning can also be used when comparing the (sample) channel covariance matrices of the different UEs. In this scenario, users that are close together will have similar covariance matrices.

Throughout this chapter, we will explore how the results presented in the previous chapters of this thesis can be applied to the clustering of wireless devices. We will primarily focus on the undersampled regime hence, we will rely on the *plug in* distances studied in Chapter 2 of this thesis. In the multi-user MIMO scenario (MU-MIMO), the undersampled regime represents the case where the number of antennas at the base station is larger than the number of users in each cluster. Simultaneously, we will also consider the case where the total number of users (i.e., the sum of the number of users in each cluster) is larger than the number of antennas at the base station, scenarios where clustering can assist in mitigating interference.

In what follows, we start by presenting the MIMO signal model considered in this chapter as well as how to apply the results derived in the previous chapters to this specific setting. We will leverage the results from the previous chapters to propose a hierarchical clustering solution designed specifically for the comparison of

non-equidimensional channel matrices. This scenario corresponds to the situation where groups with different number of users are being compared. Alternatively, one could also see this by comparing UEs that are equipped with different numbers of antennas [7]. Throughout this chapter we will focus on the first scenario. Since we are dealing with the undersampled scenario, most of the consistent estimators derived in the previous chapter are not applicable. For this reason, we will mainly focus on the *plug-in* distance. The main idea behind this chapter is to adopt these distances by connecting their asymptotic bias when the two compared covariances are the same. Specifically, we build on the asymptotic results established in Chapter 2 to introduce correction terms that converge to the asymptotic equivalents of various *plug in* distances, but only when the two covariance matrices are equal. This guarantees that the corrected distance converges to zero in this situation, which also guarantees consistency when the two covariance matrices are equal. We will show that the corrected distances effectively assist in the comparison and clustering of non-equidimensional channels matrices. Finally, we will conclude this chapter with a comparative analysis between our proposed corrected metrics and the traditional ones.

5.1 MIMO Signal Modeling

Consider a wireless scenario with a base station equipped with M antennas and $K > M$ single antenna user equipments (UEs) which are divided into G groups, each of size $N_k, k = 1, \dots, G$ and $\sum N_k = K$ together with $N_k < M$. The MU-MIMO channel of the kth group can be described by a $M \times N_k$ matrix of complex entries $\mathbf{Y}_k$. We consider the same signal model with correlation/scattering at the BS side as the one presented in Remark 2.1 in Chapter 2. In the context of wireless communications, this can be understood as a Rayleigh model for MIMO fading channels, according to which the channel matrices $\mathbf{Y}_k$ are independent among group of users and can be decomposed as

$$\mathbf{Y}_k = \mathbf{R}_k^{\frac{1}{2}} \mathbf{X}_k \tag{5.1}$$

where $\mathbf{R}_k \in \mathbb{C}^{M \times M}$ is assumed to be full rank and represents the channel spatial covariance matrix at the BS. In this case, the matrix $\mathbf{R}_k$ is inherently dependent on the scattering structure of the scenario, in particular on the multiple channel paths. Moreover, the entries of the $M \times N_k$ matrix $\mathbf{X}_k$ are assumed to be independent

and identically distributed (i.i.d.) complex circularly symmetric Gaussian random variables, with zero mean and unit variance (Rayleigh fading).

The results presented in this chapter are quite general and applicable to any kind of observation sets that follow (5.1). Nonetheless, inspired by the channel clustering application in wireless communications, we will tailor our analysis to the MIMO setting and simulate observation matrices as MU-MIMO wireless channels. In this context, the covariance matrices $\mathbf{R}_k, k = 1, \ldots, G$ are generated by averaging the contribution of random directions of arrival, which impinge on a uniform linear array with half a wavelength inter-element separation [66, Chapter 2]. We will denote by φ_k the average angle of arrival associated to each channel, we assume them to be Gaussian distributed with angular spread δ_φ^2. For two covariance matrices, $\mathbf{R}_i$ and $\mathbf{R}_j$, the separation between φ_i and φ_j can be understood as a spatial distance between the ith and jth groups of users. In other words, it is associated to the distance between the distribution from the associated channel matrices or, in a more geometric view, to the distance between the subspaces that are spanned by the channel matrices associated to each group. For instance, for small $\Delta\varphi_{ij} = |\varphi_i - \varphi_j|$, we have that two groups are close together, thus it becomes harder to distinguish between one another. In contrast, for $\Delta\varphi_{ij}$ large, both groups are generated by non-related subspaces hence are easily to distinguish from one another.

As mentioned above, we are interested in establishing whether two set of UEs channels $\mathbf{Y}_i, \mathbf{Y}_j$ belong to the same spatial sector as seen from the BS. This spatial information is completely contained in the column space of the channel matrices $\mathbf{Y}_k, k = 1, \ldots, G$. Hence, we will formally assign the two matrices to the same spatial cluster when the two column spaces are close enough, in terms of a distance/proximity measure that will be specified in what follows.

Remark 5.1. *Throughout this chapter we will simultaneously follow two intuitions to cluster the channel matrices described in (5.1). The first one, described above, considers the alignment of the column space of two channel matrices, e.g.,* $\mathrm{span}(\mathbf{Y}_k), \mathrm{span}(\mathbf{Y}_j)$. *If these subspaces are well aligned, it suggests that the two channels should be grouped together. In a clustering context, this often translates into selecting the pair of channel matrices that exhibit the best/highest alignment. This approach is commonly employed in several wireless applications, such as [7, 20, 67] to name a few, where the interest is in clustering wireless channels according to their field of view or the set of directions of arrival associated to them. A detailed analysis is conducted in the following section.*

Now, a different view of the problem considers the direct comparison of the sample covariance matrices associated to the channel matrices $\mathbf{Y}_k, \mathbf{Y}_j$*, namely,*

$$\hat{\mathbf{R}}_k = \frac{1}{N_k}\mathbf{Y}_k\mathbf{Y}_k^{\mathrm{H}} \quad \textit{and} \quad \hat{\mathbf{R}}_j = \frac{1}{N_j}\mathbf{Y}_j\mathbf{Y}_j^{\mathrm{H}}. \tag{5.2}$$

This is a more natural clustering problem in applications where we want to group users according to their physical proximity (i.e., taking into account both their directions of arrival and their power/distance to the base station). In this scenario, it is natural to consider the family of metrics described throughout this thesis. Here again, given that we are strictly examining the undersampled regime where $N_k, N_j < M$*, we will rely on the plug in distances described in Chapter 2.*

5.2 Subspace Comparison and Grassmann Manifolds

The distance between subspaces can be geometrically characterized by their principal angles. Let $\mathcal{H}_k, \mathcal{H}_j \subseteq \mathbb{C}^M$ denote the subspaces spanned by the columns of two complex-valued matrices $\mathbf{Y}_k \in \mathbb{C}^{M\times N_k}, \mathbf{Y}_j \in \mathbb{C}^{M\times N_j}$, respectively, where $\Omega_{kj} = \{\alpha_{kj}(1), \alpha_{kj}(2), \ldots, \alpha_{kj}(\min(N_k, N_j))\}$ be the principal angles between these two subspaces. It can be seen that the cosine of the principal angles between two subspaces $\mathcal{H}_k, \mathcal{H}_j$ are the singular values of the matrix $\hat{\mathbf{V}}_k^{\mathrm{H}}\hat{\mathbf{V}}_j$, where we have denoted by $\hat{\mathbf{V}}_k$ the first N_k left singular vectors associated to the kth channel matrix[1] and similarly for $\hat{\mathbf{V}}_j$. The complex (or real) Grassmannian $\mathcal{G}(N, M)$ can be seen as the complex (real) manifold built of the symmetric projection matrices of size $M \times M$ and rank N [27]. Hence, given an observation matrix $\mathbf{Y}_k$, one can rely on its (left) column space $\mathcal{H}_k$ to design an equivalence to the point $\mathbf{P}_k$ in the Grassmannian $\mathcal{G}(N_k, M)$, where $\mathbf{P}_k$ is the projection matrix

$$\mathbf{P}_k = \mathbf{Y}_k(\mathbf{Y}_k^{\mathrm{H}}\mathbf{Y}_k)^{-1}\mathbf{Y}_k^{\mathrm{H}} = \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^{\mathrm{H}}. \tag{5.3}$$

Notice that $\mathbf{P}_k$ is fully described by the first N_k left singular vectors $\hat{\mathbf{V}}_k$ of the channel matrix $\mathbf{Y}_k$. This will later become particularly useful in our MU-MIMO clustering setting, but for now, let us continue to describe the general scenario.

Remark 5.2. *For readability, in this section, we focus on the case where two subspaces have the same dimension, that is* $N_k = N_j = N$*. Nonetheless, we emphasize that these results can be trivially generalized to the more general case where* $N_k \neq N_j$ *(see Section 5.3).*

[1] A similar analysis can also be performed in the context of covariance matrices by considering $\hat{\mathbf{V}}_k$ as the eigenvectors of the sample covariance matrix associated to the kth UE.

Specifically, by using an appropriate embedding, we can still consider all the distances below by simply selecting the positive eigenvalues in the corresponding definitions (or, equivalently, by setting $N = \min(N_k, N_j)$*).*

Interesting enough, if we denote by $\hat{\lambda}_1^{(kj)} \geq \ldots \geq \hat{\lambda}_N^{(kj)}$ the non-zero eigenvalues of $\mathbf{P}_k\mathbf{P}_j\mathbf{P}_k$, we can relate the principal angles Ω_{kj} and these eigenvalues by

$$\cos^2(\alpha_{kj}(i)) = \hat{\lambda}_i^{(kj)}.$$

The main advantage of identifying a channel as a point in the Grassmannian is that we can consider conventional distance measures that define the topological structure of this manifold, see [68] for a detailed review. In particular, in this thesis we consider the squared projection-Frobenius distance, which is defined as

$$d_{\text{PF}}^2(\mathbf{Y}_k, \mathbf{Y}_j) = \sum_{i=1}^{N} \sin^2(\alpha_{kj}(i)) = N - \sum_{i=1}^{N} \hat{\lambda}_i^{(kj)} = N - \text{tr}(\mathbf{P}_k\mathbf{P}_j). \tag{5.4}$$

Analogously, another metric commonly used in wireless systems [20, 67] is the squared Fubini-Study (FS) distance

$$d_{\text{FS}}^2(\mathbf{Y}_k, \mathbf{Y}_j) = \prod_{i=1}^{N} \cos^2(\alpha_{kj}(i)) = \prod_{i=1}^{N} \hat{\lambda}_i^{(kj)} = \text{pdet}(\mathbf{P}_k\mathbf{P}_j) \tag{5.5}$$

where $\text{pdet}(\cdot)$ denotes the pseudo-determinant (product of the non-zero eigenvalues). The main advantage of distance (5.4) with respect to other metrics is the fact that it can be computed without the need for eigendecompositions. Furthermore, according to our simulations (see also results in [7]), this distance appears to outperform other metrics in the clustering application considered here.

In summary, the above framework allows us to compare the column subspace spanned by two distinct matrices $\mathbf{Y}_k, \mathbf{Y}_j$ according to the distance between their projection matrices $\mathbf{P}_k, \mathbf{P}_j$ as elements of $\mathcal{G}(N, M)$. Notice that the quantity in (5.4) is closely related to $\hat{d}_M^{(SS)}$ from the previous chapters, specifically, we have that

$$\begin{aligned}
\hat{d}_M^{SS} = \frac{1}{M}\text{tr}\left[(\mathbf{P}_k - \mathbf{P}_j)^2\right] &= \frac{1}{M}\text{tr}\left[\mathbf{P}_k^2 + \mathbf{P}_j^2 - 2\mathbf{P}_k\mathbf{P}_j\right] \\
&= \frac{1}{M}\text{tr}\left[\mathbf{P}_k^2 + \mathbf{P}_j^2\right] - \frac{2}{M}\text{tr}\left[\mathbf{P}_k\mathbf{P}_j\right] \\
&= 2 - \frac{2}{M}\text{tr}\left[\mathbf{P}_k\mathbf{P}_j\right] \\
&= 2 - \frac{2N}{M} + \frac{2}{M}d_{PF}^2
\end{aligned} \tag{5.6}$$

where we use the fact that projection matrices are idempotent, i.e., $\mathbf{P}^2 = \mathbf{P}$. Hence, in order to study the similarity measure

$$s_{kj} = \frac{1}{M}\mathrm{tr}\left[\mathbf{P}_k\mathbf{P}_j\right], \tag{5.7}$$

it is sufficient to study (5.6) and vice versa. To simplify the explanation, we chose to use s_{kj} in the remainder of this chapter when studying the alignment between the column space of two channel matrices.

Theorem 5.1. *Consider two independent (complex) channel matrices $\mathbf{Y}_k \in \mathbb{C}^{M\times N_k}$, $\mathbf{Y}_j \in \mathbb{C}^{M\times N_j}$ defined according to (5.1), assume that M, N_k, N_j increase to infinity at the same rate, so that $M/N_k \to c_k$, $M/N_j \to c_j$ for $c_k, c_j > 1$. Additionally, assume that the spectral norms of $\mathbf{R}_k$ and $\mathbf{R}_j$ are uniformly bounded in M. Then, the projection-Frobenius based similarity s_{kj} in (5.7) converges almost surely to*

$$\bar{s}_{kj}^{PF} = \frac{1}{M}\mathrm{tr}\left[\mathbf{R}_k\mathbf{Q}_k\left(\mu_0^{(k)}\right)\mathbf{R}_j\mathbf{Q}_j\left(\mu_0^{(j)}\right)\right] \tag{5.8}$$

where now, $\mu_0^{(l)} < 0, l = k, j$ is the only negative solution to the equation

$$1 = \frac{1}{N_l}\mathrm{tr}\left[\mathbf{R}_l(\mathbf{R}_l - \mu_0^{(l)}\mathbf{I}_M)^{-1}\right]. \tag{5.9}$$

Furthermore, defining

$$\begin{aligned}(\sigma_{kj}^{PF})^2 = &\left(\mu_0^{(k)}\right)^2\sigma_k^2\left(\mu_0^{(k)},\mu_0^{(k)};\mathbf{R}_j\mathbf{Q}_j\left(\mu_0^{(j)}\right),\mathbf{R}_j\mathbf{Q}_j\left(\mu_0^{(j)}\right)\right)\\ &+\left(\mu_0^{(j)}\right)^2\sigma_j^2\left(\mu_0^{(j)},\mu_0^{(j)};\mathbf{R}_k\mathbf{Q}_k\left(\mu_0^{(k)}\right),\mathbf{R}_k\mathbf{Q}_k\left(\mu_0^{(k)}\right)\right)\\ &+\left(\mu_0^{(k)}\mu_0^{(j)}\right)^2\frac{\mathrm{tr}^2\left[\mathbf{R}_k\mathbf{Q}_k^2(\mu_0^{(k)})\mathbf{R}_j\mathbf{Q}_j^2(\mu_0^{(j)})\right]}{N_kN_j\left(1-\Gamma_k(\mu_0^{(k)})\right)\left(1-\Gamma_j(\mu_0^{(j)})\right)}\end{aligned} \tag{5.10}$$

we have that $(\sigma^{PF})^{-1}M(s_{kl}^{PF} - \bar{s}_{kl}^{PF})$ converges in law to a standard Gaussian random variable.

Proof. The asymptotic quantities $\bar{s}_{kl}^{PF}$ and $(\sigma_{kl}^{PF})^2$ presented here are obtained in a similar manner as the ones in Chapter 2 due to (5.6). □

The characterization of the above quantity has an interest beyond the framework of this thesis and can be used to characterize independence tests based on canonical correlation analysis, which typically use $\mathrm{tr}\left[\mathbf{P}_1\mathbf{P}_2\right]$ as the relevant statistic to determine whether the two sets of observations are statistically independent (see [48–50] for the problem formulation and the asymptotic characterization when

the observations are spatially white; results in this paper extend this characterization to the general spatially colored case). In the remainder of this chapter, we will see how the above line of reasoning can assist in the comparison and clustering of different MIMO channels.

5.3 Hierarchical Clustering

We consider the use of an agglomerative hierarchical clustering approach using one of the *plug in* distances described in Chapter 2. In this bottom–up approach, the goal is to combine different observations into larger collections based on how similar their corresponding subspaces are. Initially, every observation forms a singleton, i.e., a cluster of one observation. The idea is to consecutively merge clusters based on one of the similarity/distance measures described above. At each merging step, the pair with the highest similarity (or, equivalently, the lowest distance) is merged and forms a new cluster. The algorithm finishes whenever a certain number of clusters is reached or the highest similarity falls below a pre-defined threshold.

In order to study the merging problem from the statistical perspective, we will follow the same procedure as in Chapter 4 and formulate each merging step in the hierarchical clustering procedure as a binary hypothesis test. We assume that two channel matrices belong to the same cluster (in terms of proximity of their corresponding column subspaces) if they are generated from the same left covariance matrix, that is

$$\begin{aligned} \mathrm{H}_0(i,j) &: \mathbf{R}_i = \mathbf{R}_j \\ \mathrm{H}_1(i,j) &: \mathbf{R}_i \neq \mathbf{R}_j. \end{aligned} \tag{5.11}$$

If the null hypothesis is accepted, then the new cluster is formed by the concatenation $\mathbf{Y}_{[kj]} = [\mathbf{Y}_k, \mathbf{Y}_j]$, which is equivalent to stacking, side by side, the eigenvectors associated to their respective column subspaces. Notice that this linkage method is distinct from the ones considered in the previous chapters of this thesis. While in the previous chapter we average the contribution of the different channels, here we generate a new channel based on the concatenation of the channel matrices being merged. Moreover, the concatenated matrix $\mathbf{Y}_{[kj]}$ – of dimensions $M \times (N_k + N_j)$ – can still be modeled according to (5.1), where $\mathbf{R}_k = \mathbf{R}_j$ is the common covariance matrix of the new cluster. Naturally, as contributions are averaged, this has little

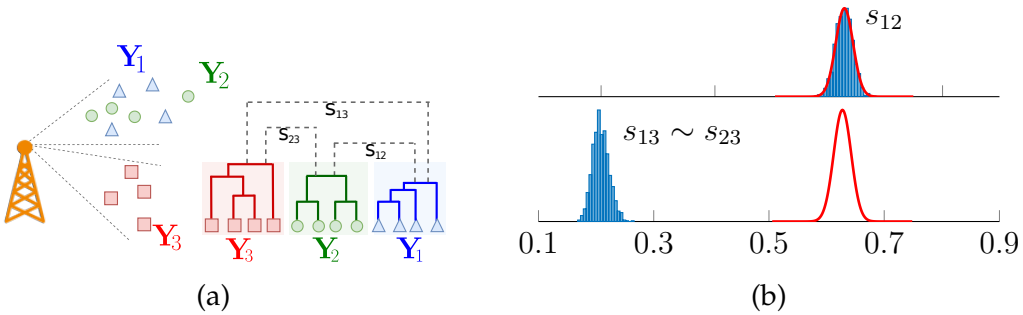


Figure 5.1: Merging point in agglomerative hierarchical clustering where groups have equal number of elements $N_1 = N_2 = N_3 = 4$ and $M = 10$. (a) Groups spread in the spatial domain and their respective dendrogram connectivity. (b) Empirical behavior (represented by the blue histogram) and asymptotic descriptor under the null hypothesis (depicted by the red curves) of similarity measures for various channel realizations within different groups.

impact when considering the sample covariance matrices, however becomes beneficial in the subspace similarity. This, however, becomes beneficial when considering the construction of the projection matrix $\mathbf{P}_k$. Notice that the larger the number of columns $N_k < M$, the better the approximation of the intrinsic subspace is. In the following sections, we will examine this merging procedure according to two scenarios: the first and simpler scenario describes cases where the subspaces (e.g., $\text{span}(\mathbf{Y}_k), \text{span}(\mathbf{Y}_j)$) being compared have the same dimension ($N_k = N_j$), while the second and more general one represents situations where the compared subspaces are non-equidimensional ($N_k \neq N_j$). Particularly, we will tailor our examples to the subspace similarity, but it is worth emphasizing the intuitions provided below are also applicable when considering the general class of distances between sample covariance matrices.

5.3.1 Comparison of Equidimensional Subspaces

We first exemplify the merging process in the specific scenario where every group is of the same size (we will later describe the more general case). Figure 5.1 illustrates this idea: it represents a merging point in the hierarchical algorithm where we compare three groups (denoted here by $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$) with the same number of elements, i.e., $N_1 = N_2 = N_3 = 4$. The number of antennas at the BS stations is $M = 10$. At this point, the clustering algorithm needs to decide which pair of groups to merge. To do so, it needs to choose – among the three possible options

$(\mathbf{Y}_1, \mathbf{Y}_2)$, $(\mathbf{Y}_1, \mathbf{Y}_3)$ and $(\mathbf{Y}_2, \mathbf{Y}_3)$ – the pair with the highest similarity or, alternatively, the pair with the smallest distance.

Let us assume $\mathbf{R}_1 = \mathbf{R}_2 \neq \mathbf{R}_3$. In this scenario, because $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are associated to the same left covariance (i.e., $\mathbf{R}_1 = \mathbf{R}_2$), the similarity measure between them should be the highest, so that we should in principle have $s_{12} > s_{13}$ and $s_{12} > s_{23}$ with very high probability. Figure 5.1(a) visually represents this scenario in the spatial domain (left) and in a hierarchical structure (right). Indeed, by comparing different realizations of users' channels[2] ($\varphi_1 = \varphi_2 = 45°$ and $\varphi_3 = 60°$), which belong to the groups $g = 1, 2, 3$, and analyzing the empirical distribution of their subspace similarities (blue histograms in Figure 5.1(b)), we notice that the similarity s_{12}, between $(\mathbf{Y}_1, \mathbf{Y}_2)$, is always higher than the similarity between any of the other possible pairs $(\mathbf{Y}_1, \mathbf{Y}_3)$ and $(\mathbf{Y}_2, \mathbf{Y}_3)$. Moreover, we observe that, because $N_1 = N_2$ and $\mathbf{R}_1 = \mathbf{R}_2$, we have that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ have equal subspace structures. Hence, the other two statistics s_{13} and s_{23} will have the same distribution, i.e., $\mathbb{E}[s_{13}] = \mathbb{E}[s_{23}]$ and $\mathrm{var}[s_{13}] = \mathrm{var}[s_{23}]$. This statistical behavior is illustrated by the overlapping histograms in the second line of Figure 5.1(b) and denoted by $s_{13} \sim s_{23}$.

We also highlight that the solid lines in Figure 5.1(b) represent the theoretical probability density function (PDF) of the similarity (5.7). These PDFs are obtained from Theorem 5.1 under the assumption $\mathbf{R}_k = \mathbf{R}_j$, which corresponds to the null hypothesis in (5.11). By employing this method, when the null hypothesis is valid, we obtain a perfect match between the theoretical PDF and its associated empirical distribution (first line of Figure 5.1(b)). Conversely, when in the alternative hypothesis, there exists a mismatch between the theoretical PDF and its associated empirical distribution. A similar behavior can also be obtained using any of the *plug in* distances, hence omitted here.

5.3.2 Non-equidimensional Subspaces

One of the main problems with the use of the similarity measure in (5.7) or, generally, any of the *plug in* distances, is that their statistics are inherently dependent on the pair of observations (groups) that are being compared, specially with regard to the dimensionality of the associated subspaces [7]. This problem is a direct consequence of the fact that the distance measures are substantially related to the dimensionality of the underlying manifold. For instance, in the SS case, by comparing

[2]See Section 5.1 for details on the modeling of the true covariance matrices.

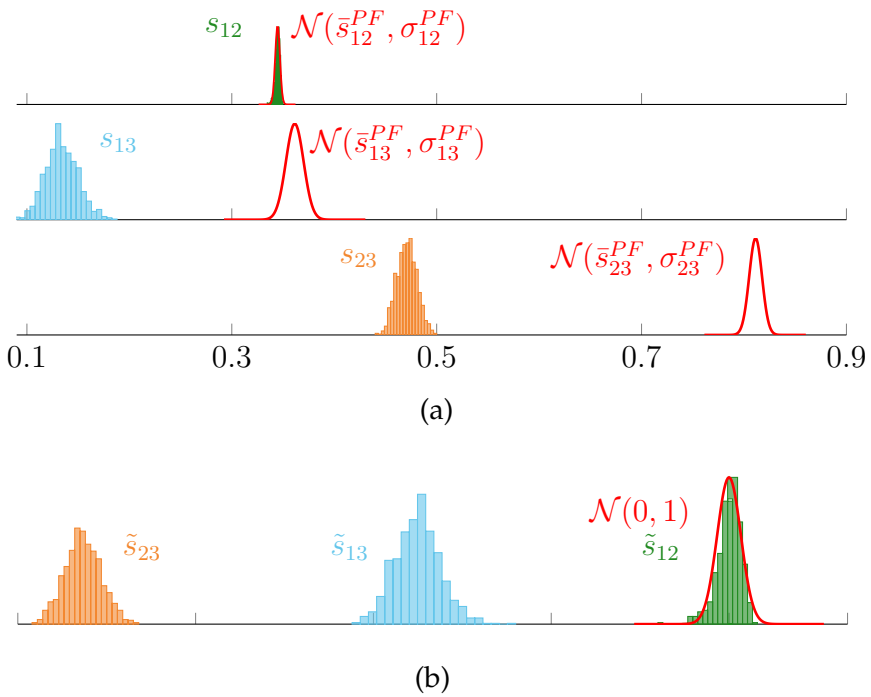


Figure 5.2: Behavior of similarity measure for comparison of non-equidimensional subspaces of dimensions $N_1 = 4, N_2 = 24, N_3 = 32$ before (a) and after (b) normalization with respect to the null hypothesis, for $M = 50$.

subspaces of different dimensions (non-equidimensional subspaces) one is implicitly comparing measures among points defined in different Grassmann manifolds. Hence, in the more generic case, where each group has a different number of observations (i.e., $N_1 \neq N_2 \neq N_3$), a direct application of any of the *plug in* distances might potentially lead to a misleading comparison metric.

To illustrate this idea, let us again consider the subspace similarity, but now the scenario where we have the projections $\mathbf{P}_1 \in \mathcal{G}(N_1, M)$, $\mathbf{P}_2 \in \mathcal{G}(N_2, M)$ and $\mathbf{P}_3 \in \mathcal{G}(N_3, M)$. Recall that the projection $\mathbf{P}_\ell$ is closely related to the subspace $\text{span}(\mathbf{Y}_\ell)$ and to its dimensionality N_ℓ. Moreover, because $N_1 \neq N_2 \neq N_3$, we also have that

$$\mathcal{G}(N_1, M) \neq \mathcal{G}(N_2, M) \neq \mathcal{G}(N_3, M).$$

Hence, in this new scenario, the measures s_{12}, s_{23} and s_{13} are no longer comparable. Figure 5.2(a) represents the same results as in Figure 5.2(b) but for the case where $N_1 = 4, N_2 = 24, N_3 = 32$ and $M = 50$. Based only on the comparisons of the similarities s_{12}, s_{13} and s_{23} (represented by histograms), one would wrongly conclude the pair $(\mathbf{Y}_1, \mathbf{Y}_3)$ to be the merge of choice, i.e., their similarity seems to be the largest one when compared to the other possible pairs.

Notice that, whenever $N_k \neq N_j$, the similarity measure s_{kj} only considers the first $\min(N_k, N_j)$ principal angles. Consequently, comparisons between two large clusters (e.g., $\mathbf{Y}_2$ and $\mathbf{Y}_3$) may yield a higher similarity than comparisons between a small and a large cluster (e.g., $\mathbf{Y}_1$ and $\mathbf{Y}_2$), regardless of the true group assignments. As illustrated in Figure 5.2(a), even if s_{12} follows the theoretical PDF obtained under the null hypothesis (red curve), the highest similarity turns out to be between clusters $\mathbf{Y}_2$ and $\mathbf{Y}_3$. This is clearly undesirable. To ensure accurate clustering, the algorithm should prioritize merges between elements belonging to the same group, i.e., those conforming to the theoretical PDF associated with the null hypothesis. Similar conclusions can be derived by considering any of the *plug in* distances between covariance matrices. In this scenario, this undesirable behavior may occur as a consequence of comparing two sample covariance matrices obtained from a large amount of samples against two other estimators obtained from a small amount of samples.

To circumvent the problem of non-equidimensional subspace comparison, in [7], we showed that, when considering the projection-Frobenius based similarity s_{kj} in (5.7), it is useful to consider the normalized measurement

$$\tilde{s}_{kj} = \frac{s_{kj} - \bar{s}_{kj}^{PF}}{\sigma_{kj}^{PF}} \tag{5.12}$$

where $\bar{s}_{kj}^{PF}$ and $\left(\sigma_{kj}^{PF}\right)^2$ are defined as in Theorem 5.1, but tailored to the null hypothesis in (5.11), namely $\mathbf{R}_k = \mathbf{R}_j$. The idea is to implicitly promote larger values for comparisons under the null hypothesis while penalizing comparisons under the alternative hypothesis. The main advantage of the proposed metric space is the fact that all similarities are asymptotically comparable regardless of the inherent subspace dimensions and that it benefits comparisons under the null hypothesis, i.e., when the subspaces are generated with the same covariance matrix.

As a result, this normalized measure allows us to effectively compare the degree of alignment of multiple non-equidimensional subspaces. Figure 5.2(b) exemplifies this case when comparing $\mathbf{Y}_1$ and $\mathbf{Y}_2$ in $\tilde{s}_{12}$ (green histogram). In the cases where two clusters have different spatial covariance matrices – e.g., $(\mathbf{R}_1, \mathbf{R}_3)$ and $(\mathbf{R}_2, \mathbf{R}_3)$ – their normalized similarities – $\tilde{s}_{13}$ (orange histogram) and $\tilde{s}_{23}$ (blue histogram), respectively – are moved far away from the standard normal distribution (red curve). This effect is visually represented by the shifting of the distributions over the x-axis. As a conclusion, notice that, in this normalized metric space,

groups that have the same covariance matrix will usually have higher similarity than groups with different covariance matrices, therefore allowing the comparison of subspaces of different dimensions. Moreover, because each set of UE is associated to a subspace, we can directly infer the UEs clustering association based on their subspaces cluster assignments.

Finally, notice that a similar behavior could also be obtained by considering the corrected term $s_{kj} - \bar{s}_{kj}^{PF}$. The main difference between $\tilde{s}_{kj}$ in (5.12) and this corrected term is that $\tilde{s}_{kj}$ is normalized by the variance, so that one can also perform other statistical analysis to this quantity, such as the one described in Chapter 4. Let us recall that to solve the problem of comparing non-equidimensional subspaces, it is sufficient to properly shift the quantities s_{12}, s_{13}, s_{23} along the x-axis so that the one that is associated to the correct merge (s_{12} in our example) can be selected. The problem of rearranging the quantities in the x-axis, thus selecting the correct cluster, can be tackled by considering the correction terms $s_{kl} - \bar{s}_{kl}^{PF}$ which does not depend on the variance. Before exploring the numerical implications of this idea (see Section 5.5), in the following section, we build upon the results above and present a correction term, denoted by $\hat{\bar{d}}_M$, which can be directly derived from the data. Notice that this is essentially different from what we have been doing so far. Specifically, in Chapter 3, we proposed consistent estimators for the distance between sample covariance matrices, which were consistent under both null and alternative hypotheses but primarily applicable to the oversampled regime (at least for the KL and LE distances). In the remainder of this chapter, we propose a different approach which is specifically designed for the undersampled regime and the *plug in* distances. It consists in estimating the asymptotic equivalent of $\hat{d}_M$ under the null hypothesis and creating the corrected term denoted as $\hat{d}_M - \hat{\bar{d}}_M$ which can be obtained directly from the channel matrices.

5.4 Correction Terms Under $\mathbf{R}_1 = \mathbf{R}_2$

Let us start by observing that the construction of the normalized statistic in (5.12) requires the perfect knowledge of the covariance matrices of the channels that are being compared. In low dimensional scenarios (small M and large N), this is typically approximated by computing the required parameters with the sample covariance matrix instead of the true one. However, in large dimensional scenarios (M and N large and comparable) this is far from optimal, mainly because the sample covariance matrix can hardly be regarded as a consistent estimate of the true

one. In this section, we will propose estimators for the asymptotic equivalents $\bar{d}_M^{KL}, \bar{d}_M^{E}, \bar{d}_M^{SS}$ under the null hypothesis, which can be obtained directly from the channel (or observation) matrices. To improve clustering, these correction terms are tailored to correctly approximate their asymptotic equivalents under the null hypothesis while penalizing comparisons under the alternative hypothesis. For visualization purposes, these asymptotic equivalents, derived in Chapter 2 using the true covariance matrices $\mathbf{R}_k, \mathbf{R}_j$, are recalled in Table 5.1. In what follows, we will propose estimators to each of them.

Table 5.1: Asymptotic equivalents of KL, E and SS *plug in* distances in the undersampled regime ($N_k, N_j < M$).

Name	**Asymptotic Equivalents of $\hat{d}_M$ ($N_k, N_j < M$)**
Euclidean distance	$\frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_k - \mathbf{R}_j)^2\right] + \frac{1}{MN_1}\mathrm{tr}^2[\mathbf{R}_k] + \frac{1}{MN_2}\mathrm{tr}^2[\mathbf{R}_j]$
Subspace Similarity	$\frac{1}{M}\mathrm{tr}\left[\mathbf{R}_k\mathbf{Q}_k\left(\mu_0^{(k)}\right)\mathbf{R}_j\mathbf{Q}_j\left(\mu_0^{(j)}\right)\right]$
Symmetrized KL divergence	$\frac{\mathrm{tr}\left[\mathbf{R}_k\mathbf{Q}_k^2(\mu_0^{(k)})\mathbf{R}_j\right]}{2M\left(1-\Gamma_k\left(\mu_0^{(1)}\right)\right)} + \frac{\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j^2(\mu_0^{(j)})\mathbf{R}_k\right]}{2M\left(1-\Gamma_j\left(\mu_0^{(j)}\right)\right)} - 1$

5.4.1 Euclidean distance

Let us start by the Euclidean distance. Notice that, by definition when $\mathbf{R}_k = \mathbf{R}_j$, we can re-write $\bar{d}_M^E$ as

$$\bar{d}_M^E = \frac{1}{MN_k}\mathrm{tr}^2[\mathbf{R}_k] + \frac{1}{MN_j}\mathrm{tr}^2[\mathbf{R}_j].$$

Moreover, using the tools described in Section 3.1.1 of this thesis, we have that

$$\frac{1}{MN_k}\mathrm{tr}^2[\mathbf{R}_k] \asymp \frac{1}{MN_k}\mathrm{tr}^2[\hat{\mathbf{R}}_k]$$

where the equivalence should be understood as in Remark 3.1 of Chapter 3, that is, for two Hermitian matrices $\mathbf{B}_M$ and $\mathbf{C}_M$ and two analytic functions $f, h : \mathbb{C} \to \mathbb{C}$ applied to the eigenvalues of these matrices, we write $f(\mathbf{B}_M) \asymp h(\mathbf{C}_M)$ if

$$\frac{1}{M}\mathrm{tr}\left[\mathbf{A}_M\left(f(\mathbf{B}_M) - h(\mathbf{C}_M)\right)\right] \to 0$$

almost surely as $M \to \infty$, where $\mathbf{A}_M$ is any sequence of deterministic $M \times M$ matrices with bounded norm.

The above allows us to write the correction term of the *plug in* Euclidean distance as

$$\hat{\bar{d}}_M^E = \frac{1}{MN_k}\mathrm{tr}^2[\hat{\mathbf{R}}_k] + \frac{1}{MN_j}\mathrm{tr}^2[\hat{\mathbf{R}}_j].$$

It is worth noticing that the corrected quantity $\hat{d}_M^E - \hat{\bar{d}}_M^E$ is equivalent[3] to the definition of the consistent estimator $\tilde{d}_M^E$ introduced in Section 3.1.1 of Chapter 3. We recall that, for $\mathbf{R}_1 = \mathbf{R}_2$, the consistent estimator $\tilde{d}_M^E \to 0$ and, for $\mathbf{R}_1 \neq \mathbf{R}_2$, $\tilde{d}_M^E \to M^{-1}\mathrm{tr}[(\mathbf{R}_1 - \mathbf{R}_2)^2]$ (i.e., it tends to the true distance between covariance matrices d_M^E). Then, as the corrected term $\hat{d}_M^E - \hat{\bar{d}}_M^E$ is equivalent to the consistent estimator $\tilde{d}_M^E$, they will also follow the same behavior meaning that, under the null hypothesis, $\hat{d}_M^E - \hat{\bar{d}}_M^E$ correctly approaches zero and, under the alternative hypothesis, the corrected term tends to some positive value which is bounded away from zero.

Finally, we also observe that both quantities share the same CLT, with $\tilde{\mathfrak{m}}_M^E = 0$ (due to the complex nature of the observations, i.e., $\varsigma = 0$) and

$$\tilde{\sigma}_M^2 = 2\left(\frac{1}{N_k}\mathrm{tr}\left[\mathbf{R}_k^2\right]\right)^2 + 2\left(\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j^2\right]\right)^2 + 4\frac{1}{N_kN_j}\mathrm{tr}^2\left[\mathbf{R}_k\mathbf{R}_j\right].$$

then, by using once more the results from Section 3.1.1, we readily obtain the estimators

$$\begin{aligned}\hat{\tilde{\sigma}}_M^2 =& 2\left(\frac{1}{N_k}\mathrm{tr}\left[\hat{\mathbf{R}}_k^2\right] - \frac{1}{N_k^2}\mathrm{tr}^2\left[\hat{\mathbf{R}}_k\right]\right)^2 + 2\left(\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j^2\right] - \frac{1}{N_j^2}\mathrm{tr}^2\left[\hat{\mathbf{R}}_j\right]\right)^2 \\ &+ 4\frac{1}{N_kN_j}\mathrm{tr}^2\left[\hat{\mathbf{R}}_k\hat{\mathbf{R}}_j\right].\end{aligned} \tag{5.13}$$

Obviously, the three terms are positive, so that in order to show that $\liminf_{M\to\infty} \hat{\tilde{\sigma}}_M^2 > 0$ it is sufficient to see that any of these are bounded away from zero. In particular, the eigenvalues of $\hat{\mathbf{R}}_k, \hat{\mathbf{R}}_j$ are located inside a compact subset of $\mathbb{R}^+$ independent of M, and one trivially sees that the last term is bounded away from zero.

5.4.2 Subspace similarity

The proposed estimator of the subspace similarity, which is defined under the null hypothesis, is denoted as $\hat{\bar{s}}_{k,j}$ and can be expressed (see Appendix C.1 for details)

[3] This is not a general rule. The consistent estimator will often have different a CLT than its associated corrected term.

by

$$\hat{\bar{s}}_{kj}^{PF} = \begin{cases} \dfrac{N_k}{2M}\left(1 - \dfrac{\hat{\kappa}_k^2(1)}{\hat{\kappa}_k(2)}\right) + \dfrac{N_j}{2M}\left(1 - \dfrac{\hat{\kappa}_j^2(1)}{\hat{\kappa}_j(2)}\right) & N_k = N_j \quad (5.14) \\ \dfrac{1}{2M}\left(\dfrac{N_k\hat{\upsilon}_k(k) - N_j\hat{\upsilon}_j(k)}{\hat{\upsilon}_k(k) - \hat{\upsilon}_j(k)} + \dfrac{N_k\hat{\upsilon}_k(j) - N_j\hat{\upsilon}_j(j)}{\hat{\upsilon}_k(j) - \hat{\upsilon}_j(j)}\right) & N_k \neq N_j \quad (5.15) \end{cases}$$

where we have defined, for $l \in \{k, j\}$,

$$\hat{\kappa}_l(m) = \frac{1}{N_l}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_l^{\#}\right)^m\right],\ m \in \mathbb{N}$$

with $(\cdot)^{\#}$ denoting the Moore-Penrose pseudoinverse. Furthermore, for $N_k = N_j$, we have $\hat{\upsilon}_\ell(l) = -\hat{\kappa}_\ell^{-1}(1)$, whereas for $N_k \neq N_j$ we take

$$\hat{\upsilon}_j(k) = \gamma\left(1 - \frac{N_j}{N_k}\right)$$

where γ is the smallest solution to

$$\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_k\left(\hat{\mathbf{R}}_k - \gamma\mathbf{I}_M\right)^{-1}\right] = 1.$$

This distinction is particularly useful to penalize the alternative hypothesis $\mathrm{H}_1(k, j)$ based on $\hat{\upsilon}_j$ and $\hat{\mathbf{R}}_j^{\#}$ by relating the smaller sample eigenvalue distribution in terms of the larger one.

The estimator of the asymptotic variance $(\sigma_{kj}^{PF})^2$ is denoted by $(\hat{\sigma}_{kj}^{PF})^2$ and can also be derived in a similar manner, but, as it is not directly applied to this thesis (see Remark 5.3), we leave its description in the Appendix C.2. Nevertheless, it's worth emphasizing that this estimator for the second-order moment holds potential relevance in various domains. One notable example is its applicability to sensor fusion methods, commonly utilized in signal processing, where it facilitates the combination of different observations or measurements based on their respective variances [69]. This, however, is outside the scope of this thesis and is left for future work.

Remark 5.3. *After having introduced the derivations above, one can rush into the conclusion that they can follow similar steps as the one conducted for the Euclidean distance and simply plug the estimators $\hat{\bar{s}}_{kj}^{PF}$ and $(\hat{\sigma}_{kj}^{PF})^2$ directly into (5.12). Unfortunately, this is not possible for the subspace similarity. When applying the new estimator of the deterministic equivalent $\hat{\bar{s}}_{kl}^{PF}$ under the null hypothesis, the corrected value $s_{kl} - \hat{\bar{s}}_{kl}^{PF}$ will produce a new*

statistic that behaves differently from the correction based on its true asymptotic counterpart $s_{kl} - \bar{s}_{kl}^{PF}$, meaning that these random quantities will have different asymptotic variances. In other words, it is important to emphasize that, despite

$$s_{kl} - \bar{s}_{kl}^{PF} \sim \mathcal{N}\left(0, \left(\sigma_{kj}^{PF}\right)^2\right),$$

in general, when using the first order mean estimator for s_{kl}^{PF} under $\mathrm{H}_0(k,l)$, we have that

$$s_{kl} - \hat{\bar{s}}_{kl}^{PF} \not\sim \mathcal{N}\left(0, \left(\sigma_{kj}^{PF}\right)^2\right)$$

meaning that the corrected quantity $s_{kl} - \hat{\bar{s}}_{kl}^{PF}$ cannot be described by neither of the variances discussed in this chapter. That is because the first order estimator $\hat{\bar{s}}_{kl}^{PF}$ is directly obtained from the data and may vary depending on the channel realization which, in turn, will directly affect the distribution of the corrected value $s_{kl} - \hat{\bar{s}}_{kl}^{PF}$. Notice, however, that we can still estimate $(\sigma_M^{(PF)})^2$ by $(\hat{\sigma}_M^{(PF)})^2$, but this estimator only holds for s_{kl} and $s_{kl} - \bar{s}_{kl}^{PF}$.

5.4.3 Symmetrized KL divergence

Under the null hypothesis $\mathbf{R}_k = \mathbf{R}_j = \mathbf{R}$, the asymptotic equivalent of $\hat{d}_M^{KL}$ becomes

$$\bar{d}_M^{KL} = \frac{\mathrm{tr}\left[\mathbf{R}^2\mathbf{Q}_k^2(\mu_0^{(k)})\right]}{2M\left(1-\Gamma_k\left(\mu_0^{(k)}\right)\right)} + \frac{\mathrm{tr}\left[\mathbf{R}^2\mathbf{Q}_j^2(\mu_0^{(j)})\right]}{2M\left(1-\Gamma_j\left(\mu_0^{(j)}\right)\right)} - 1$$

where we recall that, under the undersampled ($N_j < M$) regime, $\mu_0^{(j)}$ is the only negative solution to the equation (respectively to k)

$$0 = \mu^{(j)}\left(1 - \frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j\left(\mu^{(j)}\right)\right]\right). \tag{5.16}$$

Then, by using the results from Appendix C, and after some algebra, we readily obtain

$$\hat{\bar{d}}_M^{KL} = \frac{N_k}{2M}\left(\frac{N_j^{-1}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^2\right]}{\left(N_j^{-1}\mathrm{tr}\left[\hat{\mathbf{R}}_j^{\#}\right]\right)^2} - 1\right) + \frac{N_j}{2M}\left(\frac{N_k^{-1}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_k^{\#}\right)^2\right]}{\left(N_k^{-1}\mathrm{tr}\left[\hat{\mathbf{R}}_k^{\#}\right]\right)^2} - 1\right) - 1.$$

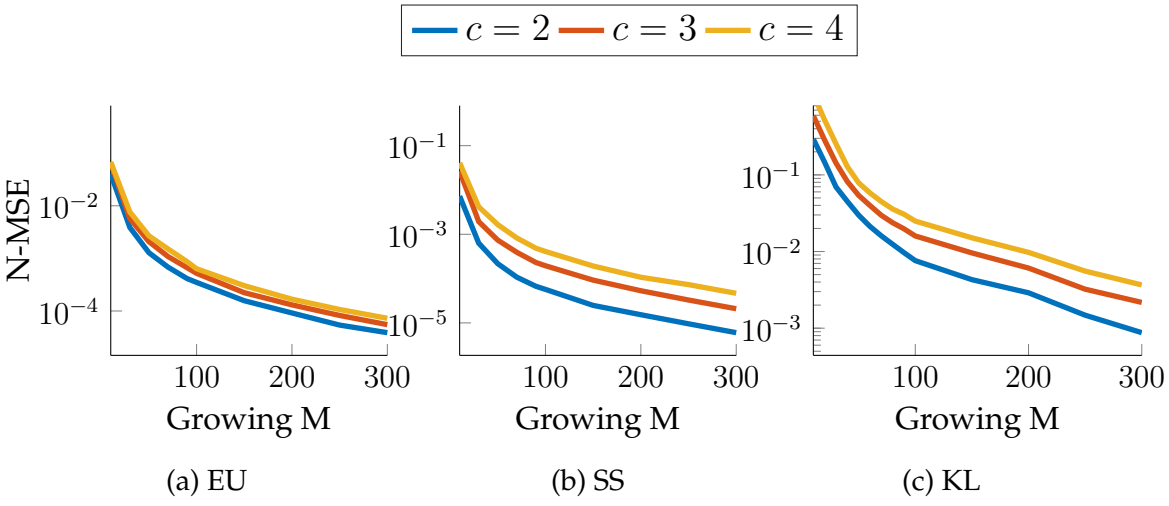
5.4.4 Consistency of Correction Terms

Before delving into the numerical implications of the correction terms described above, in this section we analyze (under the null hypothesis) how accurate these

correction terms are. We will assess the asymptotic consistency of our proposed estimators by comparing the Normalized Mean Squared Error (N-MSE) between the correction terms $\hat{\bar{d}}_M$ and their asymptotic equivalents $\bar{d}_M$ for each of the metrics considered in this chapter, namely, SS, EU and symmetrized KL divergence. The N-MSE for each of these correction terms is defined as

$$\varepsilon_{\text{cor}} = \hat{\mathbb{E}}\left[\left(\frac{\hat{\bar{d}}_M - \bar{d}_M}{\bar{d}_M}\right)^2\right]$$

where the empirical expectation ($\hat{\mathbb{E}}[\cdot]$) is averaged over a large number of different channel realizations. In what follows, we will consider the covariance matrices $\mathbf{R}_1 = \mathbf{R}_2$, as described in Section 5.1. These matrices both correspond to angle $\varphi_1 = \varphi_2 = 30°$ with angular spread of $\delta_\varphi^2 = 15°$. We will use these covariance matrices to define $\bar{d}_M$ (under the null hypothesis) and generate $S = 10^3$ realization of channel matrices, namely $\mathbf{Y}_1^{(i)} \in \mathbb{C}^{M\times N_1}$, $\mathbf{Y}_2^{(i)} \in \mathbb{C}^{M\times N_2}$, $i = 1, \ldots, S$. Notice that, by construction (see definition (5.1)), these channel matrices are not equal to one another due to the fact that $\mathbf{X}_1^{(i)} \neq \mathbf{X}_2^{(i)}$ and $\mathbf{X}_l^{(i)} \neq \mathbf{X}_l^{(j)}$, for $l = 1, 2$ and $i \neq j$. These channel matrices will be subsequently used to build the (estimated) correction term $\hat{\bar{d}}_M$.



(a) EU (b) SS (c) KL

Figure 5.3: Measured N-MSE (y-axis) between asymptotic and estimators descriptors for growing number of antennas (x-axis).

Figure 5.3 shows the N-MSE for the asymptotic estimations of $\bar{s}_M^{PF}$, $\bar{d}_M^E$ and $\bar{d}_M^{KL}$. Note that, for every metric and constant $M/N_1 = M/N_2 = c$, as $M \to \infty$ (x-axis), the N-MSE (y-axis) between the estimated corrections and their respective asymptotic equivalents tends to zero, meaning that they become very close or equal to

one another. Moreover, we note that the closer $N_j, j = 1, 2$ is from M the easier it becomes to reconstruct the intrinsic subspace or, alternatively, to estimate the true covariance matrix $\mathbf{R}$. Hence, in our scenario, the closer the number of antennas at the receiver M is to the number of antennas at the transmitters $N_j, j = 1, 2$, the faster this convergence happens. This convergence is visually depicted in the figure, with smaller values of c corresponding to a faster convergence rate. These results corroborate the accuracy of the estimators proposed in this chapter, demonstrating that they are reliable approximations of their asymptotic counterparts.

5.5 Clustering of MIMO Channels

In order to validate the results presented above, we assess the clustering performance of the proposed corrected measures $\hat{d}_M - \hat{\bar{d}}_M$. Here $\hat{d}_M$ can be any of the metrics discussed in this chapter, namely SS distance (equivalently $\hat{s}_{kj}^{PF}$ and $\hat{\bar{s}}_{kj}^{PF}$ for similarity), EU distance, or symmetrized KL divergence. Moreover, $\hat{\bar{d}}_M$ represents the correction term derived as per its respective definition from Section 5.4. The performance of the proposed measures are denoted as "COR" in the figures and are represented by dashed lines. The original (non-corrected) statistics are depicted using the same color but are displayed as a solid line and are denoted as "TRAD" in the figures. We also compare the performance against four other more common measures: the Projection–Frobenius similarity (denoted here by "PF"), the Fubini–Study based similarity [68] ("FS"), together with the minimal ($d_{\text{C–S}}$) and average ($d_{\text{C–AVG}}$) columnwise cosine distance between two clusters $\mathbf{Y}_1$, $\mathbf{Y}_2$. Specifically, the minimal columnwise cosine distance $d_{\text{C–S}}$ is defined as the minimum between every column pair $\mathbf{y}_1(i)$, $\mathbf{y}_2(j)$. Similarly, the average columnwise distance is defined as

$$d_{\text{C–AVG}} = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \left(1 - \frac{\left| \mathbf{y}_1^{\text{H}}(i) \mathbf{y}_2(j) \right|}{||\mathbf{y}_1(i)||_2 ||\mathbf{y}_2(j)||_2} \right)^2$$

where $\mathbf{y}_i(k)$ represents the kth column of matrix $\mathbf{Y}_i$. For all these distances, we consider an agglomerative hierarchical clustering algorithm that, at each step, merges the pair of clusters with highest similarity or, equivalently, smallest distance.

Moreover, rather than characterizing the behavior of the whole hierarchical clustering method, we focus here on individual merging steps in the clustering process. More specifically, we evaluate a simplified scenario with three clusters (set of channel matrices), $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ and assume that $\mathbf{Y}_1, \mathbf{Y}_2$ are generated with the same left covariance matrix ($\mathbf{R}_1 = \mathbf{R}_2$), and should therefore be merged into a

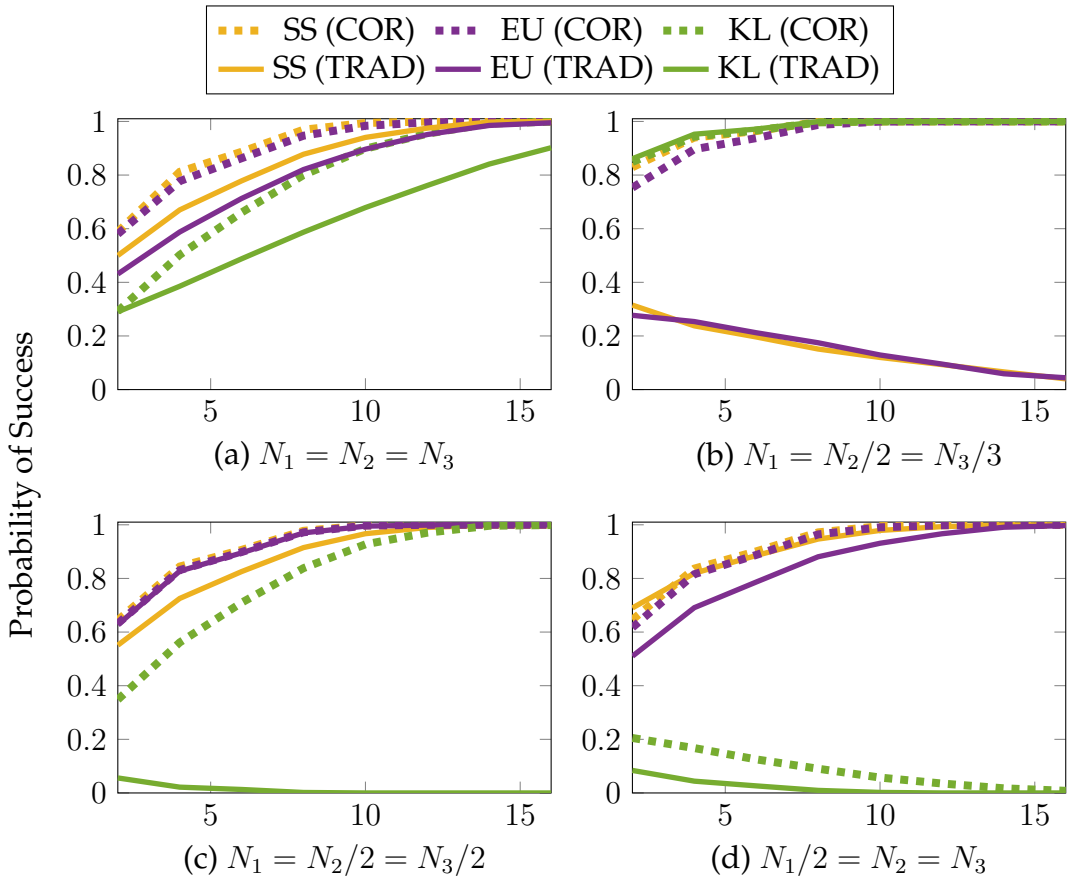


Figure 5.4: Comparison of probability of success for corrected and non-corrected *plug in* metrics in four different scenarios (a)-(d) with respect to the growth of N_1 (x–axis).

(new) single cluster. The left covariance matrix of the third observation is different from the other two ($\mathbf{R}_3 \neq \mathbf{R}_1$). Notice that the hierarchical clustering of K groups boils down to $\frac{K!}{(K-3)!3!}$ triplet comparisons. In this sense, analyzing the merging of pairs based on triplets is closely related to the behavior of the hierarchical clustering in large scenarios, i.e., $K > 3$. Moreover, by comparing clusters of different sizes, we are essentially simulating different levels of the hierarchical clustering. Finally, for each of the metrics described above, the algorithm first computes the three similarity measures between pairs of matrices, and chooses to merge the pair that has the highest similarity. We define the probability of success (POS) as the probability of making the right merge and evaluate it by considering a collection of 10^3 realizations of these clusters.

Particularly, we will simulate the scenario where the first two groups are associated to $\varphi_1 = \varphi_2 = 30°$ and the third one to $\varphi_3 = 60°$. Moreover, we will always

consider the case where $N_1, N_2, N_3 < M$ and $M = N_1+N_2+N_3-\min(N_1, N_2, N_3)+1$. The first condition simulates the undersampled regime, while the second one ensures it is not possible to spatially separate UEs and hence, a clustering solution is desired. Figure 5.4 illustrates the POS obtained in four distinct scenarios for different choices of the dimensions (M, N_1, N_2, N_3), where in all cases these quantities are taken to increase proportionally in the x–axis. Notice that in all four scenarios, the proposed corrected metric (dashed lines) generally outperform or equals their respective non-corrected measure (solid lines) especially in situations where the compared subspaces have very different dimensions. The proposed corrected metrics usually outperform their non-corrected counterpart in the case where the traditional metric cannot properly distinguish between a correct merge and a wrong one this is illustrated, for instance, when considering the SS and EU distances in Figure 5.4(b) or the KL distance in Figure 5.4(c). Unfortunately, there exist scenarios, for instance the one depicted in Figure 5.4(d), in which also the corrected symmetrized KL divergence fails to correctly select the two clusters that should be merged together. It turns out that, in some cases, $\hat{\bar{d}}_M^{KL}$ is also a close approximation of the statistic $\hat{d}_M^{KL}$ even under the alternative hypothesis. As a consequence, instead of separating the two PDFs associated to the null and alternative hypothesis, the corrected term $\hat{d}_M^{KL} - \hat{\bar{d}}_M^{KL}$ brings them close together[4], i.e., center both in zero. Nonetheless, this is primarily present when considering the symmetrized KL divergence and over specific scenarios.

We also compare the proposed corrected metrics against other traditional metrics, namely FS, C-S and C-AVG. These traditional metrics can serve as baselines to our proposed corrected metrics. This comparison is depicted in Figure 5.5 for the same scenarios as in Figure 5.4. Notice that in all four scenarios, the proposed corrected metric of the EU and SS metrics generally outperform the traditional metrics, especially in situations where the compared subspaces have very different dimensions. Furthermore, we observe that a similar behavior is observed even in the scenarios with a relatively low number of dimensions, in spite of the fact that the statistic is designed to perform well in large dimensional settings.

Finally, we also study the relationship between the mean angular distance in the covariance generation $\Delta\varphi_{13} = |\varphi_1 - \varphi_3|$ and the POS for each of the considered similarity measures. Notice that this is the same as moving the elements from

[4]It is worth noting that there might exist some alternative definition for $\bar{d}_M^{KL}$ and $\hat{\bar{d}}_M^{KL}$ which further penalizes the alternative hypothesis. The specific study of such alternatives is left as future work.

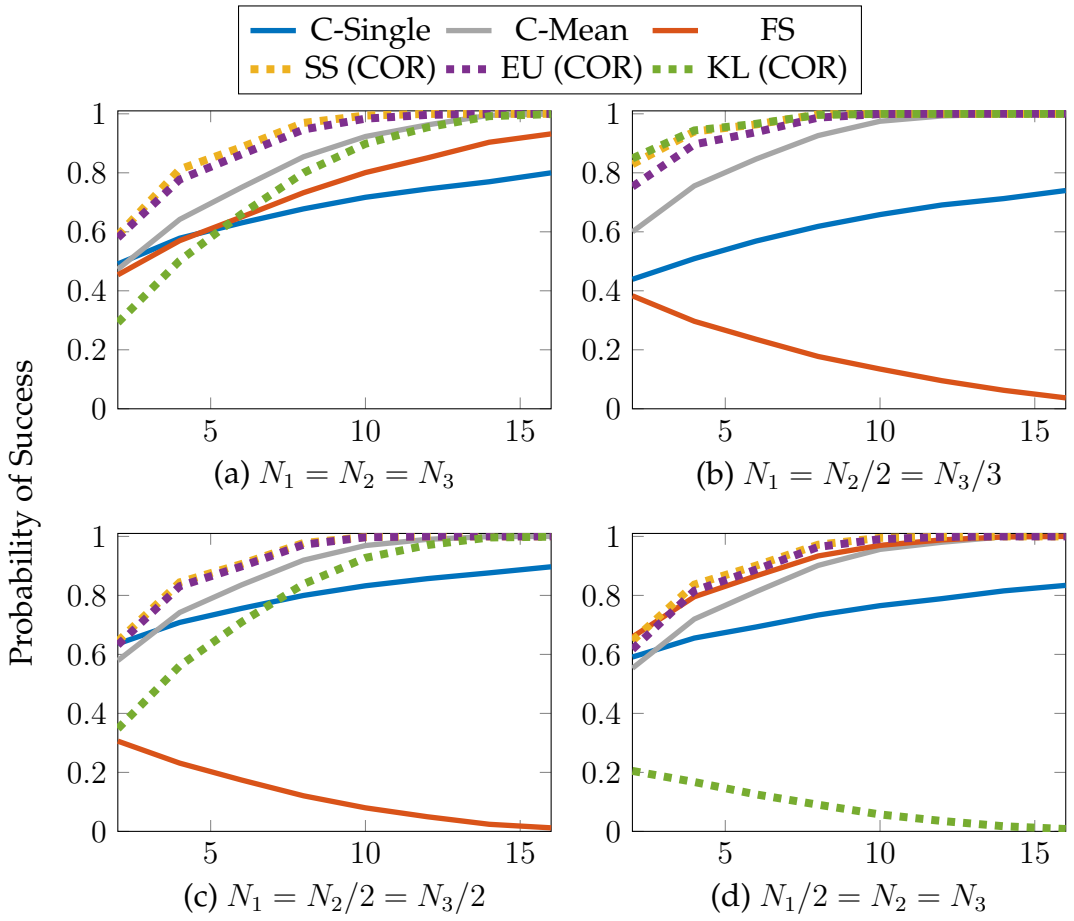


Figure 5.5: Probability of success related to the different metrics in four different scenarios (a)-(d) with respect to the growth of N_1 (x–axis).

the desired merging groups $\mathbf{Y}_1, \mathbf{Y}_2$ far away from the other group $\mathbf{Y}_3$. Figure 5.6 illustrates the relationship between the mean angular distance in the covariance generation $|\varphi_1 - \varphi_3|$, for $\varphi_1 = 20°$, $M = 10$ and the POS for each of the considered similarity measures. Observe that, once again, both normalized metrics outperform all the other metrics, especially in the region where the two left covariance matrices are close, which corresponds to the region where $\Delta\varphi$ is small.

5.6 Conclusions

An agglomerative hierarchical clustering of UEs has been analyzed in this chapter. Particularly, we have considered merging elements that are similar to each other based on their intrinsic subspaces. We have discussed two main intuitions: the first, which results from the direct comparison of the sample covariance matrices

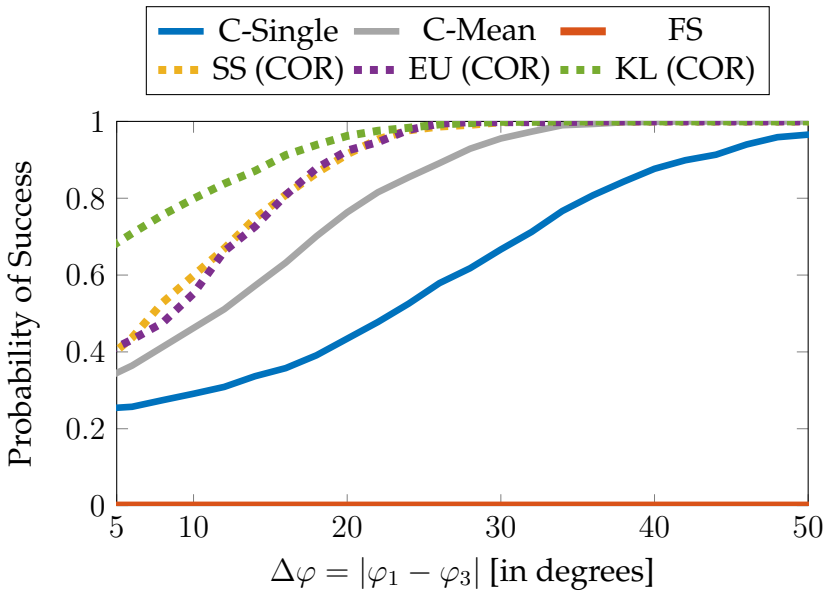


Figure 5.6: Probability of success for different $\Delta\varphi$. The dimensions of each observation can be described by $N_1/2 = N_2/2 = N_3 = 3$.

associated to each channel matrices; and the second one, which has been the focus of this chapter, and considers how well aligned the subspaces spanned by their channel matrices are. Naturally, one is closely related to the other. Moreover, it has been shown that by using the asymptotic behavior of the statistic $\hat{d}_M$ under the null hypothesis $\mathrm{H}_0(i,j)$, one can re-normalize this measure into a corrected equivalent, namely $\hat{d}_M - \hat{\bar{d}}_M$, such that the comparison between channels or group of channels with different dimensionality becomes possible. This correction measurement relies solely on the observations and their sample covariance matrices.

Particularly, these correction terms are estimators of the asymptotic equivalents, under $\mathrm{H}_0(i,j)$, of the different metrics *plug in* distances. The primary advantage of these estimators, as opposed to those presented in the previous chapters, is that these estimators can be directly obtained from the sample covariance matrices. Through numerical simulations, we have confirmed the correctness of the presented results and how one can rely on the estimated asymptotic equivalent to improve clustering. Our results have shown the better performance of the proposed correction term when compared against their original (non-corrected) statistics and three other commonly used metrics.

Chapter 6

Clustering for Rate Splitting and MIMO

In the previous chapter, we described how to take advantage of the different *plug in* metrics to correctly cluster wireless equipments. Particularly, we have introduced correction terms, built directly from the channel matrices, that assist in the clustering of these devices. One of the downsides of this approach is that it assumes that the base station has access to the true channel of each user equipment. However, in practice, it is often the case that the base station only has access to estimators of these channel matrices. In this chapter, we will build upon the results presented in the previous chapter, but this time we propose a shallow neural network based clustering technique to learn and group different UEs according to their instantaneous noisy channels. Moreover, we also take a more wireless focused approach and study the direct impact of clustering in the spectral efficiency of communication systems. Furthermore, once again, we focus on the undersampled regime for which, in the downlink, the total number of UEs exceeds the number of antennas at the base station, making it difficult to effectively mitigate the interference from multiple transmitters.

Rate splitting is a flexible and robust scheme which effectively manages interference among multiple transmitters and receivers. Its versatility and efficacy make rate splitting a promising solution for next-generation wireless networks. Different from other traditional non-orthogonal techniques which try to fully mitigate interference from multiple sources (e.g., NOMA), rate splitting takes advantage of the possible interference by partially decoding it and partially treating it as noise. This has shown to further improve multiplexing gains [70]. In its most simple design, 1-layer rate splitting divides the total data rate into two components: a common message ($s^{(c)}$) shared among all users and K private messages

($s_k^{(p)}, k < K$) intended for the individual users. At the receiver side, one often first decodes the common message while treating the remaining signals as interference, and then apply SIC to retrieve the private message. Notice that the common message is treated as a shared resource that needs to be decoded by all users in the system. This is often tackled by allocating a larger fraction of the total power to the common message. In the presence of a large number of receivers, this condition limits the total rate by the minimal common rate achieved in the whole system. As a consequence, the power assigned to each $s^{(p)}$ is reduced, leading to a degradation in communication rate. This limitation is irrespective of the number of antennas at the transmitter and arises from power allocation strategies aimed at minimizing interference among different users.

A possible solution is to adopt multiple common streams (generalized rate splitting) which leads to higher multiplexing gains. However, this approach comes with the drawback of increased complexity at the decoder due to the presence of several layers of SIC [70]. To address the escalating complexity while minimizing the loss in multiplexing gains, [23] proposes a 2-layer hierarchical rate splitting (HRS) transmission mechanism. In this scenario, users are now considered to be divided into G groups. The kth user of the gth group is required to decode three messages: a common ($s^{(c)}$) one, the semi-private message associated to its group ($s_g^{(sp)}$) and its private message $s_{gk}^{(p)}$. The idea is similar to the one deployed in the 1-layer RS, with the difference that now the system contains a semi-private message which is encoded by a codebook shared only among users in a specific group. Moreover, at the receiver side, each user is required to perform a hierarchical SIC. This involves decoding the common stream ($s^{(c)}$) first and treating the all semi-private streams and the private streams as interference. Next, the user proceeds to decode its associated semi-private common stream $s_g^{(sp)}$, considering the private streams as interference. Finally, the user retrieves its private message from the private stream ($s_{gk}^{(p)}$). This sequential decoding process allows the user to separate and extract the desired information from the different encoded streams. One of the major challenges for such HRS schemes is the necessity to know the optimal clustering of the users based only on their channel state information (CSI). As described in throughout this thesis, this clustering problem is known to be NP hard. Hence, in this chapter, we use the finding from [7] to train a neural network capable of directly learning (or approximating) the optimal clustering option from the imperfect CSI.

6.1 System and Transmission Model

Consider the scenario where a base station (BS) equipped with M antennas transmits messages to K single-antenna user equipments (UEs) over a downlink channel. Additionally, let us assume these UEs to be divided into G disjoint groups, with each group consisting of K_g UEs, $g = 1, \ldots, G$. The signal $\mathbf{y} \in \mathbb{C}^K$ received by all the users can be described by

$$\mathbf{y} = \mathbf{H}^{\mathrm{H}}\mathbf{x} + \mathbf{n} \tag{6.1}$$

where, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]^{\mathrm{T}} \in \mathbb{C}^{M \times K}$ contains the stacked channels of all the $k \in \{1, \ldots, K\}$ UEs, $\mathbf{n} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{I}_K)$ is an additive white Gaussian noise vector and $\mathbf{x} \in \mathbb{C}^M$ is the combined signal

$$\mathbf{x} = \sqrt{p^{(c)}}\mathbf{w}^{(c)}s^{(c)} + \sum_{g=1}^{G} \mathbf{B}_g \left(\sqrt{p_g^{(sp)}}\mathbf{w}_g^{(sp)}s_g^{(sp)} + \sqrt{p_{gk}}\mathbf{W}_g\mathbf{s}_g \right) \tag{6.2}$$

where $p^{(c)}$, $p_g^{(sp)}$ and p_{gk} are the power allocated to the common ($s^{(c)} \in \mathbb{C}$), semi-private ($\mathbf{s}_g^{(sp)} \in \mathbb{C}$, for all g) and private ($\mathbf{s}_g \in \mathbb{C}^{K_g}$) messages, respectively. $\mathbf{B}_g \in \mathbb{C}^{M \times b_g}$ is the group precoder designed from the gth group's long term channel's second order statistics. The number of columns b_g is a design parameter and is related to the rank of the channel's covariance matrix (see [22, 23] for a detailed explanation). Finally, $\mathbf{w}^{(c)}$, $\mathbf{w}_g^{(sp)}$ and $\mathbf{w}_{gk} = [\mathbf{W}_g]_k$ are the unit-norm precoders associated to the instantaneous common, semi-private and private messages, respectively. These terms, together with their impact in the model, are further detailed in the following section.

Remark 6.1. *In this chapter we adopt a slightly different notation from the one used in other chapters of this dissertation. Specifically, we use the notation* $\mathbf{H}_g = [\mathbf{h}_{g,1}, \ldots, \mathbf{h}_{g,K_g}]^{\mathrm{T}} \in \mathbb{C}^{M \times K_g}$ *to denote the matrix which contains the stacked channels of all the* K_g *UEs that are associated to the* g*th cluster, and* $\mathbf{y} \in \mathbb{C}^K$ *to denote the signal received by all the users as in (6.1). Additionally, in this chapter, we also consider the case where, due to limited feedback, the BS only observes an imperfect estimation of the channel. We follow [22] and model the imperfect channel for each UE as the sum of the (perfect) channel and an additional noise component generated from the same subspace which is given by*

$$\mathbf{h}_{g,k} = \mathbf{R}_g\hat{\mathbf{g}}_k = \mathbf{U}_g\boldsymbol{\Lambda}_g^{\frac{1}{2}}\hat{\mathbf{g}}_k = \mathbf{U}_g\boldsymbol{\Lambda}_g^{\frac{1}{2}} \left(\sqrt{1-\tau^2}\mathbf{g}_k + \tau\mathbf{z}_k \right) \tag{6.3}$$

where $\mathbf{U}_g \in \mathbb{C}^{M\times M}$ *a is unitary matrix containing the eigenvectors of the covariance matrix* $\mathbf{R}_g$ *associated to the* g*th group,* $\mathbf{\Lambda}_g \in \mathbb{C}^{M\times M}$ *is a diagonal matrix with its associated eigenvalues, and* $\mathbf{g}_k, \mathbf{z}_k \in \mathbb{C}^M$ *are unrelated random variables that contain i.i.d. entries with zero mean and unit variance. Specifically,* $\mathbf{g}_k$ *describes the complex path gains present in the environment while* $\mathbf{z}_k$ *is associated to the channel's imperfect estimation. Finally,* $\tau \in [0, 1]$ *controls the trade-off between these quantities, i.e., the quality of the instantaneous channel. For instance,* $\tau = 0$ *leads to a perfect channel estimation, i.e.,* $\hat{\mathbf{h}}_{g,k} = \mathbf{R}_g^{\frac{1}{2}}\mathbf{g}_k$ *while* $\tau = 1$ *leads to an uncorrelated channel in the subspace spanned by* $\mathbf{U}_g$*, i.e.,* $\hat{\mathbf{h}}_{g,k} = \mathbf{R}_g^{\frac{1}{2}}\mathbf{z}_k$ *for uncorrelated* $\mathbf{g}_k$ *and* $\mathbf{z}_k$*.*

6.1.1 Hierarchical Rate Splitting Transmission Model

The hierarchical rate splitting transmission design is defined based on the combined transmission signal $\mathbf{x}$ from (6.2). Specifically, to achieve the maximum sum-rate it is required that there exists zero interference among clusters. Particularly, the group precoder $\mathbf{B}_g$ aims to reduce this interference by minimizing the inter-group interference (i.e., $\mathbf{B}_g^{\mathrm{H}}\mathbf{H}_l \approx \mathbf{0}, \forall l \neq g$) while enhancing the signal intended to the gth group [71]. This interference can be understood as the leakage of power from each of the lth interference groups into the gth intended group, for all $l \neq g$. We follow the approaches in [22, 23] and consider $r_g^* = \sum_{l\neq g}^{G} r_l$ singular vectors associated with each one of the r_l largest eigenvalues of these interference groups ($l \neq g$) to build the precoder $\mathbf{B}_g$ orthogonal to them[1]. As a result, we end up with the gth group effective channel $\tilde{\mathbf{H}}_g^{\mathrm{H}} = \mathbf{H}_g^{\mathrm{H}}\mathbf{B}_g$ of dimensionality $b_g \times K_g$. This effective channel represents the projection of $\mathbf{H}_g$ onto the b_g–dimensional subspace orthogonal to the r_g^* strongest components of the interference groups. It is important to note that, in order to properly distinguish the K_g users within the gth group, we must have $b_g > K_g$. Moreover, it is also not possible to set both b_g and r_g indiscriminately large, as they impose constraints on each other. Specifically, since there are a maximum of M singular vectors available in each group, this means that the number of users in a group should be less than (or equal to) the dimension of the subspace of $\tilde{\mathbf{H}}_g$, so that we can ensure $K_g \leq b_g \leq M - r_g^*$. Consequently, a large number of groups leads to less freedom on the choice of both b_g and r_g which can harm the overall communication rate.

Moreover, $\mathbf{w}^{(c)}$, $\mathbf{w}_g^{(sp)}$ and $\mathbf{w}_{gk} = [\mathbf{W}_g]_k$ are the unit-norm precoders associated to the common, semi-private and private messages, respectively. To address the

[1] Generally, r_l is a design parameter.

interference among the private messages within the gth group, we adopt a Regularized Zero Forcing (RZF) precoder. Following the approach described in [22], for a given total transmission power P, we can define

$$\mathbf{W}_g = \xi_g \left(\tilde{\mathbf{H}}_g \tilde{\mathbf{H}}_g^{\mathrm{H}} + \varepsilon \mathbf{I}_{b_g} \right)^{-1} \tilde{\mathbf{H}}_g,$$

where ξ_g is used to ensure that the precoder operates within a desired power level, i.e., to fix $||\mathbf{W}_g||_F = 1$. The parameter ε is also a normalization factor (see [23] for a detailed analysis on the choice of normalization). Furthermore, we build $\mathbf{w}_g^{(sp)}$ as the equally weighted Matched Beamforming (MBF) vector given by

$$\mathbf{w}_g^{(sp)} = \xi_{ic,g} \sum_{k=1}^{K_g} \mathbf{w}_{gk}$$

where $\xi_{ic,g}$ is a normalization parameter. The purpose of this MBF vector is to provide a beamforming scheme that evenly distributes the weight among the private precoders within the group, facilitating a balanced contribution from each user. Finally, the common precoder

$$\mathbf{w}^{(c)} = \xi_{oc} \sum_{g=1}^{G} \sum_{k=1}^{K_g} \mathbf{B}_g \tilde{\mathbf{h}}_{gk}$$

is a weighted MBF, but it is designed to handle inter-group power leakage where ξ_{oc} is another normalization parameter.

To distribute power among the different messages, we introduce two parameters, α and β, both in the interval $(0, 1]$. The first parameter α represents the fraction of the total power P that is allocated to the common message. The parameter β represents the fraction of the remaining power (after allocating power to the common message) that is allocated to all the semi-private messages. By combining these parameters, we can determine the power allocation for each message as follows: $p^{(c)} = \alpha P$ for the common message, $p_g^{(sp)} = \frac{(1-\alpha)\beta P}{G}$ for the semi-private message in the gth group, and $p_{gk} = \frac{(1-\alpha)(1-\beta)P}{K_g}$ for the private message of user k in the gth group. As we are primarily concerned with the clustering aspect of HRS solutions, in this work we adopt a brute force search to find the optimal values of α and β. This is done for each channel realization.

As previously mentioned, at the receiver side, the kth user belonging to the gth group employs a 2-step successive interference cancellation (SIC) technique to decode its message. In the first step, the user decodes the common message ($s^{(c)}$) and removes it from the received signal, thereby eliminating its interference to the

other messages. Subsequently, in the second step, the user proceeds to decode the semi-private common codeword of the group after applying SIC to further mitigate interference. Once both common messages have been successfully decoded, the user extracts its own private message by treating the remaining private messages as interference. The Signal-to-Interference Plus-Noise Ratio (SINR) for each of these messages can be expressed as follows:

$$\gamma_{gk}^{(c)} = \frac{p^{(c)}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}^{(c)}|^2}{1+I_{gk}} \tag{6.4}$$

$$\gamma_{gk}^{(sp)} = \frac{p_g^{(sp)}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_g^{(sp)}|^2}{1+I_{gk}-p_g^{(sp)}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_g^{(sp)}|^2} \tag{6.5}$$

$$\gamma_{gk}^{(p)} = \frac{p_{gk}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{gk}|^2}{1+I_{gk}-\left(p_g^{(sp)}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_g^{(sp)}|^2+p_{gk}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{w}_{gk}|^2\right)} \tag{6.6}$$

where

$$I_{gk} = \sum_{l=1}^{G} p_l^{(sp)}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{B}_l\mathbf{w}_l^{(sp)}|^2 + \sum_{l=1}^{G}\sum_{k=1}^{K_g} p_{lk}|\mathbf{h}_{gk}^{\mathrm{H}}\mathbf{B}_l\mathbf{w}_{lk}|^2$$

is the combination of all interference leaked from other users and groups.

Finally, we can describe the achievable rate as the combination of the smallest achievable common rate among all users $R^{(c)} = \min_{gk} \log_2(1+\gamma_{gk}^{(c)})$, the minimal semi-private common rate per group $R^{(sp)} = \sum_{g=1}^{G} \min_{k}(\log_2(1+\gamma_{gk}^{(sp)}))$ and the sum of the rate achievable at all private messages $R^{(p)} = \sum_{g=1}^{G}\sum_{k=1}^{K_g} \log_2(1+\gamma_{gk}^{(p)})$. Then, the total achievable rate is obtained by summing up these components: $R = R^{(c)} + R^{(sp)} + R^{(p)}$.

6.2 User Clustering for HRS

The selection of an appropriate grouping mechanism is of utmost importance to fully leverage the benefits of two-tier precoding techniques like Hierarchical Rate Splitting (HRS) [22, 23]. However, as extensively explained throughout this thesis, finding the optimal clustering is a challenging task as it involves searching through a large number of potential partitions. The number of ways to partition a set into non-empty subsets is given by the Bell number which grows exponentially with the size of the set, K in our case. Moreover, many of these partitions may result in poor communication rates due to high interference. To address this challenge, we leverage the hierarchical clustering solution based on the subspace similarity (see

Chapter 5) to narrow down the number of available options. This time around, we also re-calculate the centroid of the new cluster after every merge. By using an agglomerative approach, we can systematically merge UEs and generate a hierarchy of possible clustering configurations which naturally consider their similarity. The advantage of this approach is that it reduces the search space from an exponential number of possibilities to a more manageable and feasible number of choices. Instead of considering all possible combinations, we can focus on evaluating the $K+1$ clustering options obtained from the clustering hierarchy (each of the $K+1$ levels of the hierarchy produces a new clustering solution). By analyzing these distinct clustering options, we can assess the performance of different clustering configurations and select the ones that yield more promising results. While this approach may not guarantee the optimal clustering solution, it provides a practical and effective way to explore a subset of clustering options and make informed decisions based on their achieved communication rates.

We dedicate the remainder of this chapter to illustrate this idea. Specifically, we will begin by introducing a general synthetic dataset definition, which we will use in our simulations to showcase the advantages of clustering in HRS communication systems. This dataset will include downlink communication channels, and their (best) clustering solution obtained from performing a hierarchical clustering of these channels. Here, the best solution refers to the grouping that yields the highest communication rate when using the hierarchical rate splitting scheme described above. Figure 6.1 illustrates the generation of one sample of this dataset. The idea is that after every merge we compute the communication rate obtained using the HRS scheme and in the end select the class that achieves the maximum rate (this is represented by a dashed line in the figure). Moreover, we also use this best clustering solution to train a shallow neural network capable of predicting the grouping of different UEs based solely on their estimated channel matrices. By doing so, we assume a more realistic scenario where the BS only has access to estimates of these channel matrices. Moreover, this can become particularly useful in scenarios where there exist a large number of UEs and the hierarchical clustering procedure might be expensive depending on the linkage method used.

6.2.1 Simulated Scenario & Dataset Definition

To illustrate our method, we start by generating a dataset of channel matrices according to (6.3) and clustering them according to the hierarchical clustering scheme described above. Specifically, we run the algorithm using the subspace similarity

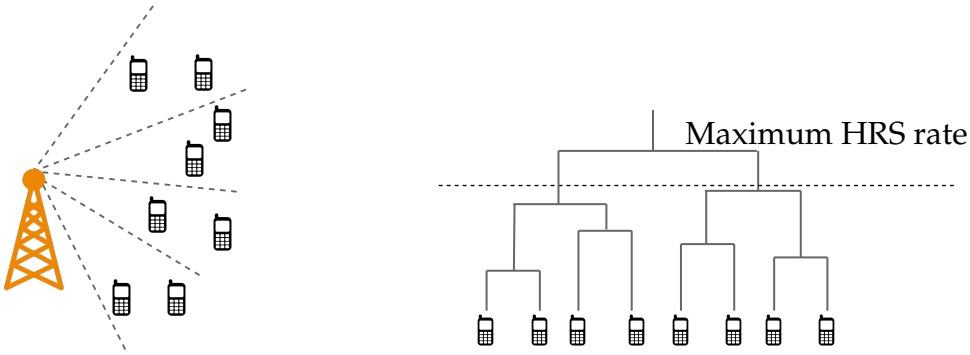


Figure 6.1: Scheme of generation of one sample of the dataset. (a) A illustrative downlink communication scenario. (b) Hierarchical clustering solution.

metric $\tilde{s}_{kl}$ defined in (5.12) and concatenate the clustered channel matrices side by side to represent new formed clusters, similar to the approach described in Section 5.3. To build our dataset, we consider four different covariance matrices and randomly generate MIMO channels using each of them. We will denote this processes of generating and clustering the MIMO channels as the generation (or realization) of one sample. We assume that the number of users associated to a particular covariance matrix may vary across different samples. Additionally, this association does not necessarily represent cluster assignments, instead, it is solely a method for generating random channels. The cluster assignment is done based on the clustering solution (different levels of the hierarchy) that achieves maximum communication rate using the HRS communication scheme. In other words, we estimate the HRS communication rate for each level of the hierarchy and choose the clustering solution that yields the highest rate. Moreover, these covariance matrices are derived by considering azimuth angles, denoted as $\theta_g = -\frac{\pi}{2} + \frac{\pi}{3}(g-1)$, and a constant angular spread of $\Delta_g = \frac{\pi}{6}$, $g \in \{1, 2, 3, 4\}$. Finally, we assume that the base station (BS) is equipped with a Uniform Circular Array (UCA) antenna. By considering these parameters and configurations, we can create a representative scenario to evaluate the performance of our method in handling multi-group communications with different channel characteristics.

Particularly, we design three different configurations based on the relationship between the number of UEs K and the number of antennas at the BS M, namely, the overloaded ($K > M$), balanced ($K = M$) and underloaded ($K < M$) scenarios. We evaluate these scenarios for different values of K and M. Specifically, we consider $M \in \{4, 8, 12\}$ for $K = 8$ and $M \in \{6, 12, 16\}$ for $K = 12$ so that we have

six different scenarios. We generate a total of 10.000 random samples for each scenario, which includes both imperfect (for fixed $\tau^2 = 0.4$) and perfect channel state information. Additionally, for each sample in each scenario, we determine the clustering scheme that maximizes the transmission rate based on the hierarchical clustering mechanism and the true CSI. Apart from this scenario, the perfect channel information is used solely to evaluate the performance of the methods being compared.

Due to the random nature of the samples, for each of the six scenarios, we obtain more than $G^* = 200$ possible clustering options (accumulated over different sample realizations) that lead to optimal results. This results in highly imbalanced datasets[2] which becomes particularly problematic when training a NN. To address this, we perform sub-sampling by discarding classes that achieve less than 25% of the average rate of the scenario and have fewer than 50 samples associated to them. Moreover, to further balance the data, we limit the maximum number of samples in each class to 200. As a result, for each scenario, we still obtain an imbalanced dataset with approximately $G^* = 50$ classes, each containing at least 50 samples and at most 200 samples. To address the decrease in the number of samples after sub-sampling, we perform additional data augmentation by randomly shuffling the users within each cluster. This augmentation technique leads to a natural extension of this dataset as clustering is unaffected by the order in which users are arranged. Notice that this procedure is performed after we have already generated the channel matrices and clustered them using this hierarchical clustering method and the projection-Frobenius based similarity. In other words, by shuffling the users, we can generate additional variations of the same clustering configuration, thereby enriching the dataset and (possibly) improving the robustness of models trained in with this dataset.

6.2.2 Performance Analysis

We solve the classification problem presented in the previous section by designing a shallow neural network with the same structure and parameters as described in [28]. The output layer consists of G^* neurons with a *softmax* activation that correspond to each cluster where G^* is the total number of classes in the scenario. For the training procedure, we use the Adam optimizer with a learning rate of 10^{-3}, we train for 50 epochs and use a batch size of 128 samples. For our multi-class

[2]Here one dataset is associated with one of the different configurations considered in this chapter, namely overloaded ($K > M$), balanced ($K = M$), and underloaded ($K < M$) scenarios.

Table 6.1: Parameters of the Simulations

Simulation Parameter	**Simulation Value**
Antenna Configuration	Uniform Circular Array
Angular Spread (Δ_g)	$\pi/6$
Number of Unique Distributions	4
Channel Quality (τ^2)	0.4
Dominant Eigenvectors ($b_g = r_g$)	$\lfloor M/G \rfloor$
Channel Quality (r_g)	0.4
Number Shuffling	10
Number of Neurons in NN	$\{256, 128\}$
NN Learning Rate	10^{-3}
NN Training Epochs	50
NN Training Batch Size	128
Total Number of Classes (G^*)	50
NN Loss Function	Categorical Cross-entropy Loss

classification task, we aim to minimize the categorical cross-entropy loss. Each class representing a possible clustering solution. For each scenario, we divide our dataset into training, validation and test sets in a proportion of $80/10/10$. During the training procedure, we use the validation set to tune the corresponding hyper-parameters. Our model is defined as a shallow neural network following the parameters from Table 6.1.

Additionally, when evaluating the accuracy of the NN model, the conventional approach is to consider the class which is predicted with the highest probability (*top-1* accuracy). However, in our case, there can be multiple clustering options that achieve sufficiently high transmission rates. Hence, it also becomes valuable to analyze the *top-k* accuracy, which measures if the desired clustering option is among the k most probable outputs. This provides a more comprehensive assessment of the model's performance in capturing the potential clustering solutions.

In order to validate the learning of the NN, we compare the achieved rate using the NN predicted classes and different RS clustering options. To perform a complete evaluation, we determine the rate achieved by the following solutions,

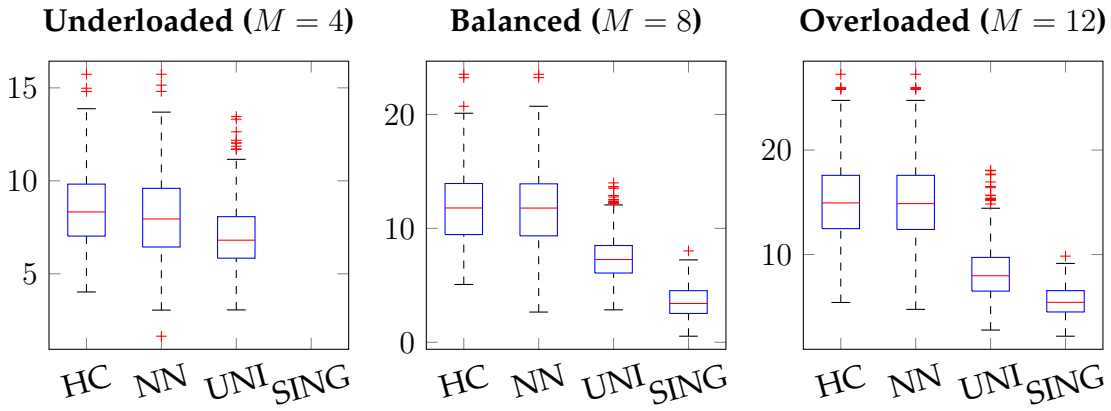
- Hierarchical Clustering - Hierarchical Rate Splitting (HC): The users are clus-

tered according to a hierarchical clustering solution, the group with higher communication performance is selected;

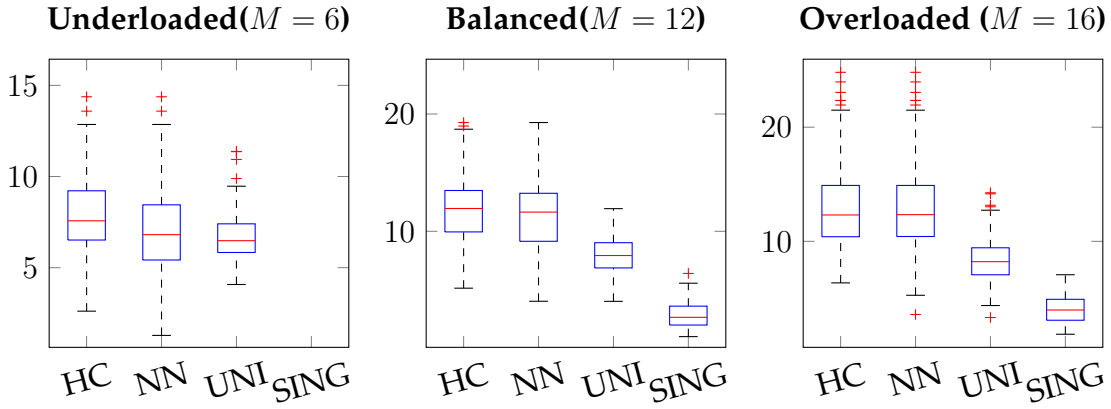
- Neural Network - Hierarchical Rate Splitting (NN): Proposed NN based clustering;
- Universal Cluster (UNI): All users are clustered into one single cluster;
- Singleton Cluster (SING): Each cluster contains only single user.

As mentioned above, we consider three scenarios to evaluate the clustering solutions 1) $M < K$, 2) $M = K$ and 3) $M > K$. Hence, for $K = 8$, we determine the rate achieved for $M \in \{4, 8, 12\}$ and for $K = 12$, we determine the rate achieved for $M \in \{6, 12, 16\}$. Then, we compare the different clustering techniques mentioned before based on the rate achieved. Figure 6.2 shows the rate achieved (in the test set) for all four clustering techniques for the different values of M and K. Each box plot shows the rate obtained for different realizations of the channel in the test dataset. The median rate is presented by a horizontal line through the box and the top and bottom of the box are the 75th and 25th percentile rate (i.e. rate achieved by 75% and 25% of the scenarios). Lastly, the extremities of the boxplot refer to the 1% and 99% and the red plus indicators in the boxplot denote the outlier rate values. Notice that the rate achieved by HC-HRS and NN-HRS is approximately similar while both clustering techniques outperform UNI and SING. This is due to the fact that with a noisy channel, it is really difficult to generate accurate precoders that can maximize the rate and minimize the inter-group and intra-group interferences. Additionally, the NN-HRS only receives the instantaneous noisy channel as an input and determines its clustering solution while HC-HRS needs to iteratively determine the similarity between different channels making it considerably slower when compared to the NN solution. Moreover, for SING, the choice of parameters b_g and r_g seems to harm the performance. We recall that both parameters are integers thus are susceptible to the trade-off between M and G. For instance, for $G = K = 8$ and $M = 12$, there exist only one viable option of r_g, i.e., $r_g = \lfloor M/G \rfloor = 1$. Alternatively, we could select four ($\mathrm{mod}(M, G)$) groups to have $r_g = 2$, but this requires further processing on the choice of these groups. As a consequence, we obtain similar rates for $K = 8$ users served with $M = 8$ or $M = 12$. Similar consequences are obtained for $K = 12$. Moreover, for $G = K > M$, we have $r_g = \mathrm{mod}(M/K) = 0$ what makes impossible to derive meaningful precoders

Fig 6.2(a)-(b). In contrast to that, the other three techniques, which consider clustering, do not suffer from this trade-off between G, r_g and M. Instead, even for $K > M$ we still achieve reasonable spectral efficiency.



(a) $K = 8$



(b) $K = 12$

Figure 6.2: Spectral efficiency (bps/Hz) achieved for clustering mechanisms using HRS.

Finally, we analyze the capability of the shallow NN to learn the grouping classification task as described above. To do so, we first analyze the accuracy of the network for class prediction. Recall that, here, a class represents a different clustering option. Table 6.2 presents, in percentage, the results obtained by training different NN according to the configuration parameters in Table 6.1 for different number of users (K) and antennas in the BS (M). The validation column contains the final classification accuracy in the validation dataset and indicates some learning capability in untrained data. During our experiments, we noticed that different points of the same dendrogram might result in similar communication rates, i.e.,

Table 6.2: *Top-k* Accuracy of validation and test sets

K/M	**Validation** ***(top-1)***	**Test** ***(top-1)***	**Test** ***(top-3)***	**Test** ***(top-5)***	**Test** **Relative Rate**
8 / 4	65.38%	65.37%	85.22%	90.48%	94.12%
8 / 8	98.3%	92.0%	96.3%	97.7%	99.0%
8 / 12	96.9%	92.2%	97.0%	98.2%	99.5%
12 / 6	71.45%	35.6%	65.62%	77.75%	89.99%
12 / 12	98.7%	86.2%	96.8%	98.9%	93.5%
12 / 16	99.18%	95.62%	98.32%	93.32%	99.77%

there might exist different clustering options which achieve the same rate. Therefore, for the test dataset, we show the *top-1*, *top-3* and *top-5* classification accuracy. Despite the fact that performance in *top-1* accuracy might be considered poor, the *top-5* results are, often, above 90%. Finally, the last column compares (in %) the relative communication rate if using the *top-1* option from the NN. Results show that, except in the cases where $K > M$, on average, the rate drops 2.5% which is an acceptable loss when compared to the complexity of the original problem. Moreover, we can infer from these results that the NN is capable of learning the maximum clustering option or clusters that approximate this option. In other words, it is capable of learning the relationship between different users directly from their channel matrices and cluster the users with a high degree of accuracy for most scenarios and finally achieve a rate comparable to more complicated similarity-based HC-HRS.

6.3 Conclusions

In this chapter, we have addressed the challenge of estimating the optimal grouping of users in a two-layer hierarchical rate splitting scenario. Due to the exponential growth of possible clustering solutions, finding the optimal grouping is an NP-hard problem. Moreover, most of these groupings result in high interference between users and should be disregarded. To overcome this, we have proposed the use of clustering techniques to identify a subset of relevant clustering solutions that achieve high communication rates by mitigating interference between

users. By leveraging these feasible clustering solutions, we have trained an NN architecture to learn the relationship between the channel matrices and the optimal clustering configurations. Numerical results have shown that the trained neural network achieves similar rates to those obtained using the hierarchical clustering solution. This demonstrates the effectiveness of machine learning techniques in learning the grouping of users in complex communication scenarios. Overall, the approach offers a practical and efficient solution for determining the clustering of users, enabling the design of efficient communication strategies that maximize the achievable rates while mitigating interference.

Chapter 7

Conclusions and Future Directions

This thesis has addressed a diverse range of challenges that arise from applying machine learning to large-dimensional observations, specifically by exploring statistical properties of the distances between positive definite matrices. Our contributions encompass deep theoretical insights into unsupervised learning with high-dimensional data, along with practical solutions to improve common unsupervised learning approaches. In particular, we have shown how to estimate the distance between covariances of high-dimensional observations and how to exploit these results in learning tasks. We have also exemplified the applicability of our findings in several applications in the field of wireless communications. Below, we provide a summary of contributions discussed throughout this thesis together with insights on possible future works.

In Chapter 2, we studied the asymptotic characterization of a general class of distances between sample covariance matrices, which consider the Riemmanian geometry of the set of positive definite matrices. These distances can be expressed as the sum of traces of analytic functions applied to each matrix separately. The results are established in the asymptotic regime where both the sample size and the observation dimension tend to infinity at the same rate. These generic results generally hold for both the undersampled and oversampled regimes, as well as for complex- and real-valued observations. Furthermore, we have specialized these results to three commonly used distances between covariance matrices: the Euclidean distance, a symmetrized version of the Kullback–Leibler divergence, and Subspace Similarity based on the principal angles of the compared subspaces. Always in Chapter 2, we considered the traditional *plug-in* estimators, where the sample covariance matrices are directly plugged into the definition of the distance between covariance matrices. Numerical simulations validated the presented re-

sults, illustrating the accuracy of the asymptotic behavior of these metrics when compared to their empirical distributions.

Moving on to Chapter 3, we further improved these *plug-in* estimators to consistently approximate the distance between true covariance matrices. We proposed a general form for the consistent estimator of this particular family of distances, along with a central limit theorem that describes its asymptotic behavior. We also provided the closed-form solution for the consistent estimator of three distances between covariance matrices: the symmetrized version of the Kullback–Leibler divergence, the Euclidean distance, and the Log-Euclidean distance. Additionally, we also presented closed-form solutions for the mean and variance of these metrics, except for the Log-Euclidean distance, whose fluctuations are expressed in terms of an integral.

A general numerical evaluation of both the traditional *plug-in* and the consistent estimators of distances between covariance matrices was provided in Chapter 4. Particularly, we demonstrated the utility of the CLTs formulated in this thesis to further account for the impact of measurement (e.g., distance/similarities between samples) uncertainty in clustering solutions. Furthermore, we also illustrated how these results can assist in properly designing clustering solutions. Specifically, we showed that the CLTs derived in this work become useful when assessing the quality of clustering solutions based on different distance metrics. It is important to emphasize that the outcomes outlined in this study are not limited to the specific clustering methods examined in this thesis. Instead, they provide a broader analysis and are also applicable to other clustering approaches that rely on comparing pairwise similarities and for distances that fall into the family of distances investigated in this thesis.

In the remainder of the thesis, specifically in Chapters 5 and 6, we provided practical examples demonstrating the application of these analytical results to the field of wireless communications. Particularly, in Chapter 5, we analyzed an agglomerative hierarchical clustering of user equipments based on the alignment of the subspaces spanned by their channel matrices. We further explored the applications of the *plug in* distances in the undersampled regime, for which consistent distance estimators are not generally available. We showed that, properly correcting the *plug-in* distances using their respective asymptotics, we can better compare subspaces of different dimensions (in our scenario, groups of UEs of different sizes). The correction terms were particularly tailored to converge to the

corresponding deterministic equivalent when we have identical covariance matrices, which facilitates the clustering of UEs. Finally, in Chapter 6, we built upon these results to train a neural network to select the (almost) optimal UEs clustering scheme that enables the application of a (hierarchical) rate splitting coding strategy. These practical examples illustrate the relevance of our analytical findings in real-world wireless communications scenarios. By applying our research insights, one can enhance clustering techniques and improve the overall efficiency and performance of wireless communications systems.

7.1 Future Directions

Finally, there are several promising directions for future research that can build upon the foundations laid out in this thesis. Firstly, throughout this thesis, we have only provided hints on how to devise, directly from the data, general asymptotic estimators for the first and second order moments of the *plug-in* and consistent distance estimators, leaving the task of finding their closed-form solutions for further investigation. These solutions can offer a deeper understanding of their statistical properties and potentially increase the applicability of our results to scenarios where limited data samples are available. Particularly, areas such as manifold learning, meta learning, and explainable AI can greatly benefit from the insights and methodologies developed in this thesis. Specially, these fields could leverage the statistical tools we have introduced to gain a better understanding of the inherent data structures, enhance transfer learning processes, and provide more interpretable explanations for machine learning solutions.

Additionally, considering the evolving needs of wireless communications, further research can be conducted to adapt and enhance our techniques to address the unique challenges of 5G and beyond communications systems. This could involve incorporating the insights from this thesis into new large dimensional antennas configurations (e.g., Large and Reconfigurable Intelligent Surfaces); advanced communication protocols, including UAV routing protocols based on position uncertainty and clustering-aware multiple access schemes; and network optimization strategies, by exploiting manifold structures for improved efficiency. By doing so, one can potentially take advantage of the findings from this thesis and improve the overall efficiency and performance of future wireless networks.

Lastly, extending the application of our analytical results beyond wireless communications holds considerable potential. For instance, in areas like natural lan-

guage processing, where analyzing large volumes of text data can be challenging due to the high-dimensional nature of text features, researchers could adapt the statistical tools developed in this thesis to derive efficient distance metrics between text representations (e.g., word embeddings) and apply them to measure the similarity between documents or parts of text. This could lead to improved accuracy in sentiment analysis, text translation and classification tasks. Similarly, applications in the medical domain, where statistical analysis is crucial for disease prevention, early detection, and treatment, can also benefit from these methodologies.

In summary, while this thesis has significantly contributed to the theoretical aspects of machine learning with practical applications in wireless communication, it has also laid the foundation for future research to delve deeper into the discussed theoretical aspects and extend applications to diverse domains with specific requirements. Together, these can further contribute to advancing the field of machine learning algorithms in large dimensional settings.

Appendix A

Appendix for Chapter 2

A.1 Proof of Theorem 2.2

We recall here the definitions introduced in Theorem 2.1 and (2.4), namely

$$\hat{\mathbf{Q}}_j(z_j) = (\hat{\mathbf{R}}_j - z_j \mathbf{I}_M)^{-1} \qquad \text{and} \qquad \bar{\mathbf{Q}}_j(z_j) = \frac{\omega_j(z_j)}{z_j} \mathbf{Q}_j(\omega_j(z_j)).$$

Moreover, to simplify the notation in this appendix, we will drop the dependence on z_j whenever it is obvious from the context. A similar notation will be used when the argument of these functions is an alternative variable z_j'. More specifically, we will write (for $j \in \{1,2\}$), $\hat{\mathbf{Q}}_j \equiv \hat{\mathbf{Q}}_j(z_j)$, $\bar{\mathbf{Q}}_j \equiv \bar{\mathbf{Q}}_j(z_j)$, $\omega_j \equiv \omega_j(z_j)$ and $\mathbf{Q}_j \equiv \mathbf{Q}_j(\omega_j(z_j))$. We will use a similar short-hand notation for the "prime" quantities, namely $\hat{\mathbf{Q}}_{j'} \equiv \hat{\mathbf{Q}}_j(z_j')$, $\bar{\mathbf{Q}}_j \equiv \bar{\mathbf{Q}}_{j'}(z_j')$, $\omega_{j'} \equiv \omega_j(z_j')$ and $\mathbf{Q}_{j'} \equiv \mathbf{Q}_j(\omega_j(z_j'))$. Let us start by considering the random variable

$$\hat{\zeta}_M = \frac{1}{(2\pi \mathrm{j})^2} \oint_{C_1^-} \oint_{C_2^-} g(z_1, z_2) \mathrm{tr}\left[\hat{\mathbf{Q}}_1 \hat{\mathbf{Q}}_2 - \bar{\mathbf{Q}}_1 \bar{\mathbf{Q}}_2\right] dz_1 dz_2$$

In order to derive the CLT, we will consider the function

$$\Psi_M(u) = \mathrm{e}^{\mathrm{j}u\hat{\zeta}_M}.$$

The objective is to show that, in the limit when $M, N_j \to \infty$, we have

$$\mathbb{E}\left[\Psi_M(v)\right] - \mathrm{e}^{\mathrm{j}u\mathfrak{m}_M - (u\sigma_M)^2/2} \to 0 \tag{A.1}$$

pointwise in u, where $\mathfrak{m}_M$ and σ_M^2 will be as defined in Theorem 2.2. Given the boundedness assumptions in the statement of the theorem, the result will follow from a trivial modification of [72, Proposition 6]. The rest of the section is therefore devoted to showing (A.1).

Unfortunately, the random variable $\hat{\zeta}_M$ above does not need to have a characteristic function for all M. This is because there might exist realizations for which the positive eigenvalues of $\hat{\mathbf{R}}_j$ become dangerously close to the contour C_j. In order to overcome this difficulty, we will follow the approach in [73,74] and consider an equivalent (large-M) representation of $\hat{\zeta}_M$ that is guaranteed to have a characteristic function for all M. Indeed, let us define the interval $\mathcal{S}_j^\epsilon$ as an ϵ-blowup of the set $\mathcal{S}_j \cup \{0\} = [\theta_j^-, \theta_j^+] \cup \{0\}$, that is $\mathcal{S}_j^\epsilon = [\theta_j^- - \epsilon, \theta_j^+ + \epsilon] \cup [-\epsilon, \epsilon]$ for some small $\epsilon > 0$. Assume that ϵ is small enough such that $\mathcal{S}_j^{2\epsilon}$ does not intersect with the contour C_j. Let ϕ denote a smooth function $\phi_j : \mathbb{R} \to [0,1]$ such that $\phi_j(x) = 1$ for $x \in \mathcal{S}_j^\epsilon$ and $\phi_j(x) = 0$ for $x \in \mathbb{R} \backslash \mathcal{S}_j^{2\epsilon}$. We will write $\phi_j = \det \phi_j(\hat{\mathbf{R}}_j)$. By [58], we know that $\phi_j = 1$ with probability one for all M sufficiently large. Therefore, we have $\hat{\zeta}_M = \tilde{\zeta}_M$ almost surely for all M sufficiently large, for

$$\tilde{\zeta}_M = \frac{-1}{4\pi^2} \oint_{C_1^-} \oint_{C_2^-} g(z_1, z_2) \mathrm{tr} \left[\hat{\mathbf{Q}}_1 \hat{\mathbf{Q}}_2 \phi_1 \phi_2 - \bar{\mathbf{Q}}_1 \bar{\mathbf{Q}}_2 \right] dz_1 dz_2. \tag{A.2}$$

The characteristic function of $\tilde{\zeta}_M$ exists for every realization and every possible M. Having introduced this regularization parameter ϕ_j, we are now in the position of introducing the main technical tools that will be used in the proof of this theorem. Following the approach in [72], our derivations will be based on the partial integration formula for Gaussian functionals, together with the Poincaré-Nash inequality. We introduce these tools in the following proposition.

Remark A.1. *In what follows, the symbol $\mathcal{O}(M^{-k})$ will denote a general multivariate complex function that is bounded in magnitude by $\epsilon(z_1, \ldots, z_4) M^{-k}$, where $\epsilon(z_1, \ldots, z_4)$ does not depend on M and is such that*

$$\max_{m,n} \sup_{(z_1,\ldots,z_4) \in C_m \times \ldots \times C_n} \| \epsilon(z_1, \ldots, z_4) \| < +\infty. \tag{A.3}$$

The function itself may be different from one line to another, and it may be matrix valued, in which case (A.3) is understood as the spectral norm. Moreover, $\mathcal{O}(M^{-\mathbb{N}})$ should be understood as a multivariate complex function that can be written as $\mathcal{O}(M^{-\ell})$ for every $\ell \in \mathbb{N}$.

Proposition A.1. *Consider function $\Omega(\mathbf{X}, \mathbf{X}^*, z) : \mathbb{R}^{2MN_1} \to \mathbb{C}$ to be continuously differentiable and such that both itself and its partial derivatives are polynomically bounded. If $\mathbf{X}$ is real valued, simply consider Ω as a function on $\mathbb{R}^{MN_\ell}$, with the same properties. If*

$\mathbf{X}$ is a matrix of i.i.d. standard Gaussian random variables and X_{ij} denotes its i,jth entry, we have

$$\mathbb{E}\left[X_{ij}\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\right]=\mathbb{E}\left[\frac{\partial\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)}{\partial X_{ij}^*}\right] \tag{A.4}$$

where

$$\frac{\partial}{\partial X_{ij}^*}=\frac{1+\varsigma}{2}\frac{\partial}{\partial\mathrm{Re}\left[X_{ij}\right]}+\mathrm{j}\frac{1-\varsigma}{2}\frac{\partial}{\partial\mathrm{Im}\left[X_{ij}\right]}.$$

On the other hand, we can also write

$$\begin{aligned}\mathrm{var}\left[\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\right]\leq\sum_{i=1}^{M}\sum_{j=1}^{N_\ell}&\mathbb{E}\left[\left|\frac{\partial\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)}{\partial X_{ij}}\right|^2\right]\\&+\left(1-\varsigma\right)\mathbb{E}\left[\left|\frac{\partial\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)}{\partial X_{ij}^*}\right|^2\right]\end{aligned} \tag{A.5}$$

where now

$$\frac{\partial}{\partial X_{ij}}=\frac{1+\varsigma}{2}\frac{\partial}{\partial\mathrm{Re}\left[X_{ij}\right]}-\mathrm{j}\frac{1-\varsigma}{2}\frac{\partial}{\partial\mathrm{Im}\left[X_{ij}\right]}.$$

Assume that, for each fixed $z\in\mathbb{C}$ and $\ell\in\{1,2\}$. The function ϕ_ℓ is continuously differentiable (on $\mathbb{R}^{2MN_\ell}$) with polynomically bounded partial derivatives. If, in addition, $\sup_{z\in C_\ell}\mathbb{E}\left(\left|\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\phi_\ell\right|^2\right)<C$ for some positive deterministic C independent of M, then

$$\mathbb{E}\left[\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\phi_\ell^r\right]=\mathbb{E}\left[\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\phi_\ell\right]+\mathcal{O}\left(M^{-\mathbb{N}}\right) \tag{A.6}$$

for any $r\in\mathbb{N}$, and also

$$\mathbb{E}\left[\Omega\left(\mathbf{X},\mathbf{X}^*,z\right)\frac{\partial\phi_\ell}{\partial X_{ij}}\right]=\mathcal{O}\left(M^{-\mathbb{N}}\right) \tag{A.7}$$

where the term $\mathcal{O}\left(M^{-\mathbb{N}}\right)$ should be understood as in Remark A.1 above.

The above results are well known in the random matrix literature and the proof is therefore omitted. One of the conclusions of Proposition A.1 is the fact that we can basically ignore the presence of the regularization term ϕ_ℓ up to an error of order $\mathcal{O}\left(M^{-m}\right)$ for every $m\in\mathbb{N}$, which will be irrelevant for the purposes of our derivations.

From now on we will therefore consider the definition of $\tilde{\zeta}_M$ in (A.2). Furthermore, we will denote $\mathbf{Y}_\ell=\mathbf{R}_\ell^{1/2}\mathbf{X}_\ell$ for $\ell\in\{1,2\}$ and we will use the symbol $\mathbb{E}_\ell\left[\cdot\right]$ to denote the expectation with respect to the entries of $\mathbf{X}_\ell$, which are all i.i.d. standard Gaussian random variables. The jth column vector of $\mathbf{X}_\ell$ will be denoted as $\mathbf{x}_j$ and the (i,j)th entry as X_{ij} (in both cases, the dependence on ℓ will be obvious from the context). We will also drop the dependence on M in $\Psi_M(u)$ to further

simplify the notation. Finally, using the Dominated Convergence Theorem, we can establish that $\mathbb{E}\left[\Psi(u)\right]$ is a differentiable function with derivative

$$\frac{d\mathbb{E}\left[\Psi(u)\right]}{du} = \mathrm{j}\mathbb{E}[\tilde{\zeta}_M \Psi(u)].$$

We start by noting that we can decompose the sample covariance matrix $\hat{\mathbf{R}}_1$ as where $\mathbf{e}_i$ is the ith column of the $M \times M$ identity matrix. By denoting as $\mathbb{E}_\ell$ the expectation with respect to the elements of $\mathbf{X}_\ell$, we have

$$\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\Psi(u)\phi_1\right] = \sum_{i=1}^{M}\sum_{j=1}^{N_1}\mathbb{E}_1\left[X_{ij}\hat{\mathbf{Q}}_1\mathbf{R}_1^{1/2}\frac{\mathbf{e}_i\mathbf{x}_j^H}{N_1}\mathbf{R}_1^{1/2}\Psi(u)\phi_1\right]$$

which can be further developed using the integration by parts [72], together with the identity

$$\frac{\partial}{\partial X_{ij}^*}\hat{\mathbf{Q}}_\ell = -\hat{\mathbf{Q}}_\ell\mathbf{R}_\ell^{1/2}\frac{\left(\mathbf{x}_j\mathbf{e}_i^H + \varsigma\mathbf{e}_i\mathbf{x}_j^H\right)}{N_\ell}\mathbf{R}_\ell^{1/2}\hat{\mathbf{Q}}_\ell.$$

A direct application of these techniques and the resolvent identity $z_\ell\hat{\mathbf{Q}}_\ell = \hat{\mathbf{Q}}_\ell\hat{\mathbf{R}}_\ell - \mathbf{I}_M$ allows us to write

$$\begin{aligned}\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\Psi(u)\phi_1\right] = &-\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\right]\Psi(u)\phi_1\right]\\ &-\frac{\varsigma}{N_1}\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\Psi(u)\phi_1\right] + \mathbb{E}_1\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\Psi(u)\phi_1\right]\\ &-\frac{1+\varsigma}{N_1}u\mathrm{j}\frac{1}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-} g(z_1', z_2')\times\\ &\times\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\Psi(u)\phi_1\right]dz_1'dz_2' + \mathcal{O}\left(M^{-\mathbb{N}}\right)\end{aligned}$$

where we recall that $\hat{\mathbf{Q}}_{j'} \equiv \hat{\mathbf{Q}}_j(z_j')$. Hence, by using the fact that

$$\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\bar{\mathbf{Q}}_j\right] = \frac{\omega_j}{z_j} - 1,$$

for $j \in \{1, 2\}$, together with the resolvent identity from above and the error quantity

$$\alpha_1(z_1) = \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}\right]\phi_1 - \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\right],$$

we can then manipulate the above expectation into

$$\begin{aligned}\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\Psi(u)\phi_1\right] &= \bar{\mathbf{Q}}_1\mathbb{E}_1\left[\Psi(u)\phi_1\right] + \\ &+ \frac{z_1}{\omega_1}\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\Psi(u)\phi_1\alpha_1(z_1)\right] \\ &+ \frac{z_1}{\omega_1}\frac{\varsigma}{N_1}\mathbb{E}_1\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\Psi(u)\phi_1\right] \\ &+ \frac{z_1}{\omega_1}\frac{1+\varsigma}{N_1}\mathrm{j}u\frac{1}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-} g(z_1',z_2')\times \\ &\times \mathbb{E}_1\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\Psi(u)\phi_1\right]dz_1'dz_2' + \mathcal{O}\left(M^{-\mathbb{N}}\right). \quad \text{(A.8)}\end{aligned}$$

Moreover, multiplying (A.8) by $\hat{\mathbf{Q}}_2\phi_2$, taking expectation with respect to $\mathbf{X}_2$ and using again (A.8) after interchanging the two indices, we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathrm{tr}\left(\hat{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\phi_1\phi_2 - \bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_2\right)\Psi(u)\right] \\ &= \frac{z_1}{\omega_1}\mathbb{E}\left[\frac{1}{N_1}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\right]\phi_1\phi_2\left[N_1\alpha_1(z_1)\right]\Psi(u)\right] \\ &+ \frac{z_2}{\omega_2}\mathbb{E}\left[\frac{1}{N_2}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\hat{\mathbf{R}}_2\bar{\mathbf{Q}}_2\right]\phi_1\phi_2\left[N_2\alpha_2(z_2)\right]\Psi(u)\right] \\ &+ \varsigma\frac{z_2}{\omega_2}\mathbb{E}\left[\frac{1}{N_2}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\mathbf{R}_2\hat{\mathbf{Q}}_2\hat{\mathbf{R}}_2\bar{\mathbf{Q}}_2\phi_1\phi_2\right]\Psi(u)\right] \\ &+ \varsigma\frac{z_1}{\omega_1}\mathbb{E}\left[\frac{1}{N_1}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\phi_1\phi_2\right]\Psi(u)\right] \\ &+ \mathrm{j}u\frac{z_1}{\omega_1}\frac{1+\varsigma}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-} g(z_1',z_2')\times \\ &\times \mathbb{E}\left[\frac{1}{N_1}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\right]\Psi(u)\phi_1\phi_2\right]dz_1'dz_2' \\ &+ \mathrm{j}u\frac{z_2}{\omega_2}\frac{1+\varsigma}{(2\pi\mathrm{j})^2}\oint_{C_1^-}\oint_{C_2^-} g(z_1',z_2')\times \\ &\times \mathbb{E}\left[\frac{1}{N_2}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\mathbf{R}_2\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{R}}_2\bar{\mathbf{Q}}_2\right]\Psi(u)\phi_1\phi_2\right]dz_1'dz_2' \\ &+ \mathcal{O}\left(M^{-\mathbb{N}}\right).\end{aligned}$$

where we have defined $\alpha_2(z_2)$ in the same way as $\alpha_1(z_1)$.

Now, we still need to further investigate the two terms that depend on $\alpha_\ell(z_\ell)$.

By multiplying (A.8) by $\mathbf{R}_1$ and taking traces, we immediately obtain

$$\begin{aligned}\mathbb{E}_1\left[\Psi(u)N_1\alpha_1(z_1)\right] = &\frac{\mathbb{E}_1\left[\beta_1\left(z_1,z_1\right)\Psi(u)N_1\alpha_1(z_1)\right]}{1-\gamma_1\left(z_1,z_1\right)}\\ &+\varsigma\frac{z_1}{\omega_1}\frac{\mathbb{E}_1\left[N_1^{-1}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\phi_1\right]\Psi(u)\right]}{1-\gamma_1\left(z_1,z_1\right)}+\\ &+\frac{1}{1-\gamma_1\left(z_1,z_1\right)}\frac{z_1}{\omega_1}\frac{1+\varsigma}{N_1}\mathrm{j}u\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{C_1^-}\oint_{C_2^-}g(z_1',z_2')\times\\ &\times\mathbb{E}_1\left[\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_1\right]\Psi(u)\phi_1\right]dz_1'dz_2'+\\ &+\mathcal{O}\left(M^{-\mathbb{N}}\right)\end{aligned}\tag{A.9}$$

where we have introduced the definitions

$$\begin{aligned}\beta_j\left(z_j,z_j'\right) &= \frac{1}{N_j}\mathrm{tr}\left[\mathbf{Q}_j\mathbf{R}_j\hat{\mathbf{Q}}_{j'}\hat{\mathbf{R}}_j\right]\phi_j-\gamma_j\left(z_j,z_j'\right)\\ \gamma_j\left(z_j,z_j'\right) &= \frac{1}{N_j}\mathrm{tr}\left[\mathbf{Q}_j\mathbf{R}_j\mathbf{Q}_{j'}\mathbf{R}_j\right]\end{aligned}$$

and where we have used the well known fact that $\sup_M\sup_{(z,z')\in C_j\times C_j}\left|\gamma_j\left(z,z'\right)\right|<1$ (from Cauchy-Schwarz and Lemma D.2) so that the quantity $1-\gamma_j\left(z_j,z_j'\right)$ is always invertible. In order to further simplify the notation, we will denote from now on $\gamma_{jj}\equiv\gamma_j(z_j,z_j)$, $\gamma_{jj'}\equiv\gamma_j(z_j,z_j')$ and $\gamma_{j'j'}\equiv\gamma_j(z_j',z_j')$. According to [75, Lemma 11], the expectations of $\alpha_\ell(z_\ell)$ and $\beta_\ell(z_\ell,z_\ell)$ are $\mathcal{O}(M^{-1})$, and their variance decays as $\mathcal{O}(M^{-2})$. Hence, we can write (by Cauchy-Schwarz)

$$\left|N_\ell\mathbb{E}\left[\alpha_\ell(z_\ell)\beta_\ell(z_\ell,z_\ell)\Psi(u)\right]\right|^2\leq N_\ell^2\mathbb{E}\left[\left|\alpha_\ell(z_\ell)\right|^2\right]\mathbb{E}\left[\left|\beta_\ell(z_\ell,z_\ell)\right|^2\right]=\mathcal{O}(M^{-2})$$

so that the first term on the right hand side of (A.9) decays as $\mathcal{O}(M^{-1})$ and therefore

$$\begin{aligned}\mathbb{E}\left[N_1\alpha_1(z_1)\Psi(u)\right] = &\varsigma\frac{\mathbb{E}\left[\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\mathbf{Q}_1\phi_1\right]\right]}{N_1(1-\gamma_{11})}\mathbb{E}[\Psi(u)]\\ &+\frac{1}{1-\gamma_{11}}\frac{1+\varsigma}{N_1}\mathrm{j}u\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{C_1^-}\oint_{C_2^-}g(z_1',z_2')\times\\ &\times\mathbb{E}\left[\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\mathbf{Q}_1\phi_1\phi_2\right]\right]dz_1'dz_2'\mathbb{E}\left[\Psi(u)\right]+\mathcal{O}\left(M^{-1}\right)\end{aligned}$$

where we have used the variance bounds in Lemmas D.4 and D.5 in Appendix D.2.

Now, using the results in Lemmas D.3 to D.5 (regarding the bound on the corresponding variances) together with the above result, one can readily see that

$$\begin{aligned}\frac{d\mathbb{E}\left[\Psi(u)\right]}{du} = \frac{-\mathrm{j}\varsigma}{4\pi^2}\oint_{C_1^-}\oint_{C_2^-} g(z_1,z_2)\mu\left(z_1,z_2\right)dz_1dz_2\mathbb{E}\left[\Psi(u)\right] \\ - u\frac{1+\varsigma}{(2\pi\mathrm{j})^4}\oint_{C_1^-}\oint_{C_2^-}\oint_{C_1^-}\oint_{C_2^-} g(z_1,z_2)g(z_1',z_2')\times \\ \times\sigma^2\left(z_1,z_1',z_2,z_2'\right)dz_1dz_2dz_1'dz_2'\mathbb{E}\left[\Psi(u)\right] + \mathcal{O}\left(M^{-1}\right) \quad \text{(A.10)}\end{aligned}$$

where

$$\begin{aligned}\mu\left(z_1,z_2\right) &= \mu_{12}\left(z_1,z_2\right) + \mu_{21}\left(z_1,z_2\right) \\ \sigma^2\left(z_1,z_1',z_2,z_2'\right) &= \chi_{12}^2\left(z_1,z_1',z_2,z_2'\right) + \chi_{21}^2\left(z_1,z_1',z_2,z_2'\right)\end{aligned}$$

and

$$\begin{aligned}\mu_{12}\left(z_1,z_2\right) = \frac{1}{1-\gamma_{12}}\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\mathbf{Q}_1\hat{\mathbf{Q}}_2\phi_1\phi_2\right]\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1\phi_1\hat{\mathbf{R}}_1^2\hat{\mathbf{Q}}_1^2\right] \\ + \frac{\mathbb{E}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\hat{\mathbf{R}}_1\hat{\mathbf{Q}}_1^2\hat{\mathbf{Q}}_2\phi_1\phi_2\right]}{N_1} \quad \text{(A.11)} \\ \mu_{21}\left(z_1,z_2\right) = \frac{1}{1-\gamma_{22}}\frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\hat{\mathbf{R}}_2\mathbf{Q}_2\phi_1\phi_2\right]\frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\mathbf{R}_2^2\mathbf{Q}_2\hat{\mathbf{R}}_2\hat{\mathbf{Q}}_2^2\phi_2\right] \\ + \frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\mathbf{R}_2\mathbf{Q}_2\hat{\mathbf{R}}_2\hat{\mathbf{Q}}_2^2\phi_1\phi_2\right] \quad \text{(A.12)}\end{aligned}$$

$$\begin{aligned}\chi_{12}^2\left(z_1,z_1',z_2,z_2'\right) = \frac{1}{(1-\gamma_{11})N_1}\mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\mathbf{Q}_1\hat{\mathbf{Q}}_2\phi_1\phi_2\right]\times \quad \text{(A.13)} \\ \times\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1^2\hat{\mathbf{Q}}_{1'}^2\hat{\mathbf{Q}}_{2'}\phi_1\phi_2\right] \\ + \frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_1\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\phi_2\right] \quad \text{(A.14)} \\ \chi_{21}^2\left(z_1,z_1',z_2,z_2'\right) = \frac{1}{1-\gamma_{22}}\frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Q}}_2\hat{\mathbf{R}}_2\mathbf{Q}_2\bar{\mathbf{Q}}_1\phi_2\right]\times \quad \text{(A.15)} \\ \times\frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_2\mathbf{R}_2^2\hat{\mathbf{Q}}_2\hat{\mathbf{Q}}_{2'}^2\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_2\phi_1\phi_2\right] \\ + \frac{1}{N_2}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_2\bar{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\mathbf{R}_2\hat{\mathbf{Q}}_{2'}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{Q}}_{2'}\hat{\mathbf{R}}_2\phi_1\phi_2\right] \quad \text{(A.16)}\end{aligned}$$

At this point, it is worth noting that this definition of $d\mathbb{E}\left[\Psi(u)\right]/du$ is closely related to the results presented in Section 2.3. Particularly, $\mathfrak{m}_M$ and σ_M are associated to the first and, respectively, second integrals on the right hand side of (A.10). Then, it is also interesting to note that whenever $\varsigma = 0$ (complex-valued observations)

the first term goes to zero and also $\mathfrak{m}_M = 0$. In what follows, we analyze the terms from (A.11)-(A.16). Observe that these four terms are essentially different and should be analyzed separately.

Regarding the functions $\mu_{12}(z_1, z_2)$ and $\mu_{21}(z_1, z_2)$, a direct application of Lemmas D.3 and D.4 shows

$$\left(\frac{z_1}{\omega_1}\right)^2 \frac{\mathbb{E}\mathrm{tr}\left[\bar{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\phi_1\right]}{N_1} = \gamma_{11} + \frac{\omega_1 \mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3\right]}{(1-\gamma_{11})N_1} + \mathcal{O}(M^{-1})$$
$$= \frac{\mathrm{tr}\left[\bar{\mathbf{Q}}_2\mathbf{R}_1\bar{\mathbf{Q}}_1^2\right]}{N_1} + \frac{z_1}{1-\gamma_{11}}\frac{\mathrm{tr}\left[\bar{\mathbf{Q}}_2\mathbf{R}_1\bar{\mathbf{Q}}_1^3\right]}{N_1} + \mathcal{O}(M^{-1})$$

and equivalent expressions can be obtained by swapping the indices $1, 2$. A direct application of these identities together with the fact that $\mathbf{Q}_j(\mathbf{R}_j - \omega_j\mathbf{I}_M) = \mathbf{I}_M$ shows that

$$\mu_{12}(z_1, z_2) = \frac{\omega_1}{z_1}\frac{\omega_2}{z_2}\mathfrak{m}_1(\omega_1, \mathbf{Q}_2) + \mathcal{O}(M^{-1})$$
$$\mu_{21}(z_1, z_2) = \frac{\omega_1}{z_1}\frac{\omega_2}{z_2}\mathfrak{m}_2(\omega_2, \mathbf{Q}_1) + \mathcal{O}(M^{-1})$$

where we notice that $\gamma_{jj} = \Gamma(\omega_j, \omega_j)$ in the statement of the theorem.

In order to deal with the variance terms, we consider the following random function

$$\psi_\ell(z_\ell, z'_\ell; \mathbf{A}, \mathbf{B}) = \frac{1}{N_\ell}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_\ell\mathbf{A}\hat{\mathbf{Q}}_\ell\mathbf{R}_\ell\hat{\mathbf{Q}}_{\ell'}\mathbf{B}\hat{\mathbf{Q}}_{\ell'}\hat{\mathbf{R}}_\ell\phi_\ell\right]$$
$$+ \frac{\mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathbf{Q}_\ell^2\right]}{(1-\gamma_{\ell\ell})N_\ell}\frac{1}{N_\ell}\mathbb{E}\mathrm{tr}\left[\mathbf{Q}_\ell\mathbf{R}_\ell\hat{\mathbf{Q}}_\ell\mathbf{R}_\ell\hat{\mathbf{Q}}_{\ell'}\mathbf{B}\hat{\mathbf{Q}}_{\ell'}\hat{\mathbf{R}}_\ell\phi_\ell\right]$$

where $\mathbf{A}, \mathbf{B}$ are two $M \times M$ matrices that are either deterministic or independent of $\mathbf{X}_\ell$. Using Lemma D.3, one can express the two variance terms as specific instances of the above function, namely

$$\chi^2_{12}(z_1, z'_1, z_2, z'_2) = \psi_1(z_1, z'_1; \hat{\mathbf{Q}}_2\phi_2, \hat{\mathbf{Q}}_{2'}\phi_2) + \mathcal{O}(M^{-1})$$
$$\chi^2_{21}(z_1, z'_1, z_2, z'_2) = \psi_2(z_2, z'_2; \bar{\mathbf{Q}}_1, \bar{\mathbf{Q}}_1) + \mathcal{O}(M^{-1})$$

It is therefore sufficient to analyze the behavior of $\psi_\ell(z_\ell, z'_\ell; \mathbf{A}, \mathbf{B})$ for general bounded

matrices $\mathbf{A}, \mathbf{B}$. By Lemma D.5 presented below, we readily obtain

$$\begin{aligned}\frac{z_\ell z'_\ell \psi_\ell(z_\ell, z'_\ell; \mathbf{A}, \mathbf{B})}{\omega_\ell \omega_{\ell'}} &= \left(\frac{\gamma_{\ell\ell'}^{(1,2)}\gamma_{\ell\ell'}^{(2,1)}}{(1-\gamma_{\ell\ell'})^2} + \frac{\gamma_{\ell\ell'}^{(2,2)}}{1-\gamma_{\ell\ell'}}\right)\phi_\ell(\mathbf{A})\,\phi_{\ell'}(\mathbf{B}) \\ &+ \frac{\phi_\ell(\mathbf{A})}{1-\gamma_{\ell\ell'}}\frac{\gamma_{\ell\ell'}^{(2,1)}}{1-\gamma_{\ell\ell'}}\frac{\mathrm{tr}\left[\mathbf{B}\mathbf{R}_\ell\mathbf{Q}_\ell\mathbf{R}_\ell\mathbf{Q}_{\ell'}^2\right]}{N_\ell} \\ &+ \frac{\phi_\ell(\mathbf{A})}{1-\gamma_{\ell\ell'}}\frac{\mathrm{tr}\left[\mathbf{B}\mathbf{R}_\ell\mathbf{Q}_\ell^2\mathbf{R}_\ell\mathbf{Q}_{\ell'}^2\right]}{N_\ell} \\ &+ \frac{\phi_{\ell'}(\mathbf{B})}{1-\gamma_{\ell\ell'}}\frac{\gamma_{\ell\ell'}^{(1,2)}}{1-\gamma_{\ell\ell'}}\frac{\mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathbf{Q}_\ell^2\mathbf{R}_\ell\mathbf{Q}_{\ell'}\right]}{N_\ell} \\ &+ \frac{\phi_{\ell'}(\mathbf{B})}{1-\gamma_{\ell\ell'}}\frac{\mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathbf{Q}_\ell^2\mathbf{R}_\ell\mathbf{Q}_{\ell'}^2\right]}{N_\ell} + \frac{\mathrm{tr}[\mathbf{A}\mathbf{Z}_\ell^2\mathbf{Q}_{\ell'}]\mathrm{tr}[\mathbf{B}\mathbf{Z}_\ell^{(2,1)}\mathbf{Q}_{\ell'}]}{N_\ell^2(1-\gamma_{\ell\ell'})^2} \\ &+ \frac{\mathrm{tr}[\mathbf{A}\mathbf{Z}_\ell\mathbf{Q}_{\ell'}\mathbf{B}\mathbf{Z}_\ell\mathbf{Q}_{\ell'}]}{N_\ell(1-\gamma_{\ell\ell'})} + \mathcal{O}(M^{-1}) \quad \text{(A.17)}\end{aligned}$$

where we have defined

$$\phi_\ell(\mathbf{A}) = \frac{\omega_\ell \mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathbf{Q}_\ell^2\right]}{N_\ell(1-\gamma_{\ell\ell})} \text{ and } \gamma_{\ell\ell'}^{(r,s)} = \frac{\mathrm{tr}\left[\mathbf{R}_\ell\mathbf{Q}_\ell^r\mathbf{R}_\ell\mathbf{Q}_{\ell'}^s\right]}{N_\ell}$$

so that in particular $\gamma_{\ell\ell'} = \gamma_{\ell\ell'}^{(1,1)}$ and where we have used the identities

$$\frac{1}{N_\ell}\mathrm{tr}\left[\mathbf{R}_\ell^2\mathbf{Q}_\ell^2\right] = \gamma_{\ell\ell}$$

and

$$\frac{1}{N_\ell}\mathrm{tr}\left[\mathbf{R}_\ell\bar{\mathbf{Q}}_\ell^2\right] = \frac{1}{z_\ell^2}[z_\ell - \omega_\ell(1-\gamma_{\ell\ell})].$$

Particularizing the above expression we see that

$$\chi_{21}^2(z_1, z'_1, z_2, z'_2) = \frac{\omega_1\omega_{1'}\omega_2\omega_{2'}}{z_1 z'_1 z_2 z'_2}\sigma_2^2(\omega_2, \omega'_2, \mathbf{Q}_1, \mathbf{Q}_{1'}) + \mathcal{O}(M^{-1}).$$

On the other hand, using again Lemma D.4,

$$\chi_{12}^2(z_1, z'_1, z_2, z'_2) = \frac{\omega_1}{z_1}\frac{\omega_{1'}}{z'_1}\frac{\omega_2}{z_2}\frac{\omega_{2'}}{z'_2} \times \left[\sigma_1^2(\omega_1, \omega'_1, \mathbf{Q}_2, \mathbf{Q}_{2'}) + \frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_{1'}\mathbf{Q}_1\mathbf{R}_2\mathbf{Q}_2\mathbf{Q}_{2'})\right]}{N_1 N_2(1-\gamma_{11'})(1-\gamma_{22'})}\right] + \mathcal{O}(M^{-1})$$

which concludes the derivation.

In order to finalize the proof of the theorem, we need to prove that

$$\limsup_{M\to\infty} |\mathfrak{m}_M| < \infty$$

and

$$\limsup_{M\to\infty} \sigma_M^2 < \infty.$$

Since we are assuming that

$$\liminf_{M\to\infty} \sigma_M^2 > 0,$$

the CLT will follow from [72]. To show that $\mathfrak{m}_M$ is asymptotically bounded we see from (2.17) that it is sufficient to show that

$$\liminf_{M\to\infty} \sup_{C_1\times C_2} |\omega_1\omega_2/(z_1z_2)\mathfrak{m}(\omega_1,\omega_2)| < \infty.$$

This follows easily from Lemmas D.1 and D.2 since, for instance

$$\begin{aligned}|\mathfrak{m}_1(\omega_1,\mathbf{Q}_2)| &\leq \frac{|\omega_1|\,|\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3\right]|\,|\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1^2\mathbf{Q}_2\right]|}{N_1^2\left(1-|\Gamma_1(\omega_1)|\right)^2} + \frac{1}{1-|\Gamma_1(\omega_1)|}\left|\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^3\mathbf{Q}_2\right]\right| \\ &\leq \left(\frac{|\omega_1|\,\|\mathbf{Q}_1\|^5\,\|\mathbf{Q}_2\|}{N_1\left(1-|\Gamma_1(\omega_1)|\right)^2}\mathrm{tr}\left[\mathbf{R}_1\right] + \frac{\|\mathbf{Q}_1\|^3\,\|\mathbf{Q}_2\|}{\left(1-|\Gamma_1(\omega_1)|\right)}\right)\frac{\mathrm{tr}\left[\mathbf{R}_1^2\right]}{N_1}\end{aligned}$$

where the second inequality follows from the fact that $\mathrm{tr}\left[\mathbf{AB}\right] \leq \|\mathbf{B}\|\,\mathrm{tr}\left[\mathbf{A}\right]$ for Hermitian positive definite $\mathbf{A}$, together with $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$. All the terms on the right hand side of the above inequality are bounded by Lemmas D.1 -D.2, leading to the conclusion that

$$\limsup_{M\to\infty} |\mathfrak{m}_M| < \infty.$$

Regarding the upper bound on σ_M^2 , we can use the expression in (2.19) to see that it is sufficient to show that

$$\liminf_{M\to\infty} \sup_{C_1^2\times C_2^2} \left|\Sigma^2(\omega_1,\omega_2,\omega_1',\omega_2')\right| < \infty,$$

which amounts to finding an upper bound on each of the three terms in (2.20). The bounds on these three terms can be found using again Lemmas D.1-D.2 together with the fact that, by Cauchy-Schwarz, $\left|\Gamma_j(\omega_j,\omega_j')\right| \leq \sqrt{\Gamma_j(\omega_j)\Gamma_j(\omega_j')}$.

A.2 Derivation of the Asymptotic Second-Order Mean and Variances

A.2.1 Euclidean distance

Let us evaluate the asymptotic mean and variance of the Euclidean distance between sample covariance matrices. To that effect, one must carry out the integrals

in (2.17)-(2.19) for $g(z_1, z_2) = (z_1 - z_2)^2$ under the assumption that all the contours enclose $\{0\}$. Hence, one can follow exactly the same approach that was used in Section 2.2.1. The main idea is to first use the change of variable $z \mapsto \omega = \omega_j(z)$. The resulting contour $\mathcal{C}^-_{\omega_j} = \omega_j(\mathcal{C}^-_j)$ encloses all the singularities of the integrand, so that one can apply a second change of variables $\omega \mapsto \zeta(\omega) = \omega^{-1}$ in a way that $\zeta(\mathcal{C}^-_{\omega_j})$ encloses zero and no other singularity. By direct application of this technique we may find

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}^-_1} \frac{\omega_1}{z_1} \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 = \frac{\mathrm{tr}\left[\mathbf{R}_2^2 \mathbf{Q}_2^3\left(\omega_2\right) \Omega_2\left(\omega_2; \mathbf{I}_M\right)\right]}{N_2\left(1 - \Gamma_2(\omega_2)\right)} \tag{A.18}$$

together with

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}^-_1} \omega_1 \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 = \frac{1}{N_2} \frac{\mathrm{tr}\left[\mathbf{R}_2^2 \mathbf{Q}_2^3\left(\omega_2\right) \Omega_2\left(\omega_2; \mathbf{R}_1\right)\right]}{1 - \Gamma_2(\omega_2)}.$$

Repeating the approach with respect to the second variable z_2 we find

$$\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathcal{C}^-_1} \oint_{\mathcal{C}^-_2} z_1 z_2 \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 dz_2 = 0$$

and

$$\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathcal{C}^-_1} \oint_{\mathcal{C}^-_2} z_2^2 \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 dz_2 = \frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right].$$

We can therefore conclude that $\mathfrak{m}_M$ takes the form in (2.23).

The same approach can be followed to evaluate the asymptotic variance. In this case, we will use the fact that, for any $\omega' \in \mathcal{C}^-_{\omega_j} = \omega_j(\mathcal{C}^-_j)$, the equation $\Gamma_j(\omega, \omega') = 1$ has all its solutions inside $\mathcal{C}^-_{\omega_j}$, see [75]. Therefore, all the singularities of the term $\sigma_j^2\left(\omega, \omega'; \mathbf{A}, \mathbf{B}\right)$ are located inside $\mathcal{C}^-_{\omega_j}$. We can therefore use the same integration technique as before by first applying the change of variables $z_j \mapsto \omega_j = \omega_j(z_j)$ and noting that the resulting contour $\mathcal{C}^-_{\omega_j} = \omega_j(\mathcal{C}^-_j)$ encloses all the singularities of the integrand, because $\mathcal{C}^-_j$ is built to enclose zero. Applying then a second change of variables $\omega_j \mapsto \zeta(\omega_j) = \omega_j^{-1}$ one can see that $\zeta(\mathcal{C}^-_{\omega_j})$ encloses zero and no other

singularity. Computing the corresponding residue, one easily finds that

$$\begin{aligned}\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_1^-}(z_1-z_2)^2\frac{\omega_1}{z_1}\sigma_1^2(\omega_1,\omega_1';\mathbf{A},\mathbf{B})\,dz_1 = \\ = \left(2z_2+\omega_1'\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\omega_1')\right]\right)\times \\ \times\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\omega_1')\,\mathbf{A}\mathbf{R}_1\mathbf{Q}_1(\omega_1')\,\Omega_1(\omega_1';\mathbf{B})\right] \\ -\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1(\omega_1')\,\mathbf{A}\mathbf{R}_1\mathbf{Q}_1(\omega_1')\,\Omega_1(\omega_1';\mathbf{B})\right] \\ -\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\omega_1')\,\mathbf{A}\mathbf{R}_1^2\mathbf{Q}_1(\omega_1')\,\Omega_1(\omega_1';\mathbf{B})\right] \\ +\omega_1'\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\omega_1')\,\mathbf{A}\right]\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^2(\omega_1')\,\Omega_1(\omega_1';\mathbf{B})\right].\end{aligned}$$

We can now multiply by $(z_1'-z_2')^2\,\omega_1'/z_1'$ and integrate with respect to z_1'. Following exactly the same approach as above, after some algebra, we obtain

$$\begin{aligned}\frac{1}{(2\pi\mathrm{j})^2}\oint_{\mathcal{C}_1^-}\oint_{\mathcal{C}_1^-}(z_1-z_2)^2(z_1'-z_2')^2\frac{\omega_1}{z_1}\frac{\omega_1'}{z_1'}\sigma_1^2(\omega_1,\omega_1';\mathbf{A},\mathbf{B})\,dz_1dz_1' = \\ = \frac{1}{N_1}\mathrm{tr}\Big[\mathbf{R}_1\left(\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\right]-2z_2\right)\mathbf{A}+\mathbf{R}_1\mathbf{A}+\mathbf{A}\mathbf{R}_1+\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{A}\right]\mathbf{I}_M\right)\times \\ \times\mathbf{R}_1\left(\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\right]-2z_2'\right)\mathbf{B}+\mathbf{B}\mathbf{R}_1+\mathbf{R}_1\mathbf{B}+\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{B}\right]\mathbf{I}_M\right)\Big] \\ +\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{A}\right]\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{B}\right]+\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\right]\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{A}\mathbf{R}_1\mathbf{B}\right]. \quad \text{(A.19)}\end{aligned}$$

Now, to solve the integrals with respect to z_2, we apply the following result

$$\begin{aligned}\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_2^-}\mathbf{Q}_2(\omega_2)\frac{\omega_2}{z_2}dz_2 = \mathbf{I}_M \\ \frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_2^-}z_2\mathbf{Q}_2(\omega_2)\frac{\omega_2}{z_2}dz_2 = \mathbf{R}_2.\end{aligned}$$

A direct application of the above identities allows us to obtain

$$\begin{aligned}\frac{1}{(2\pi\mathrm{j})^4}\oint_{\mathcal{C}_1^-}\oint_{\mathcal{C}_1^-}\oint_{\mathcal{C}_2^-}\oint_{\mathcal{C}_2^-}(z_1-z_2)^2(z_1'-z_2')^2\left(\frac{\omega_1}{z_1}\frac{\omega_1'}{z_1'}\frac{\omega_2}{z_2}\frac{\omega_2'}{z_2'}\right)\sigma_1^2(\omega_1,\omega_1';\mathbf{A},\mathbf{B})\,dz_1dz_1'dz_2dz_2' \\ = 2\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\right]\right)^2+4\frac{1}{N_1}\mathrm{tr}\left[\left(\mathbf{R}_1\left(\mathbf{R}_1+\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\right]\mathbf{I}_M-\mathbf{R}_2\right)\right)^2\right]. \quad \text{(A.20)}\end{aligned}$$

Obviously, the integral of the term $\sigma_2^2(\omega_2,\omega_2';\mathbf{Q}_1(\omega_1),\mathbf{Q}_1(\omega_1'))$ can be obtained from the above by simply swapping the two indices.

We finally evaluate the last integral, first by noting that, using the same integration technique, we obtain

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_1^-} (z_1 - z_2)^2 \frac{\omega_1}{z_1} \frac{\mathrm{tr}^2 \left[\mathbf{A}\mathbf{Q}_1(\omega_1)\right]}{1 - \Gamma_1(\omega_1, \omega_1')} dz_1 =$$
$$= \left(2z_2 - \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\right]\right) \mathrm{tr}^2\left[\mathbf{A}\right] - 2\mathrm{tr}\left[\mathbf{A}\right]\mathrm{tr}\left[\mathbf{A}\mathbf{R}_1\right]$$
$$+ \mathrm{tr}^2\left[\mathbf{A}\right] \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2 \mathbf{Q}_1(\omega_1')\right]$$

and therefore replacing $\mathbf{A}$ with $\mathbf{B}\mathbf{Q}_2(\omega_2)\mathbf{A}$ and using the same technique we find

$$\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathcal{C}_1^-} \oint_{\mathcal{C}_2^-} (z_1 - z_2)^2 \frac{\omega_1}{z_1}\frac{\omega_2}{z_2} \times$$
$$\times \frac{\mathrm{tr}^2\left[\mathbf{A}\mathbf{Q}_1(\omega_1)\mathbf{B}\mathbf{Q}_2(\omega_2)\right]}{(1 - \Gamma_1(\omega_1, \omega_1'))(1 - \Gamma_2(\omega_2, \omega_2'))} dz_1 dz_2 = -2\mathrm{tr}^2\left[\mathbf{A}\mathbf{B}\right]$$

and consequently, after some manipulation,

$$\frac{1}{(2\pi \mathrm{j})^4} \oint_{\mathcal{C}_1^-} \oint_{\mathcal{C}_1^-} \oint_{\mathcal{C}_2^-} \oint_{\mathcal{C}_2^-} (z_1 - z_2)^2 (z_1' - z_2')^2 \frac{\omega_1 \omega_1' \omega_2 \omega_2'}{z_1 z_1' z_2 z_2'} \times$$
$$\times \varrho(\omega_1, \omega_1', \omega_2, \omega_2') dz_1 dz_2 dz_1' dz_2' = 4\frac{\mathrm{tr}^2[\mathbf{R}_1\mathbf{R}_2]}{N_1 N_2}.$$

Adding the three integrals, we can conclude that the asymptotic variance takes the expression in (2.24).

A.2.2 Symmetrized KL distance

Let us consider the integrals in (2.17)-(2.19) for $g(z_1, z_2) = z_1/(2z_2) + z_2/(2z_1) - 1$, where only the contours of the last integral enclose $\{0\}$. The third term can be evaluated by considering the integral in (A.18) and integrating with respect to z_2. Using the change of variable $z_2 \mapsto \zeta = \omega_2^{-1}(z_2)$ one can readily see that

$$\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathcal{C}_1^-} \oint_{\mathcal{C}_2^-} \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2) dz_1 dz_2 = 0$$

and we only need to evaluate the other two terms. We begin by noticing that, by applying the change of variable $z_1 \mapsto \omega = \omega_1(z_1)$

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} z_1 \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2) dz_1 \tag{A.21}$$
$$= \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_1}^-} \frac{\omega_1 \omega_2}{z_2} \mathfrak{m}(\omega_1, \omega_2) \left(1 - \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2 \mathbf{Q}_1^2(\omega_1)\right]\right) d\omega_1 \tag{A.22}$$

where $\mathrm{C}_{\omega_1}^- = \omega_1(\mathrm{C}_1^-)$. It can readily be seen that the integrand is holomorphic at $\mu_0^{(1)}$ and therefore $\mathrm{C}_{\omega_1}^-$ contains all the singularities. This means that we can enlarge the contour and apply the change of variables $\omega \mapsto \zeta(\omega) = \omega^{-1}$ so that the transformed contour only encloses zero and no other singularity. This readily shows that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} z_1 \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 = \frac{\omega_2}{z_2} \mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right).$$

Now, multiplying by z_2^{-1} and by applying the change of variable $z_2 \mapsto \omega = \omega_2\left(z_2\right)$ and using the integration by parts formula we see that

$$\begin{aligned}
&\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \frac{z_1}{z_2} \frac{\omega_1 \omega_2}{z_1 z_2} \mathfrak{m}\left(\omega_1, \omega_2\right) dz_1 dz_2 \\
&= \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_2}^-} \frac{\omega_2}{z_2^2} \mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right)\left(1 - \Gamma_2(\omega_2)\right) d\omega_2 \\
&= \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_2}^-} \frac{\mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right)}{z_2} d\omega_2 + \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_2}^-} \frac{\omega_2}{z_2} \mathfrak{m}_2'\left(\omega_2, \mathbf{R}_1\right) d\omega_2
\end{aligned}$$

where we have introduced the notation

$$\mathfrak{m}_j'\left(\omega, \mathbf{A}\right) = \frac{d\mathfrak{m}_j\left(\omega, \mathbf{A}\right)}{d\omega}.$$

Let us now evaluate these two integrals separately. The first one can be solved by using the fact that $\mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right)$ is holomorphic at $\mu_0^{(2)}$ so that we can expand the contour as

$$\begin{aligned}
&\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_2^-} \frac{1}{z_2} \mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right) d\omega_2 \\
&= \frac{\mathfrak{m}_2\left(\mu_0^{(2)}, \mathbf{R}_1\right)}{1 - \Gamma_2\left(\mu_0^{(2)}\right)} + \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_2}^-} \frac{1}{z_2} \mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right) d\omega_2.
\end{aligned}$$

Now, the integral on the right hand side can be shown to be zero by using the change of variables $\omega_2 \mapsto \zeta(\omega_2) = \omega_2^{-1}$. Following the same procedure, we find that

$$\begin{aligned}
&\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_2}^-} \frac{\omega_2}{z_2} \frac{d\mathfrak{m}_2\left(\omega_2, \mathbf{R}_1\right)}{d\omega_2} d\omega_2 \\
&= \frac{\mu_0^{(2)} \mathfrak{m}_2'\left(\mu_0^{(2)}, \mathbf{R}_1\right)}{1 - \Gamma_2\left(\mu_0^{(2)}\right)} + \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_2}^-} \frac{\omega_2}{z_2} \mathfrak{m}_2'\left(\omega_2, \mathbf{R}_1\right) d\omega_2
\end{aligned}$$

where the last integral is shown to be zero by using the change of variables $\omega_2 \mapsto \zeta(\omega_2) = \omega_2^{-1}$. This directly leads to (2.25).

Let us now consider the computation of the variance. We begin by noticing that

$$\begin{aligned}
&\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_1^-}\frac{\omega_1}{z_1}\sigma_1^2\left(\omega_1,\omega_1';\mathbf{A},\mathbf{B}\right)dz_1\\
&=\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_{\omega_1}^-}\sigma_1^2\left(\omega_1,\omega_1';\mathbf{A},\mathbf{B}\right)\frac{1-\Gamma_1(\omega_1)}{1-\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\right]}d\omega_1\\
&=0
\end{aligned}$$

where the last integral is evaluated by using the change of variables $\omega_1 \mapsto \zeta = \omega_1^{-1}$. Proceeding in the same way, we find

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_1^-}\frac{\omega_1}{z_1}\varrho(\omega_1,\omega_1',\omega_2')dz_1=0$$

which allows us to conclude that

$$\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{\mathcal{C}_1^-}\oint_{\mathcal{C}_2^-}\frac{\omega_1}{z_1}\frac{\omega_2}{z_2}\Sigma^2\left(\omega_1,\omega_2,\omega_1',\omega_2'\right)dz_1dz_2=0 \tag{A.23}$$

and that the same equality holds if we integrate with respect to z_j' instead of z_j. On the other hand, following the same procedure we can show

$$\begin{aligned}
&\frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_1^-}\omega_1\sigma_1^2\left(\omega_1,\omega_1';\mathbf{A},\mathbf{B}\right)dz_1\\
&=-\phi_1\left(\omega_1';\mathbf{B}\right)\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^2\left(\omega_1'\right)\mathbf{A}\right]-\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{A}\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{B}\right]
\end{aligned}$$

and

$$\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_1^-}\omega_1\omega_1'\sigma_1^2\left(\omega_1,\omega_1';\mathbf{A},\mathbf{B}\right)dz_1dz_1'=\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{A}\mathbf{R}_1\mathbf{B}\right].$$

Similarly, we obtain

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_1^-}\omega_1\varrho(\omega_1,\omega_1',\omega_2,\omega_2')dz_1=-\frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_2\left(\omega_2'\right)\right]}{N_1N_2\left(1-\Gamma_2(\omega_2,\omega_2')\right)}$$

and

$$\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_1^-}\omega_1\omega_1'\varrho(\omega_1,\omega_1',\omega_2,\omega_2')dz_1dz_1'=\frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_2\left(\omega_2'\right)\right]}{N_1N_2\left(1-\Gamma_2(\omega_2,\omega_2')\right)}.$$

We can conclude that

$$\begin{aligned}
&\frac{1}{\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_1^-}\omega_1\omega_1'\Sigma^2\left(\omega_1,\omega_2,\omega_1',\omega_2'\right)dz_1dz_1'=\\
&\qquad=\sigma_2^2\left(\omega_2,\omega_2';\mathbf{R}_1,\mathbf{R}_1\right)+\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_2\left(\omega_2\right)\mathbf{R}_1\mathbf{Q}_2\left(\omega_2'\right)\right]\\
&\qquad\qquad+\frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_2\left(\omega_2'\right)\right]}{N_1N_2\left(1-\Gamma_2(\omega_2,\omega_2')\right)}
\end{aligned} \tag{A.24}$$

and

$$\frac{1}{(2\pi\mathrm{j})^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\omega_1\omega_2'\Sigma^2\left(\omega_1,\omega_2,\omega_1',\omega_2'\right)dz_1dz_2' = \\ = -\phi_1\left(\omega_1';\mathbf{R}_2\right)\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\mathbf{Q}_1^2\left(\omega_1'\right)\mathbf{Q}_2\left(\omega_2\right)\right] \\ -\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{Q}_2\left(\omega_2\right)\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{R}_2\right] \\ -\phi_2\left(\omega_2;\mathbf{R}_1\right)\frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2^2\mathbf{Q}_2^2\left(\omega_2\right)\mathbf{Q}_1\left(\omega_1'\right)\right] \\ -\frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_1\left(\omega_1'\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{R}_1\right] \\ +\frac{1}{N_1N_2}\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1'\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\right]. \quad \text{(A.25)}$$

With the above results, let us now focus on the evaluation of the variance. Using (A.23) we can simplify the evaluation of σ_M^2 to

$$\sigma_M^2 = \frac{1+\varsigma}{(2\pi\mathrm{j})^4}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\left(\frac{z_1}{2z_2}+\frac{z_2}{2z_1}\right)\times \\ \times\left(\frac{z_1'}{2z_2'}+\frac{z_2'}{2z_1'}\right)\frac{\omega_1\omega_2}{z_1z_2}\frac{\omega_1'\omega_2'}{z_1'z_2'}\Sigma^2\left(\omega_1,\omega_2,\omega_1',\omega_2'\right)dz_1dz_2dz_1'dz_2'.$$

Now using the integration by parts formula together with the integrals in (A.24)-(A.25) we can write

$$\sigma_M^2 = \frac{1+\varsigma}{4\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_{\omega_1}^-}\oint_{\mathrm{C}_{\omega_1}^-}\frac{1}{z_1z_1'}\frac{\partial^2\left[\omega_1\omega_1'\Upsilon_{11}\left(\omega_1,\omega_1'\right)\right]}{\partial\omega_1\partial\omega_1'}d\omega_1d\omega_1' \\ +\frac{1+\varsigma}{2\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_{\omega_1}^-}\oint_{\mathrm{C}_{\omega_2}^-}\frac{1}{z_1z_2}\frac{\partial^2\left[\omega_1\omega_2\Upsilon_{12}\left(\omega_1,\omega_2\right)\right]}{\partial\omega_1\partial\omega_2}d\omega_1d\omega_2 \\ +\frac{1+\varsigma}{4\left(2\pi\mathrm{j}\right)^2}\oint_{\mathrm{C}_{\omega_2}^-}\oint_{\mathrm{C}_{\omega_2}^-}\frac{1}{z_2z_2'}\frac{\partial^2\left[\omega_2\omega_2'\Upsilon_{22}\left(\omega_2,\omega_2'\right)\right]}{\partial\omega_2\partial\omega_2'}d\omega_2d\omega_2'$$

where Υ_{11}, Υ_{22} and Υ_{12} are as defined in (2.27)-(2.28). Now, noting that the above quantities are holomorphic at $\omega_1=\mu_0^{(1)}$ and $\omega_2=\mu_0^{(2)}$ and using the fact that

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_{\omega_1}^-}\frac{1}{z_1}\frac{\partial\Upsilon_{11}\left(\omega_1,\omega_1'\right)}{\partial\omega_1}d\omega_1=0 \\ \frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_{\omega_1}^-}\frac{1}{z_1}\frac{\partial\Upsilon_{12}\left(\omega_1,\omega_2\right)}{\partial\omega_1}d\omega_1=0$$

(which follows using the integration by parts formula together with the change of variables $\omega\mapsto\zeta=\omega^{-1}$), we obtain the expression of the variance in (2.26).

A.2.3 Subspace distance

Consider next the subspace distance between two sample covariance matrices in the undersampled regime. In this case, we need to consider the integrals in (2.17)-(2.19) for $g(z_1, z_2) = 1$ where none of the contours encloses $\{0\}$. We begin by noticing that, by applying the change of variable $z_1 \mapsto \omega = \omega_1(z_1)$ we are able to write

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1\omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2)\, dz_1 = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_{\omega_1}^-} \frac{\omega_1\omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2)\,(1 - \Gamma_1(\omega_1))\, d\omega_1$$

where $\mathrm{C}_{\omega_1}^- = \omega_1(\mathrm{C}_1^-)$. This contour contains all the singularities of the integrand, except for the smallest zero of $z_1(\omega)$, that is $\mu_0^{(1)}$. Noting that $\mathfrak{m}(\omega_1, \omega_2)$ is holomorphic at $\omega_1 = \mu_0^{(1)}$ we see that the residue of the integrand at this point turns out to be $\mu_0^{(1)}\omega_2/z_2\mathfrak{m}(\mu_0^{(1)}, \omega_2)$, where we have used the fact that the derivative of $z_1(\omega_1)$ is precisely $1 - \Gamma_1(\omega_1)$. Therefore, we can write

$$\begin{aligned}\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1\omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2) dz_1 &= \mu_0^{(1)} \frac{\omega_2}{z_2} \mathfrak{m}(\mu_0^{(1)}, \omega_2) + \\ &+ \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_1}^-} \frac{\omega_1\omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2)\,(1 - \Gamma_1(\omega_1))\, d\omega_1\end{aligned}$$

where now $\mathcal{C}_{\omega_1}^-$ encloses $\mu_0^{(1)}$. The integral on the right hand side can now be solved by enlarging $\mathcal{C}_{\omega_1}^-$ and using the change of variables $\zeta(\omega) = \omega^{-1}$ so that $\zeta(\mathcal{C}_{\omega_1}^-)$ encloses only zero. Using this technique one can easily see that this integral is equal to $\frac{\omega_2}{z_2}\mathfrak{m}_2(\omega_2, \mathbf{I}_M)$. Repeating the same process with respect to z_2 we find

$$\begin{aligned}\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathrm{C}_2^-} \oint_{\mathrm{C}_1^-} \frac{\omega_1\omega_2}{z_1 z_2} \mathfrak{m}(\omega_1, \omega_2) dz_1 dz_2 &= \\ &= \mu_0^{(1)} \mathfrak{m}_1(\mu_0^{(1)}, \mathbf{R}_2\mathbf{Q}_2(\mu_0^{(2)})) + \mu_0^{(2)} \mathfrak{m}_2(\mu_0^{(2)}, \mathbf{R}_1\mathbf{Q}_1(\mu_0^{(1)}))\end{aligned}$$

which directly leads to the expression of $\mathfrak{m}_M$ in (2.29).

Now, regarding the variance we can follow the same approach as before. We begin by noting that $\mathbf{Q}_2(\omega_2)$ is holomorphic at $\omega_2 = \mu_0^{(2)}$, so that we can write

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_2^-} \frac{\omega_2}{z_2} \mathbf{Q}_2(\omega_2) dz_2 = \mu_0^{(2)} \mathbf{Q}_2(\mu_0^{(2)}) + \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_2}^-} \frac{\omega_2}{z_2} \mathbf{Q}_2(\omega_2)\,(1 - \Gamma_2(\omega_2))\, d\omega_2$$

where $\mathcal{C}_{\omega_2}^-$ now encloses $\mu_0^{(2)}$. The second integral is solved by using the change of variable $\omega_2 \mapsto \zeta = \omega_2^{-1}$, leading to

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_2}^-} \frac{\omega_2}{z_2} \mathbf{Q}_2(\omega_2)\,(1 - \Gamma_2(\omega_2))\, d\omega_2 = \mathbf{I}_M$$

so that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_2^-} \frac{\omega_2}{z_2} \mathbf{Q}_2\left(\omega_2\right) dz_2 = \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)}).$$

Since $\sigma_1^2\left(\omega_1, \omega_1'; \mathbf{Q}_2\left(\omega_2\right), \mathbf{Q}_2\left(\omega_2'\right)\right)$ is a linear function in the last two terms, we have

$$\frac{1}{(2\pi \mathrm{j})^2} \oint_{\mathrm{C}_2^-} \oint_{\mathrm{C}_2^-} \frac{\omega_2 \omega_2'}{z_2 z_2'} \sigma_1^2(\omega_1, \omega_1'; \mathbf{Q}_2(\omega_2), \mathbf{Q}_2(\omega_2')) dz_2 dz_2' = \sigma_1^2(\omega_1, \omega_1'; \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)}), \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)})). \quad \text{(A.26)}$$

On the other hand, noting that σ_1^2 is holomorphic at $\omega_1 = \mu_0^{(1)}$ and $\omega_1' = \mu_0^{(1)}$ we can write (denoting here $\mathbf{A} = \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)})$)

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1}{z_1} \sigma_1^2(\omega_1, \omega_1'; \mathbf{A}, \mathbf{A}) dz_1 dz_1' = \mu_0^{(1)} \sigma_1^2(\mu_0^{(1)}, \omega_1'; \mathbf{A}, \mathbf{A}) + + \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_1}^-} \sigma_1^2\left(\omega_1, \omega_1'; \mathbf{A}, \mathbf{A}\right) \frac{1 - \Gamma_1\left(\omega_1\right)}{1 - \frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1 \mathbf{Q}_1\left(\omega_1\right)\right]} d\omega_1$$

where now $\mathcal{C}_{\omega_1}^-$ enloses also $\mu_0^{(1)}$. Now, we consider the change of coordinates $\omega_1 \mapsto \xi = \omega_1^{-1}$ which readily shows that the integral on the right hand side of the above expresson is identically zero. Proceeding in the same way with the other variable ω_1' one can conclude that the integral of (A.26) with respect to ω_1, ω_1' is equal to $(\mu_0^{(1)})^2 \sigma_1^2(\mu_0^{(1)}, \mu_0^{(1)}; \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)}), \mathbf{R}_2 \mathbf{Q}_2(\mu_0^{(2)}))$. The integral corresponding to the term $\sigma_2^2\left(\omega_2, \omega_2'; \mathbf{Q}_1\left(\omega_1\right), \mathbf{Q}_1\left(\omega_1'\right)\right)$ is obtained by swapping indices $(1 \leftrightarrow 2)$.

It remains to compute the third integral

$$\mathcal{I} = \frac{1}{(2\pi \mathrm{j})^4} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \oint_{\mathrm{C}_2^-} \frac{\omega_1 \omega_1' \omega_2 \omega_2'}{z_1 z_1' z_2 z_2'} \times \varrho(\omega_1, \omega_1', \omega_2, \omega_2') dz_2 dz_1' dz_2'.$$

Using again the approach above we can readily see that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1}{z_1} \varrho(\omega_1, \omega_1', \omega_2, \omega_2') dz_1 = \mu_0^{(1)} \varrho(\mu_0^{(1)}, \omega_1', \omega_2, \omega_2') + \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_1}^-} \frac{\omega_1}{z_1} \varrho(\omega_1, \omega_1', \omega_2, \omega_2') dz_1$$

where the integral on the right hand side can be shown to be identically zero by the change of variables $\omega_1 \mapsto \xi = \omega_1^{-1}$. Following the same reasoning over the other integration variables, one can conclude that

$$\mathcal{I} = (\mu_0^{(1)} \mu_0^{(2)})^2 \varrho(\mu_0^{(1)}, \mu_0^{(1)}, \mu_0^{(2)}, \mu_0^{(2)})$$

so that we can conclude that the variance takes the form in (2.30).

Appendix B

Appendix for Chapter 3

B.1 Proof of Proposition 3.1

We start by noticing that we can express $f_j^{(l)}(\mathbf{R}_j)$ as in (3.2), so that proving Proposition 3.1, it is the same as proving

$$\sup_{z\in\mathrm{C}}\left\{f_j^{(l)}(\omega_j(z))\mathbf{Q}_j(\omega_j(z))\,\omega_j'(z)-\hat{h}_j^{(l)}(z)\hat{\mathbf{Q}}_j(z)\right\}\asymp 0.$$

From Theorem 2.1 together with (2.4), we know that $\omega_j(z)\,\mathbf{Q}_j(\omega_j(z))\asymp z\hat{\mathbf{Q}}_j(z)$ for $z\in\mathbb{C}^+$. Then, by using the fact that $z\hat{\mathbf{Q}}_j(z)=-\mathbf{I}_M+\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)$ (and equivalently for $\mathbf{Q}_j(\omega_j(z))$), we obtain $\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))\asymp\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)$. By Montel's theorem, one can extend this result to uniform convergence on C, so that

$$\sup_{z\in\mathrm{C}}\left|\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))\right]-\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right|\to 0$$

almost surely. Now, noting that

$$\omega_j(z)-\hat{\omega}_j(z)=\omega_j(z)\,\frac{\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))-\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]}{N_j\left(1-\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right)}$$

and using the fact that $\sup_{z\in\mathrm{C}}\sup_M|\omega_j(z)|<+\infty$, together with (see Lemma D.1)

$$\inf_{z\in\mathrm{C}}\inf_M\left|1-\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right|>0,$$

we can conclude that $\sup_{z\in\mathrm{C}}|\omega_j(z)-\hat{\omega}_j(z)|\to 0$ almost surely. Since these functions are holomorphic and the above properties can be extended to an open subset that includes C, by Montel's theorem, we automatically have

$$\sup_{z\in\mathrm{C}}|\omega_j'(z)-\hat{\omega}_j'(z)|\to 0$$

with probability one. Consequently, we end up with

$$f_j^{(l)}(\omega_j(z))\omega_j'(z)\,\mathbf{Q}_j(\omega_j(z)) - \hat{h}_j^{(l)}(z)\hat{\mathbf{Q}}_j(z) = \\ = \frac{\hat{h}_j^{(l)}(z)}{z}\left(\omega_j(z)\,\mathbf{Q}_j(\omega_j(z)) - z\hat{\mathbf{Q}}_j(z)\right) \\ + \left(\frac{f_j^{(l)}(\omega_j(z))\omega_j'(z)}{\omega_j(z)} - \frac{f_j^{(l)}(\hat{\omega}_j(z))\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\right)\omega_j(z)\,\mathbf{Q}_j(\omega_j(z)). \quad \text{(B.1)}$$

In what follows, we show that the terms on the right hand side of this equation are asymptotically equivalent to zero which is sufficient to conclude the proof.

The first term is asymptotically equivalent to zero because $\omega_j(z)\,\mathbf{Q}_j(\omega_j(z)) - z\hat{\mathbf{Q}}_j(z) \asymp 0$ and

$$\sup_{z\in\mathrm{C}}\left|\hat{h}_j^{(l)}(z)\right| = \sup_{z\in\mathrm{C}}\left|f_j^{(l)}(\hat{\omega}_j(z))\right|\sup_{z\in\mathrm{C}}\left|\frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\right| < \infty \quad \text{(B.2)}$$

with probability one for all large M. Indeed, the second term on the right hand side of (B.2) is finite (see Lemma D.1), and the first one is also bounded because $\hat{\omega}_j(z)$ belongs to a compact interval inside the analycity region of $F_j^{(l)}(\omega)$ with probability one.

Now, the second term on the right hand side of (B.1) can be studied by noting that $\sup_{z\in\mathrm{C}}\sup_M\|\omega_j(z)\,\mathbf{Q}_j(\omega_j(z))\| < +\infty$ and

$$\left|\frac{f_j^{(l)}(\omega_j(z))\omega_j'(z)}{\omega_j(z)} - \frac{f_j^{(l)}(\hat{\omega}_j(z))\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\right| \leq \\ \leq \left|f_j^{(l)}(\hat{\omega}_j(z))\right|\left|\frac{\omega_j'(z)}{\omega_j(z)} - \frac{\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\right| + \left|\frac{\omega_j'(z)}{\omega_j(z)}\right|\left|f_j^{(l)}(\omega_j(z)) - f_j^{(l)}(\hat{\omega}_j(z))\right|.$$

Then, reasoning as above, we immediately see that

$$\sup_M\sup_{z\in\mathrm{C}}|f_j^{(l)}(\hat{\omega}_j(z))| < \infty$$

and

$$\sup_{z\in\mathrm{C}}\left|\frac{\omega_j'(z)}{\omega_j(z)} - \frac{\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}\right| \to 0$$

with probability one. Regarding the second term, we have

$$\left|f_j^{(l)}(\omega_j(z)) - f_j^{(l)}(\hat{\omega}_j(z))\right| \leq |f_j^{(l)\prime}(\bar{\omega}_j(z))||\omega_j(z) - \hat{\omega}_j(z)|$$

where $f_j^{(l)\prime}(\omega)$ denotes the derivative of $f_j^{(l)}(\omega)$ and where $\bar{\omega}_j(z)$ belongs to the segment joining $\omega_j(z)$ and $\hat{\omega}_j(z)$. Clearly $\sup_M\sup_{z\in\mathrm{C}}|f_j^{(l)\prime}(\bar{\omega}_j(z))| < \infty$ as before, so that this term converges to zero uniformly in C.

Consider now the norm of $\hat{h}_j^{(l)}\left(\hat{\mathbf{R}}_j\right)$ and observe that

$$\left\|\hat{h}_j^{(l)}\left(\hat{\mathbf{R}}_j\right)\right\| = \sup_{\|\mathbf{u}\|=1} \mathbf{u}^H \hat{h}_j^{(l)}\left(\hat{\mathbf{R}}_j\right)\mathbf{u} \leq \sup_{\|\mathbf{u}\|=1} \frac{1}{2\pi}\oint_{\mathrm{C}^-} \left|\hat{h}_j^{(l)}(z)\right| \left|\mathbf{u}^H \hat{\mathbf{Q}}_j(z)\mathbf{u}\right| |dz| .$$

Now, obviously, $|\mathbf{u}^H\hat{\mathbf{Q}}_j(z)\mathbf{u}| \leq \mathrm{dist}^{-1}(z, \mathcal{S}_j \cup \{0\})$ and, as claimed in (B.2), we have $\sup_{z\in\mathrm{C}} |\hat{h}_j^{(l)}(z)| < \infty$ almost surely for all large M. It therefore follows that $\sup \|\hat{h}_j^{(l)}(\hat{\mathbf{R}}_j)\| < \infty$ with probability one for all large M.

B.2 Solving the integral in (3.10)

In order to simplify the notation, we drop the dependence on $j \in \{1, 2\}$ in all quantities (such as $\hat{\lambda}_m^{(j)}$, $\hat{\mathbf{e}}_k^{(j)}$, $\hat{\omega}_j(z)$, N_j, or $\alpha^{(j)}$) within this appendix. On the other hand, we will extensively use the fact that the eigenvalues $\hat{\lambda}_k$ are inside the contour C almost surely for all large M. Hence, all the associated statements should be understood to hold also with probability one and assuming that M is large enough (we will omit this detail throughout this appendix).

By using the expression for $\hat{\omega}(z)$ and $\hat{\omega}'(z)$ in (3.6) and (3.7) respectively, we can immediately see that we need to evaluate

$$\alpha = \frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}^-} \log^2\left(\frac{1-\hat{\Psi}(z)}{z}\right)\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right)\left(\frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{(\hat{\lambda}_m - z)^2}}{1-\hat{\Psi}(z)}\right) dz$$

where we have defined

$$\hat{\Psi}(z) = \frac{1}{N}\sum_{m=1}^{M}\frac{\hat{\lambda}_m}{\hat{\lambda}_m - z}.$$

To evaluate this integral, we first observe that $\log^2((1-\hat{\Psi}(z))/z) = \log^2(1-\hat{\Psi}(z)) - 2\log z \log(1-\hat{\Psi}(z)) + \log^2(z)$ and analyze the three integrals separately. The first integral (containing $\log^2(1-\hat{\Psi}(z))$) is the one that is simpler to evaluate, because $\log(1-\hat{\Psi}(z))$ is holomorphic everywhere except for the intervals $\cup_{k=1}^{M}[\hat{\mu}_k, \hat{\lambda}_k]$ where $\{\hat{\mu}_k, k = 1, \ldots, M\}$ are the solutions to the equation $\hat{\Psi}(\hat{\mu}) = 1$. Since these intervals are inside the contour C almost surely for all large M, one can conclude that the whole integrand is holomporphic outside C. One can therefore enlarge C and consider the change of variable $\zeta = z^{-1}$, after which the only potential singularity will be at $\zeta = 0$. It turns out that the resulting singularity at zero has residue equal to

zero, so that the corresponding integral is zero as well:

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log^2\left(1-\hat{\Psi}(z)\right)\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right)\left(\frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{\left(\hat{\lambda}_m - z\right)^2}}{1-\hat{\Psi}(z)}dz\right) = 0.$$

Now, the integral with respect to the term $\log^2(z)$ is also easily solved by evaluation of the residues at the singularities $\{\hat{\lambda}_k, \hat{\mu}_k\}$ for $k = 1, \ldots, M$, which are the only ones inside the contour C. It follows that

$$\begin{aligned}\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log^2 z\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right) \times \frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{\left(\hat{\lambda}_m - z\right)^2}}{1-\hat{\Psi}(z)}dz = \\ = \left(\frac{N}{M}-1\right)\sum_{r=1}^{M}\log^2\left(\hat{\mu}_r\right) + 2\frac{1}{M}\sum_{k=1}^{M}\log^2\left(\hat{\lambda}_k\right) \\ + 2\frac{1}{M}\sum_{k=1}^{M}\log\left(\hat{\lambda}_k\right) - \frac{1}{M}\sum_{k=1}^{M}\sum_{\substack{r=1\\r\neq k}}^{M}\log^2\left(\hat{\lambda}_r\right)\frac{\hat{\lambda}_r}{\hat{\lambda}_k - \hat{\lambda}_r} \\ - \frac{1}{M}\sum_{k=1}^{M}\log^2\left(\hat{\lambda}_k\right)\left(\sum_{m=1}^{M}\frac{\hat{\lambda}_m}{\hat{\lambda}_m - \hat{\mu}_k} - \sum_{\substack{m=1\\m\neq k}}^{M}\frac{\hat{\lambda}_m}{\hat{\lambda}_m - \hat{\lambda}_k}\right).\end{aligned}$$

It remains to compute the integral with respect to the cross term $\log z \log\left(1-\hat{\Psi}(z)\right)$. Let us denote

$$\mathcal{I} = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log z \log\left(1-\hat{\Psi}(z)\right)\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right)\left(\frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{\left(\hat{\lambda}_m - z\right)^2}}{1-\hat{\Psi}(z)}\right)dz.$$

We observe that $\mathcal{I} = \mathcal{I}(1)$ where we have defined the function $\mathcal{I}(x) : [0,1] \to \mathbb{C}$ as

$$\begin{aligned}\mathcal{I}(x) = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^-} \log z \log\left(1-x\hat{\Psi}(z)\right)\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right) \times \\ \times \frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{\left(\hat{\lambda}_m - z\right)^2}}{1-\hat{\Psi}(z)}dz. \quad \text{(B.3)}\end{aligned}$$

The above function is continously differentiable with respect to x, with derivative

$$\mathcal{I}'(x) = \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}^+} \log z\left(\frac{1}{M}\sum_{k=1}^{M}\frac{1}{\hat{\lambda}_k - z}\right)\hat{\Psi}(z)\left(\frac{1-\frac{M}{N}+\frac{1}{N}\sum_{m=1}^{M}\frac{z^2}{\left(\hat{\lambda}_m - z\right)^2}}{\left(1-x\hat{\Psi}(z)\right)\left(1-\hat{\Psi}(z)\right)}\right)dz.$$

This is a consequence of the fact that the integrand of the above function is uniformly bounded in C, so that by the dominated convergence theorem we can move the derivative with respect to x inside the integration. The above integral can easily be solved for $x \in (0,1)$ by noting that the only singularities of the integrand inside C are the sample eigenvalues $\hat{\lambda}_m$, the solutions to the equation $1 = \hat{\Psi}(\hat{\mu})$, namely $\hat{\mu}_m$, and the solutions to the equation $1 = x\hat{\Psi}(z)$, which will be denoted $\hat{\mu}_m(x)$, $m = 1, \ldots, M$. Using conventional residue calculus, we can solve for any $x \in (0,1)$, leading to

$$\begin{aligned}\mathcal{I}'(x) = \frac{1 - N/M}{1-x} \sum_{r=1}^{M} \log \hat{\mu}_r + \frac{1}{M} \sum_{r,k=1}^{M} \frac{\log \hat{\mu}_r(x)}{\hat{\lambda}_k - \hat{\mu}_r(x)} \hat{\mu}_r'(x) \\ + \frac{1}{(1-x)\,x} \left(\frac{N}{Mx} - 1 \right) \sum_{r=1}^{M} \log \hat{\mu}_r(x) \\ + \frac{1}{x} \frac{M+1}{M} \sum_{k=1}^{M} \log \hat{\lambda}_k + \frac{1}{x} - \left(\frac{1}{x} + 1 \right) \frac{1}{x} \frac{N}{M} \sum_{k=1}^{M} \log \hat{\lambda}_k\end{aligned}$$

where we have used the fact that $\hat{\mu}_m(x)$ are differentiable functions of x with probability one, with derivative

$$\hat{\mu}_k'(x) = \left(\frac{-x^2}{N} \sum_{m=1}^{M} \frac{\hat{\lambda}_m}{\left(\hat{\lambda}_m - \hat{\mu}_k(x) \right)^2} \right)^{-1}.$$

The above expression can be simplified by using the fact that (see [56,75])

$$1 - \frac{M}{N} = \frac{\prod_r \hat{\mu}_r}{\prod_r \hat{\lambda}_r} \qquad 1 - \frac{xM}{N} = \frac{\prod_r \hat{\mu}_r(x)}{\prod_r \hat{\lambda}_r} \tag{B.4}$$

so that the derivative $\mathcal{I}'(x)$ can alternatively be expressed as

$$\begin{aligned}\mathcal{I}'(x) = \frac{-1}{1-x} \left(\frac{N}{M} - 1 \right) \log \left(1 - \frac{M}{N} \right) \\ + \frac{1}{(1-x)\,x} \left(\frac{N}{Mx} - 1 \right) \log \left(1 - \frac{xM}{N} \right) \\ + \frac{1}{M} \sum_{k=1}^{M} \sum_{r=1}^{M} \frac{\log \hat{\mu}_r(x)}{\hat{\lambda}_k - \hat{\mu}_r(x)} \hat{\mu}_r'(x) + \frac{1}{x} \left(1 + \frac{1}{M} \sum_{k=1}^{M} \log \hat{\lambda}_k \right).\end{aligned} \tag{B.5}$$

From the above expression of the derivative $\mathcal{I}'(x)$ it is easy to find a primitive as follows. The primitive of the first and the fourth term are trivial, so let us first focus on the second term. In order to obtain a primitive of this term, we recall the

function $\Phi_2(x)$ introduced in (3.12). By using the change of variables $t = 1 - \frac{xM}{N}$ and partial fraction decomposition one can show that

$$\begin{aligned}\int \frac{1}{(1-x)\,x}\left(\frac{N}{Mx}-1\right)\log\left(1-\frac{xM}{N}\right)dx = \\ = -\frac{N}{xM}\log\left(1-\frac{xM}{N}\right)+\log\left(\frac{1-\frac{xM}{N}}{\frac{xM}{N}}\right) \\ +\left(\frac{N}{M}-1\right)\log\left(1-\frac{xM}{N}\right)\log\left(x\frac{1-\frac{M}{N}}{1-x}\right) \\ +\left(\frac{N}{M}-1\right)\left[\Phi_2\left(1-\frac{xM}{N}\right)-\Phi_2\left(\frac{1-\frac{xM}{N}}{1-\frac{M}{N}}\right)\right]+K\end{aligned}$$

where K is an undetermined constant (its value may change from one line to the next) and where we have used the fact that

$$\begin{aligned}\int \frac{\log t}{\lambda - t}dt &= \log t \log\left(\frac{\lambda}{|\lambda - t|}\right) - \Phi_2\left(\frac{t}{\lambda}\right) + \text{constant} \\ \int \frac{\log t}{(\lambda - t)^2}dt &= \frac{\log t}{|\lambda - t|} - \frac{1}{\lambda}\log\left(\frac{t}{|\lambda - t|}\right) + \text{constant}\end{aligned} \tag{B.6}$$

which can be readily proven by taking derivatives on both sides.

Regarding the third term of (B.5), one can easily show using (B.6) that

$$\begin{aligned}\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\int \frac{\log\hat{\mu}_r(x)}{\hat{\lambda}_k - \hat{\mu}_r(x)}\hat{\mu}_r'(x)\,dx = \frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\log\hat{\mu}_r(x)\log\frac{\hat{\lambda}_k}{\left|\hat{\lambda}_k - \hat{\mu}_r(x)\right|} \\ -\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\Phi_2\left(\frac{\hat{\mu}_r(x)}{\hat{\lambda}_k}\right)+K.\end{aligned}$$

Hence, putting everything together we can state that the primitive of $\mathcal{I}'(x)$ takes

the form

$$\begin{aligned}\mathcal{I}(x) = &\left(\frac{N}{M}-1\right)\log\left(1-\frac{M}{N}\right)\log(1-x)\\ &+\left(\frac{N}{M}-1\right)\log\left(1-\frac{xM}{N}\right)\log\left(\left(1-\frac{M}{N}\right)\frac{x}{1-x}\right)\\ &+\left(\frac{N}{M}-1\right)\left[\Phi_2\left(1-\frac{xM}{N}\right)-\Phi_2\left(\frac{1-\frac{xM}{N}}{1-\frac{M}{N}}\right)\right]\\ &-\frac{N}{xM}\log\left(1-\frac{xM}{N}\right)+\log\left(\frac{N}{M}\left(1-\frac{xM}{N}\right)\right)\\ &+\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\log\hat{\mu}_r(x)\log\frac{\hat{\lambda}_k}{\left|\hat{\lambda}_k-\hat{\mu}_r(x)\right|}\\ &+\left(\frac{1}{M}\sum_{k=1}^{M}\log\hat{\lambda}_k\right)\log x-\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\Phi_2\left(\frac{\hat{\mu}_r(x)}{\hat{\lambda}_k}\right)+K.\end{aligned}$$

The undetermined constant can be obtained by forcing $\mathcal{I}(0)=0$ (which follows from the definition of $\mathcal{I}(x)$ in (B.3)), leading to

$$\begin{aligned}K = &-\left(\frac{N}{M}-1\right)\left[\Phi_2(1)-\Phi_2\left(\frac{1}{1-\frac{M}{N}}\right)\right]-1-\log\left(\frac{N}{M}\right)+\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\Phi_2\left(\frac{\hat{\lambda}_r}{\hat{\lambda}_k}\right)\\ &-\frac{\log(N)}{M}\sum_{k=1}^{M}\log\hat{\lambda}_k-\frac{1}{M}\sum_{k=1}^{M}\sum_{\substack{r=1\\r\neq k}}^{M}\log\hat{\lambda}_r\log\left(\frac{\hat{\lambda}_k}{\left|\hat{\lambda}_k-\hat{\lambda}_r\right|}\right)\end{aligned}$$

where we have used the fact that $\hat{\mu}_k(x)\to\hat{\lambda}_k$ when $x\to 0$, and also

$$\lim_{x\to 0}\frac{x\hat{\lambda}_k}{\hat{\lambda}_k-\hat{\mu}_k(x)}=\lim_{x\to 0}\left(N-x\sum_{\substack{m=1\\m\neq k}}^{M}\frac{\hat{\lambda}_m}{\hat{\lambda}_m-\hat{\mu}_k(x)}\right)=N.$$

As a consequence of this, we can directly find

$$\begin{aligned}\mathcal{I}(1) = &\left(\frac{N}{M}-1\right)\log\left(1-\frac{M}{N}\right)\left[\frac{1}{2}\log\left(1-\frac{M}{N}\right)-1\right]-1\\ &+\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\log\hat{\mu}_r\log\left(\frac{\hat{\lambda}_k}{\hat{\lambda}_k-\hat{\mu}_r}\right)-\frac{1}{M}\sum_{k=1}^{M}\sum_{\substack{r=1\\r\neq k}}^{M}\log\hat{\lambda}_r\log\left(\frac{\hat{\lambda}_k}{\hat{\lambda}_k-\hat{\lambda}_r}\right)\\ &+\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\Phi_2\left(\frac{\hat{\lambda}_r}{\hat{\lambda}_k}\right)-\frac{1}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\Phi_2\left(\frac{\hat{\mu}_r}{\hat{\lambda}_k}\right)-\log(N)\frac{1}{M}\sum_{k=1}^{M}\log\hat{\lambda}_k\end{aligned}$$

where we have used the fact that $\mathrm{Li}_2(1) = \pi^2/6$ (so that $\Phi_2(1) = \pi^2/6$) together with the identity (for $z \in (0,1)$)

$$\Phi_2(z) + \Phi_2\left(z^{-1}\right) = \frac{\pi^2}{3} - \frac{1}{2}\log^2 z.$$

The final expression for α is obtained by putting together all the above integrals and using the fact that [76]

$$-\frac{1}{N}\hat{\lambda}_k = \frac{\prod_{r=1}^{M}\left(\hat{\mu}_r - \hat{\lambda}_k\right)}{\prod_{\substack{r=1\\r\neq k}}^{M}\left(\hat{\lambda}_r - \hat{\lambda}_k\right)}$$

and therefore

$$\log N = \sum_{r=1}^{M}\log\frac{\hat{\lambda}_k}{\left|\hat{\mu}_r - \hat{\lambda}_k\right|} - \sum_{\substack{r=1\\r\neq k}}^{M}\log\frac{\hat{\lambda}_k}{\left|\hat{\lambda}_r - \hat{\lambda}_k\right|}.$$

B.3 Proof of Theorem 3.1

The estimators $\hat{d}_M$ can all be expressed as

$$\hat{d}_M = \frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\hat{h}(z_1,z_2)\frac{1}{M}\mathrm{tr}\left[\hat{\mathbf{Q}}_1(z_1)\hat{\mathbf{Q}}_2(z_2)\right]dz_1dz_2$$

for some L, where

$$\hat{h}(z_1,z_2) = \sum_{l=1}^{L}\hat{h}_1^{(l)}(z_1)\hat{h}_2^{(l)}(z_2)$$
$$\hat{h}_j^{(l)}(z) = f_j^{(l)}(\hat{\omega}_j(z))\frac{z\hat{\omega}_j'(z)}{\hat{\omega}_j(z)}$$

for some $f_j^{(l)}(\omega)$ under consideration. Let us now denote as $h(z_1,z_2)$ the asymptotic equivalent of the random quantity $\hat{h}(z_1,z_2)$, that is

$$h(z_1,z_2) = \sum_{l=1}^{L}h_1^{(l)}(z_1)h_2^{(l)}(z_2) \tag{B.7}$$
$$h_j^{(l)}(z_1) = f_j^{(l)}(\omega_j(z_j))\frac{z_j\omega_j'(z_j)}{\omega_j(z_j)}$$

and note that, recalling the definition of $\bar{\mathbf{Q}}_j(z_j)$ and by Cauchy integration,

$$d_M = \frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}h(z_1,z_2)\frac{1}{M}\mathrm{tr}\left[\bar{\mathbf{Q}}_1(z_1)\bar{\mathbf{Q}}_2(z_2)\right]dz_1dz_2.$$

We begin by noticing that almost sure convergence in Theorem 2.1 above and the fact that all quantities inside the integral are bounded over the corresponding contours that we can write, using the short-hand notation $\bar{\mathbf{Q}}_j = \bar{\mathbf{Q}}_j(z_j)$ and $\hat{\mathbf{Q}}_j = \hat{\mathbf{Q}}_j(z_j)$,

$$M\left(\hat{d}_M - d_M\right) = \frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-} h(z_1,z_2)\hat{\xi}(z_1,z_2)dz_1dz_2 - \frac{1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-} \mathrm{tr}\left[\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_2\right]\left(\hat{h}(z_1,z_2) - h(z_1,z_2)\right)dz_1dz_2 + o_p(1) \quad \text{(B.8)}$$

where $o_p(1)$ denotes a random variable that converges in probability to zero and where we have defined

$$\hat{\xi}(z_1,z_2) = \mathrm{tr}\left[\hat{\mathbf{Q}}_1\hat{\mathbf{Q}}_2\right] - \mathrm{tr}\left[\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_2\right]. \quad \text{(B.9)}$$

Our first objective is to show that the second term on the right hand side of (B.8) can be expressed in a similar way as the first one. In a second step, we will apply a central limit theorem that was derived in [77] for this type of statistic. Let us focus on the first step of the proof, which is summarized in the proposition that is presented below.

In order to introduce this result, we need some additional definitions. For $j \in \{1,2\}$ and for a certain $M \times M$ deterministic matrix $\mathbf{A}$, consider the function $\phi_j(\omega;\mathbf{A})$, defined as

$$\phi_j(\omega;\mathbf{A}) = \frac{\omega}{1-\Gamma_j(\omega)}\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j^2(\omega)\mathbf{A}\right] \quad \text{(B.10)}$$

where $\Gamma_j(\omega)$ is defined in (3.17). With these definitions, we are now ready to present the first step in the proof, which is summarized in the following proposition.

Proposition B.1. *Under assumptions **(As1)-(As4)** we can write $M(\hat{d}_M - d_M) = \hat{\xi}_M + o_p(1)$, where*

$$\hat{\xi}_M = \frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-} h(z_1,z_2)\hat{\xi}(z_1,z_2)dz_1dz_2 + \frac{1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathcal{C}_2^-} g^{(1)}(z_1)\hat{\xi}(z_1,z_2)dz_1dz_2 + \frac{1}{4\pi^2}\oint_{\mathrm{C}_2^-}\oint_{\mathcal{C}_1^-} g^{(2)}(z_2)\hat{\xi}(z_1,z_2)dz_1dz_2 \quad \text{(B.11)}$$

where $\mathcal{C}_j^-$, $j \in \{1,2\}$, are negatively oriented simple contours enclosing $\mathcal{S}_j \cup \{0\}$ and where we have introduced the two functions

$$g^{(1)}(z_1) = \sum_{l=1}^{L} f_1^{(l)}(\omega_1)\phi_1(\omega_1; f_2^{(l)}(\mathbf{R}_2))$$
$$g^{(2)}(z_2) = \sum_{l=1}^{L} f_2^{(l)}(\omega_2)\phi_2(\omega_2; f_1^{(l)}(\mathbf{R}_1)).$$

Proof. It is sufficient to see that the second term on the right hand side of (B.8) coincides with the sum of the second and third terms in the statement of the proposition up to a sequence of random variables that converge to zero in probability. Observe that we can write

$$M(\hat{h}(z_1,z_2) - h(z_1,z_2)) = \sum_{l=1}^{L} \chi^{(l)}(z_1,z_2).$$

where we have defined

$$\chi^{(l)}(z_1,z_2) = M\left(\hat{h}_1^{(l)}(z_1)\hat{h}_2^{(l)}(z_2) - h_1^{(l)}(z_1)h_2^{(l)}(z_2)\right). \tag{B.12}$$

Using again Theorem 2.1 and the bounds in Lemmas D.1-D.2, we see that

$$\chi^{(l)}(z_1,z_2) = h_2^{(l)}(z_2)M\left(\hat{h}_1^{(l)}(z_1) - h_1^{(l)}(z_1)\right) + \\ + h_1^{(l)}(z_1)M\left(\hat{h}_2^{(l)}(z_2) - h_2^{(l)}(z_2)\right) + o_p(1)$$

where here and in the rest of this proof we should understand $o_p(1)$ as a function of z_1, z_2 which converges in probability to zero uniformly in $\mathrm{C}_1 \times \mathrm{C}_2$. Now, using a Taylor approximation of f around $\omega_j(z)$, we see that

$$N_j\left(f_j^{(l)}(\hat{\omega}_j) - f_j^{(l)}(\omega_j)\right) = f_j^{(l)\prime}(\omega_j)N_j(\hat{\omega}_j - \omega_j) + o_p(1)$$

where $f_j^{(l)\prime}(\omega_j)$ is the derivative of $f_j^{(l)}(\omega_j)$. Consequently, we see that

$$N_j\left(\hat{h}_j^{(l)}(z_j) - h_j^{(l)}(z_j)\right) = z_j\frac{f_j^{(l)}(\hat{\omega}_j)}{\omega_j}N_j\left(\hat{\omega}_j' - \omega_j'\right) \\ + z_j\left(f_j^{(l)\prime}(\omega_j) - \frac{f_j^{(l)}(\hat{\omega}_j)}{\omega_j}\right)\frac{\omega_j'}{\omega_j}N_j(\hat{\omega}_j - \omega_j) + o_p(1)$$

where we have used the short-hand notation $\omega_j' = \omega_j'(z_j)$ to denote the derivative of the function $\omega_j(z_j)$ (and equivalently for $\hat{\omega}_j'$). Now, by using the actual expressions for these two quantities, one can also express

$$N_j(\hat{\omega}_j - \omega_j) = N_j\hat{\omega}_j\omega_j\left(\frac{1}{\omega_j} - \frac{1}{\hat{\omega}_j}\right) = \hat{\omega}_j\omega_j\mathrm{tr}\left[\hat{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j\right] = \omega_j^2\mathrm{tr}\left[\hat{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j\right] + o_p(1)$$

which implies that

$$N_j\left(\hat{h}_j^{(l)}(z_j) - h_j^{(l)}(z_j)\right) = \varphi_j^{(l)}(z_j) + o_p(1)$$

where

$$\begin{aligned}\varphi_j^{(l)}(z_j) = z_j \frac{f_j^{(l)}(\omega_j)}{\omega_j}\frac{d}{dz_j}\left[\omega_j^2 \mathrm{tr}\left[\hat{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j\right]\right] \\ + z_j\left[\omega_j f_j^{(l)\prime}(\omega_j) - f_j^{(l)}(\omega_j)\right]\omega_j' \mathrm{tr}\left[\hat{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j\right].\end{aligned}$$

We can therefore express the second term on the right hand side of (B.8) as the sum for $l = 1, \ldots, L$ of the terms

$$\begin{aligned}\frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\frac{\mathrm{tr}\left[\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_2\right]}{M}\chi^{(l)}(z_1,z_2)dz_1dz_2 = \frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_1^-}\frac{1}{N_1}\mathrm{tr}\left[\bar{\mathbf{Q}}_1 f_2^{(l)}\left(\mathbf{R}_2\right)\right]\varphi_1^{(l)}(z_1)dz_1 \\ + \frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_2^-}\frac{1}{N_2}\mathrm{tr}\left[f_1^{(l)}\left(\mathbf{R}_1\right)\bar{\mathbf{Q}}_2\right]\varphi_2^{(l)}(z_2)dz_2 + o_p(1)\end{aligned}$$

where we have used the fact that (introducing the short hand notation $\mathbf{Q}_j = \mathbf{Q}_j(\omega_j(z_j))$)

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_j^-} h_j^{(l)}(z_j)\bar{\mathbf{Q}}_j dz_j = \frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_{\omega_j}^-} f_j^{(l)}(\omega_j)\mathbf{Q}_j d\omega_j = f_j^{(l)}\left(\mathbf{R}_j\right). \tag{B.13}$$

Using the integration by parts formula we can simplify the above expression to

$$\begin{aligned}\frac{-1}{4\pi^2}\oint_{\mathrm{C}_1^-}\oint_{\mathrm{C}_2^-}\frac{\mathrm{tr}\left[\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_2\right]}{M}\chi^{(l)}(z_1,z_2)dz_1dz_2 \\ = \frac{-1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_1^-}\phi_1(\omega_1; f_2^{(l)}(\mathbf{R}_2))f_1^{(l)}(\omega_1)\mathrm{tr}\left[\hat{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_1\right]dz_1 \\ - \frac{1}{2\pi\mathrm{j}}\oint_{\mathrm{C}_2^-}\phi_2(\omega_2; f_1^{(l)}(\mathbf{R}_1))f_2^{(l)}(\omega_2)\mathrm{tr}\left[\hat{\mathbf{Q}}_2 - \bar{\mathbf{Q}}_2\right]dz_2 + o_p(1)\end{aligned}$$

where we have used the fact that, for any deterministic square matrix $\mathbf{A}$, we can write

$$\begin{aligned}\frac{d}{dz_j}\left[\frac{\mathrm{tr}\left[\bar{\mathbf{Q}}_j\mathbf{A}\right]}{N_j}z_j\frac{f_j^{(l)}(\omega_j)}{\omega_j}\right] = \frac{d}{dz_j}\left[\frac{\mathrm{tr}\left[\mathbf{Q}_j\mathbf{A}\right]}{N_j}f_j^{(l)}(\omega_j)\right] \\ = \frac{1}{N_j}\mathrm{tr}\left[\mathbf{Q}_j^2\mathbf{A}\right]f_j^{(l)}(\omega_j)\omega_j' + \frac{1}{N_j}\mathrm{tr}\left[\mathbf{Q}_j\mathbf{A}\right]f_j^{(l)\prime}(\omega_j)\omega_j'.\end{aligned}$$

Now, let $\mathcal{C}_1^-$ and $\mathcal{C}_2^-$ denote two negatively oriented simple contours as in the statement of the proposition. We can now use the fact that

$$\frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_j^-}\hat{\mathbf{Q}}_j dz_j = \frac{1}{2\pi\mathrm{j}}\oint_{\mathcal{C}_j^-}\bar{\mathbf{Q}}_j dz_j = \mathbf{I}_M$$

which directly leads to

$$\mathrm{tr}[\hat{\mathbf{Q}}_1 - \bar{\mathbf{Q}}_1] = \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_2^-} \hat{\xi}(z_1, z_2) dz_2$$

(equivalently for $\mathrm{tr}[\hat{\mathbf{Q}}_2 - \bar{\mathbf{Q}}_2]$) and hence the statement of the proposition. □

It directly follows from the above proposition that the asymptotic law of $M(\hat{d}_M - d_M)$ coincides with that of $\hat{\xi}_M$ in (B.11). To study the asymptotic behavior of this new sequence of random variables, we use a CLT derived in [77], which can be directly applied to this new sequence of random variables. The CLT basically shows that random variables of the form ζ_M asymptotically fluctuate as Gaussian random variables with some predefined asymptotic (second order) mean and variance, which we introduce next.

The expression of $\hat{\xi}_M$ in (B.11) is complicated because the three integrals have to be taken with respect to different contours (i.e. $\mathcal{C}_j$ when the contour encloses $\{0\}$ and C_j when it does not). In order to simplify the presentation, we will admit a certain abuse of notation and rewrite $\hat{\xi}_M$ in (B.11) as

$$\hat{\xi}_M = \frac{-1}{4\pi^2} \oint_{C_1^-} \oint_{C_2^-} g(z_1, z_2) \hat{\xi}(z_1, z_2) dz_1 dz_2$$

where now

$$g(z_1, z_2) = h(z_1, z_2) - g^{(1)}(z_1) - g^{(2)}(z_2) \tag{B.14}$$

and where the general contour C_j should be understood to symbolize either $\mathcal{C}_j$ or C_j depending on the actual function that is integrated and following the convention in (B.11).

Following this simplified notation, we now define the asymptotic (second order) mean of $\hat{\xi}_M$ as

$$\mathfrak{m}_M = \frac{-\varsigma}{4\pi^2} \oint_{C_1^-} \oint_{C_2^-} \frac{\omega_1 \omega_2}{z_1 z_2} g(z_1, z_2) \mathfrak{m}(z_1, z_2) dz_1 dz_2 \tag{B.15}$$

where the function $\mathfrak{m}(z_1, z_2)$ is defined as follows. We let $\mathfrak{m}(z_1, z_2) = \mathfrak{m}_1(\omega_1, \mathbf{Q}_2(\omega_2)) + \mathfrak{m}_2(\omega_2, \mathbf{Q}_1(\omega_1))$, where $\mathfrak{m}_j(\omega_j, \mathbf{A})$ is defined (for a given $M \times M$ matrix $\mathbf{A}$) as

$$\mathfrak{m}_j(\omega, \mathbf{A}) = \frac{1}{N_j} \frac{\mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j^3(\omega) \Omega_j(\omega_j; \mathbf{A})\right]}{1 - \Gamma_j(\omega)} \tag{B.16}$$

with $\Omega_j(\omega; \mathbf{A})$ denoting

$$\Omega_j(\omega; \mathbf{A}) = \mathbf{A} + \phi_j(\omega; \mathbf{A}) \mathbf{I}_M \tag{B.17}$$

and where $\phi_j(\omega;\mathbf{A})$ is defined in (B.10).

In the same way, we define the asymptotic variance of $\hat{\xi}_M$ as (denoting again $\omega_j=\omega_j(z_j)$ and $\tilde{\omega}_j=\omega_j(\tilde{z}_j)$)

$$\sigma_M^2=\frac{1+\varsigma}{(2\pi\mathrm{j})^4}\oint_{C_1^-}\oint_{C_1^-}\oint_{C_2^-}\oint_{C_2^-} g(z_1,z_2)g(\tilde{z}_1,\tilde{z}_2)\times \\ \times\frac{\omega_1\omega_2\tilde{\omega}_1\tilde{\omega}_2}{z_1z_2\tilde{z}_1\tilde{z}_2}\Sigma^2\left(\omega_1,\omega_2,\tilde{\omega}_1,\tilde{\omega}_2\right)dz_1dz_2d\tilde{z}_1d\tilde{z}_2 \quad \text{(B.18)}$$

where $g(z_1,z_2)$ is defined in (B.14) and where the function $\Sigma^2\left(\omega_1,\omega_2,\tilde{\omega}_1,\tilde{\omega}_2\right)$ consists of the sum of three terms, namely

$$\Sigma^2\left(\omega_1,\omega_2,\tilde{\omega}_1,\tilde{\omega}_2\right)=\sigma_1^2\left(\omega_1,\tilde{\omega}_1;\mathbf{Q}_2\left(\omega_2\right),\mathbf{Q}_2(\tilde{\omega}_2)\right) \\ +\sigma_2^2\left(\omega_2,\tilde{\omega}_2;\mathbf{Q}_1\left(\omega_1\right),\mathbf{Q}_1(\tilde{\omega}_1)\right)++\varrho\left(\omega_1,\tilde{\omega}_1,\omega_2,\tilde{\omega}_2\right).$$

The first two terms are given by

$$\sigma_j^2\left(\omega,\tilde{\omega};\mathbf{A},\mathbf{B}\right)= \\ =\frac{1}{1-\Gamma_j(\omega,\tilde{\omega})}\frac{1}{N_j}\times\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j\left(\omega\right)\mathbf{Q}_j\left(\tilde{\omega}\right)\Omega_j(\omega;\mathbf{A})\mathbf{R}_j\mathbf{Q}_j\left(\omega\right)\mathbf{Q}_j\left(\tilde{\omega}\right)\Omega_j(\tilde{\omega};\mathbf{B})\right] \\ +\frac{1}{\left(1-\Gamma_j(\omega,\tilde{\omega})\right)^2}\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j^2\left(\omega\right)\mathbf{Q}_j\left(\tilde{\omega}\right)\Omega_j(\omega;\mathbf{A})\right]\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j^2\mathbf{Q}_j\left(\omega\right)\mathbf{Q}_j^2\left(\tilde{\omega}\right)\Omega_j(\tilde{\omega};\mathbf{B})\right] \quad \text{(B.19)}$$

whereas $\varrho\left(\omega_1,\tilde{\omega}_1,\omega_2,\tilde{\omega}_2\right)$ is as defined in (3.21). Now, assuming that **(As1)-(As3)** together with **(As5)** hold and that the observations are Gaussian distributed. It was proven in [77] that if $\liminf_{M\to\infty}\sigma_M^2>0$ we have

$$\frac{\zeta_M-\mathfrak{m}_M}{\sigma_M}\to\mathcal{N}(0,1).$$

Hence, Theorem 3.1 follows directly from this result, provided that we prove that the expressions of the asymptotic (second order) mean $\mathfrak{m}_M$ in (B.15) and the asymptotic variance in σ_M^2 (3.18) coincide with the expressions that are provided in the statement of the theorem.

B.3.1 Simplification of $\mathfrak{m}_M$ in (B.15)

The idea here again is to first use the change of variables $z_j\mapsto\omega_j=\omega_j\left(z_j\right)$ and noting that the transformed contour $\mathcal{C}_{\omega_j}^-=\omega_j(\mathcal{C}_j^-)$ encloses all the singularities of the integrand in (B.15). Consequently, one can enlarge the resulting contour $\mathrm{C}_{\omega_j}^-$ so

that it encloses $\{0\}$ and then apply a second change of variables $\omega_j \longmapsto \zeta = \omega_j^{-1}$ which only presents a singularity at zero. Following this procedure, one can easily see that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_1^-} \frac{\omega_1}{z_1} \mathfrak{m}\left(z_1, z_2\right) dz_1 = \mathfrak{m}_2(\omega_2, \mathbf{I}_M)$$

and equivalently for the integral with respect z_1. Consequently, the expression in (B.15) can be written as

$$\begin{aligned}
\mathfrak{m}_M = & \frac{-\varsigma}{4\pi^2} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \frac{\omega_1}{z_1} h(z_1, z_2) \mathfrak{m}_1\left(\omega_1, \bar{\mathbf{Q}}_2(z_2)\right) dz_1 dz_2 \\
& - \frac{\varsigma}{4\pi^2} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \frac{\omega_2}{z_2} h(z_1, z_2) \mathfrak{m}_2\left(\omega_2, \bar{\mathbf{Q}}_1(z_1)\right) dz_1 dz_2 \\
& - \frac{\varsigma}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1}{z_1} g^{(1)}(z_1) \mathfrak{m}_1(\omega_1, \mathbf{I}_M) dz_1 \\
& - \frac{\varsigma}{2\pi \mathrm{j}} \oint_{\mathrm{C}_2^-} \frac{\omega_2}{z_2} g^{(2)}(z_2) \mathfrak{m}_2(\omega_2, \mathbf{I}_M) dz_2.
\end{aligned}$$

The integral of the first (resp. second) term with respect to z_2 (resp. z_1) can easily be solved by inserting the expression of $h(z_1, z_2)$ given in (B.7) and using (B.13). At this point, we introduce the identity

$$\Omega_j\left(\omega; \mathbf{A} - \frac{\omega_j}{z_j \omega_j'} \phi_j(\omega_j; \mathbf{A}) \mathbf{I}_M\right) = \mathbf{A} \tag{B.20}$$

where $\omega_j' = (1 - \Gamma_j(\omega_j))^{-1}$ is the derivative of $\omega_j(z_j)$ with respect to z_j. A direct application of this identity, which can be proven using conventional algebra, allows converting the above expression for the second order mean into the expression given in (3.16).

B.3.2 Simplification of σ_M^2 in (3.18)

Noting that some of the terms of $g(z_1, z_2)$ only depend on one of the variables, we see that in a number of terms of σ_M^2 one of the variables will be integrated out of $\Sigma^2\left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right)$. We begin by noticing that, applying first the change of variables $z_1 \mapsto \omega_1\left(z_1\right)$, enlarging the contour and applying a second change of variables $\omega_1 \mapsto \zeta\left(\omega_1\right) = \omega_1^{-1}$, we obtain

$$\begin{aligned}
\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_1^-} \frac{\omega_1}{z_1} \varrho\left(\omega_1, \tilde{\omega}_1, \omega_2, \tilde{\omega}_2\right) dz_1 &= \frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_{\omega_1}^-} \frac{\omega_1}{z_1} \varrho\left(\omega_1, \tilde{\omega}_1, \omega_2, \tilde{\omega}_2\right)\left(1 - \Gamma_1(\omega_1)\right) d\omega_1 \\
&= \frac{1}{2\pi \mathrm{j}} \oint_{C_0^-} \frac{\zeta^{-1}\left(1 - \Gamma_1(\zeta^{-1})\right)}{z_1\left(\zeta^{-1}\right)} \varrho\left(\zeta^{-1}, \tilde{\omega}_1, \omega_2, \tilde{\omega}_2\right) \frac{d\zeta}{\zeta^2}
\end{aligned}$$

where C_0^- is a simple closed clockwise oriented contour enclosing zero and no other singularity. The last integral can be shown to be zero since the singularity at zero is in fact removable. By exactly the same procedure we see that

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathcal{C}_1^-} \frac{\omega_1}{z_1} \sigma_1^2 \left(\omega_1, \tilde{\omega}_1; \mathbf{A}, \mathbf{B}\right) dz_1 = 0$$

and, using the fact that (since $\mu_0^{(1)} = 0$)

$$\frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1}{z_1} \mathbf{Q}_1 \left(\omega_1\right) dz_1 = \mathbf{I}_M$$

we can conclude that

$$\frac{1}{2\pi \mathrm{j}} \oint_{C_1^-} \frac{\omega_1}{z_1} \Sigma^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) dz_1 = \sigma_2^2 \left(\omega_2, \tilde{\omega}_2; \mathbf{I}_M, \mathbf{Q}_1(\tilde{\omega}_1)\right).$$

Proceeding in a similar way, one can also show that

$$\frac{-1}{4\pi^2} \oint_{C_1^-} \oint_{C_1^-} \frac{\omega_1 \tilde{\omega}_1}{z_1 \tilde{z}_1} \Sigma^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) dz_1 d\tilde{z}_1 = \sigma_2^2 \left(\omega_2, \tilde{\omega}_2; \mathbf{I}_M, \mathbf{I}_M\right)$$

whereas

$$\frac{-1}{4\pi^2} \oint_{C_1^-} \oint_{C_2^-} \frac{\omega_1}{z_1} \frac{\omega_2}{z_2} \Sigma^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) dz_1 dz_2 = \\ = \frac{-1}{4\pi^2} \oint_{C_1^-} \oint_{C_2^-} \frac{\omega_1 \tilde{\omega}_2}{z_1 \tilde{z}_2} \Sigma^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) dz_1 d\tilde{z}_2 = 0.$$

Using all these identities, together with the fact that one can express

$$g^{(1)}(z_1) = \frac{\omega_1}{z_1 \omega_1'} \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_2^-} \frac{\omega_2}{z_2} h(z_1, z_2) \phi_1(\omega_1; \mathbf{Q}_2 \left(\omega_2\right)) dz_2$$
$$g^{(2)}(z_2) = \frac{\omega_2}{z_2 \omega_2'} \frac{1}{2\pi \mathrm{j}} \oint_{\mathrm{C}_1^-} \frac{\omega_1}{z_1} h(z_1, z_2) \phi_2(\omega_2; \mathbf{Q}_1 \left(\omega_1\right)) dz_1$$

we see that we can express the asymptotic variance as

$$\sigma_M^2 = \frac{1+\varsigma}{(2\pi \mathrm{j})^4} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_1^-} \oint_{\mathrm{C}_2^-} \oint_{\mathrm{C}_2^-} \frac{\omega_1}{z_1} \frac{\omega_2}{z_2} \frac{\tilde{\omega}_1}{\tilde{z}_1} \frac{\tilde{\omega}_2}{\tilde{z}_2} \times \\ \times h(z_1, z_2) h(\tilde{z}_1, \tilde{z}_2) \overline{\Sigma}^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) dz_1 dz_2 d\tilde{z}_1 d\tilde{z}_2$$

which is similar to the original expression, save for the fact that $g(z_1, z_2)$ in (3.18) is replaced by the simpler function $h(z_1, z_2)$ and $\Sigma^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right)$ is also replaced by the function

$$\overline{\Sigma}^2 \left(\omega_1, \omega_2, \tilde{\omega}_1, \tilde{\omega}_2\right) = \overline{\sigma}_1^2 \left(\omega_1, \tilde{\omega}_1; \mathbf{Q}_2 \left(\omega_2\right), \mathbf{Q}_2(\tilde{\omega}_2)\right) + \\ + \overline{\sigma}_2^2 \left(\omega_2, \tilde{\omega}_2; \mathbf{Q}_1 \left(\omega_1\right), \mathbf{Q}_1(\tilde{\omega}_1)\right) + \varrho \left(\omega_1, \tilde{\omega}_1, \omega_2, \tilde{\omega}_2\right)$$

where $\overline{\sigma}_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A}, \mathbf{B}\right)$ takes the form

$$\overline{\sigma}_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A}, \mathbf{B}\right) = \sigma_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A} - \frac{\omega_j \phi_j(\omega_j; \mathbf{A})}{z_j \omega_j'}\mathbf{I}_M, \mathbf{B} - \frac{\tilde{\omega}_j \phi_j(\tilde{\omega}_j; \mathbf{B})}{\tilde{z}_j \tilde{\omega}_j'}\mathbf{I}_M\right).$$

Now, using the identity in (B.20) one trivially finds that $\overline{\sigma}_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A}, \mathbf{B}\right)$ can also be expressed as in (B.19), but replacing $\Omega_j(\mathbf{A})$ and $\Omega_j(\mathbf{B})$ with $\mathbf{A}$ and $\mathbf{B}$ respectively. In order to obtain the above expression, one needs to use the interesting identity

$$\sigma_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A} - \frac{\omega_j \phi_j(\omega_j; \mathbf{A})}{z_j \omega_j'}\mathbf{I}_M, \mathbf{B} - \frac{\tilde{\omega}_j \phi_j(\tilde{\omega}_j; \mathbf{B})}{\tilde{z}_j \tilde{\omega}_j'}\mathbf{I}_M\right) = \tilde{\sigma}_j^2\left(\omega_j, \tilde{\omega}_j; \mathbf{A}, \mathbf{B}\right)$$

which follows directly from the identity in (B.20).

Appendix C

Appendix for Chapter 5

C.1 Estimation of mean

To derive an estimator for the first order moment $\bar{s}_{12}^{PF}$, for the subspace similarity $s_{12} = M^{-1}\text{tr}[\mathbf{P}_1\mathbf{P}_2]$, we first observe that, for two processes $\mathbf{Y}_1, \mathbf{Y}_2$ built from the same covariance $\mathbf{R}$, we have that

$$\mu_1\mathcal{Q}_1 - \mu_2\mathcal{Q}_2 = \mathcal{Q}_2(\mu_2 - \mathcal{Q}_2^{-1}\mu_2\mathcal{Q}_1) = \mathbf{R}(\mu_2 - \mu_1)\mathcal{Q}_2\mathcal{Q}_1,$$

where we have defined the short notation $\mathcal{Q}_j = \mathbf{Q}_j\left(\mu_0^{(j)}\right)$, for $j = 1, 2$. The above allows us to re-write $\mathcal{Q}_1 - \mathcal{Q}_2 = (\mu_2 - \mu_1)\mathcal{Q}_1\mathcal{Q}_2$. This is particularly useful for the case where $N_1 > N_2$, but when $N_1 = N_2$, we have $\mu_1 = \mu_2$, and the above brings no further intuition. Therefore, we need to distinguish between the two cases $N_1 > N_2$ and $N_1 = N_2$.

C.1.1 Case $N_1 > N_2$

We start by the simpler case and observe that, for $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$, we can re-write

$$\begin{aligned}\bar{s}_{12}^{PF} &= \frac{1}{M}\text{tr}\left[\mathbf{R}_1\mathcal{Q}_1\mathbf{R}_2\mathcal{Q}_2\right] \\ &= \frac{1}{\mu_1 - \mu_2}\left(\mu_1\frac{1}{M}\text{tr}\left[\mathbf{R}\mathcal{Q}_1\right] - \mu_2\frac{1}{M}\text{tr}\left[\mathbf{R}\mathcal{Q}_2\right]\right) \\ &= \frac{N_1\mu_1 - N_2\mu_2}{M}\end{aligned} \tag{C.1}$$

so that we only need to find an estimate for $\mu_j, j \in \{1, 2\}$.

Consistent estimator under null and alternative hypothesis

Consider the identity

$$\frac{1}{N_j}\text{tr}\left[\hat{\mathbf{R}}_j\left(\hat{\mathbf{R}}_j - z\mathbf{I}_M\right)^{-1}\right] \asymp \frac{1}{N_j}\text{tr}\left[\mathbf{R}_j\left(\mathbf{R}_j - z\mathbf{I}_M\right)^{-1}\right] = 1 - \frac{z_j}{\omega_j(z_j)}$$

we then recover (2.1), i.e.,

$$z = \omega_j\left(z\right)\left(1 - \frac{1}{N_j}\sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\gamma_m^{(j)}}{\gamma_m^{(j)} - \omega_j\left(z\right)}\right)$$

and because $\mu_0^{(j)} = \omega_j(0)$, we obtain that

$$\mu_0^{(j)} \asymp -\left(\frac{1}{N_j}\sum_{m=M-N_j-+1}^{M} \lambda_j\right)^{-1} = -N_j\left(\text{tr}\left[\hat{\mathbf{R}}_j^{\#}\right]\right)^{-1}. \tag{C.2}$$

Plugging the above into (C.1), we obtain, for $N_1 > N_2$,

$$\frac{1}{M}\frac{N_1^2\text{tr}\left[\hat{\mathbf{R}}_2^{\#}\right] - N_2^2\text{tr}\left[\hat{\mathbf{R}}_1^{\#}\right]}{N_1\text{tr}\left[\hat{\mathbf{R}}_2^{\#}\right] - N_2\text{tr}\left[\hat{\mathbf{R}}_1^{\#}\right]}.$$

The problem with the above estimate is that it is consistent even under the alternative hypothesis which is undesired in our scenario as we want an estimator consistent under the null hypothesis but that penalizes the alternative one.

Consistent estimator only under null hypothesis

Let us consider the estimator of both $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_1$ assuming we only have access to $\hat{\mathbf{R}}_1$, which is built using $N_1 > N_2$ samples. This means that we have

$$\omega_1\left(z_1\right) \asymp z_1\left(1 - \frac{1}{N_1}\text{tr}\left[\hat{\mathbf{R}}_1\left(\hat{\mathbf{R}}_1 - z_1\mathbf{I}_M\right)^{-1}\right]\right)^{-1}.$$

Now, by definition (2.3) $\mu_0^{(2)}$ is the only negative solution to

$$\frac{1}{N_2}\text{tr}\left[\mathbf{R}\left(\mathbf{R} - \mu_0^{(2)}\mathbf{I}_M\right)^{-1}\right] = 1$$

and we have that

$$\frac{1}{N_2}\text{tr}\left[\mathbf{R}\left(\mathbf{R} - \omega_1(z_1)\mathbf{I}_M\right)^{-1}\right] \asymp \frac{1}{N_2}\text{tr}\left[\hat{\mathbf{R}}_1\left(\hat{\mathbf{R}}_1 - z_1\mathbf{I}_M\right)^{-1}\right].$$

Therefore, we can estimate $\mu_0^{(2)}$ by first finding the negative solution to the equation

$$\frac{1}{N_2}\text{tr}\left[\hat{\mathbf{R}}_1\left(\hat{\mathbf{R}}_1 - \gamma\mathbf{I}_M\right)\right] = 1$$

and then taking

$$\hat{v}_2(1) = \gamma\left(1 - \frac{N2}{N1}\right).$$

Plugging the above back into (C.1) we obtain (5.15). This distinction is particularly useful to penalize the alternative hypothesis based on $\hat{v}_2$ and $\hat{\mathbf{R}}_2^{\#}$ by relating the smaller sample eigenvalue distribution in terms of the larger one.

C.1.2 Case $N_1 = N_2$

For the case where $N_1 = N_2 = N$, we also have that $\mathcal{Q}_1 = \mathcal{Q}_2$. Hence, the observation in the beginning of this appendix brings no further intuition. Therefore, we rely on the fact that, for $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$, we have

$$\bar{s}_{12}^{PF} = \frac{1}{M}\text{tr}\left[\mathbf{R}^2\mathcal{Q}^2\right] = \frac{N}{M}\left(1 - \frac{1}{\omega'(0)}\right)$$

where $\omega_1'(0) = \omega_2'(0) = \omega'(0)$. Hence, by taking any of the derivatives $\omega_j'(0)$, for $j \in \{1, 2\}$ we obtain

$$\bar{s}_{kj}^{PF} \asymp \frac{N_j}{M}\left(1 - \frac{\left(N_j^{-1}\text{tr}\left[\hat{\mathbf{R}}_j^{\#}\right]\right)^2}{N_j^{-1}\text{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^2\right]}\right).$$

Finally, by splitting the contributions of both the observations ($\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$), we obtain with our estimator

$$\begin{aligned}\bar{s}_{12}^{PF} \asymp \hat{\bar{s}}_{12}^{PF} &= \frac{1}{2M}\left(\frac{N_1\hat{v}_1(k) - N_2\hat{v}_2(k)}{\hat{v}_1(k) - \hat{v}_2(k)} + \frac{N_1\hat{v}_1(2) - N_2\hat{v}_2(2)}{\hat{v}_1(2) - \hat{v}_2(2)}\right) \\ &= \frac{N_1}{2M}\left(1 - \frac{\hat{\kappa}_1^2(1)}{\hat{\kappa}_1(2)}\right) + \frac{N_2}{2M}\left(1 - \frac{\hat{\kappa}_2^2(1)}{\hat{\kappa}_2(2)}\right).\end{aligned}$$

C.2 Estimation of variance of Subspace Similarity

The estimator of the asymptotic variance $(\sigma_{kj}^{PF})^2$ in (5.10) is denoted by $(\hat{\sigma}_{kj}^{PF})^2$ and can be described as

$$\left(\hat{\sigma}_{kj}^{PF}\right)^2 = \zeta_k^2(j) + \zeta_j^2(k) + \frac{\left(\text{tr}\left[\hat{\mathbf{R}}_k^{\#}\hat{\mathbf{R}}_j^{\#}\right]\right)^2}{\text{tr}\left[\left(\hat{\mathbf{R}}_k^{\#}\right)^2\right]\text{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^2\right]} \tag{C.3}$$

where we have defined

$$\begin{aligned}\zeta_j^2(k) = &\left[\frac{\hat\kappa_j(4)\hat\kappa_j^2(1)}{\hat\kappa_j^3(2)} - 2\left(\frac{\hat\kappa_j(1)\hat\kappa_j(3)}{\hat\kappa_j^2(2)}\right)^2\right]\times \\ &\left[\frac{1}{\hat\kappa_j(1)}\rho_j(\mathbf{B}_k,1) - \left(\frac{1}{\hat\kappa_j(1)}\rho_j(\mathbf{A}_k,1)\right)^2\right] \\ &+ \frac{\hat\kappa_j(1)\hat\kappa_j(3)}{\hat\kappa_j^2(2)}\left(\frac{\rho_j(\mathbf{B}_k,1)}{\hat\kappa_j(1)} + \frac{2\rho_j(\mathbf{A}_k,2)}{\hat\kappa_j(2)} - \frac{\rho_j(\mathbf{B}_k,3)}{\hat\kappa_j(3)}\right) \\ &- 2\frac{\hat\kappa_j(3)\rho_j(\mathbf{A}_k,1)}{\hat\kappa_j^2(2)}\left(\frac{2\rho_j(\mathbf{A}_k,2)}{\hat\kappa_2(2)} - \frac{\rho_j(\mathbf{A}_k,3)}{\hat\kappa_j(3)}\right) \\ &+ \left(\frac{\rho_j(\mathbf{A}_k,2)}{\hat\kappa_j(2)}\right)^2 - \frac{\rho_j(\mathbf{B}_k,2)}{\hat\kappa_j(2)}\end{aligned} \tag{C.4}$$

and have also introduced the matrices $\mathbf{A}_k = -\mathbf{P}_k^\perp$ and $\mathbf{B}_k = \mathbf{P}_k^\perp - \frac{\hat\kappa_k(1)}{\hat\kappa_k(2)}\hat{\mathbf{R}}_k^\#$, together with $\mathbf{P}_k^\perp = \mathbf{I}_M - \mathbf{P}_k$ and

$$\rho_j(\mathbf{A},n) = \frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\left(\hat{\mathbf{R}}_j^\#\right)^n\right]$$

valid for any squared matrix $\mathbf{A}$. Finally, $\zeta_k^2(j)$ is defined in the same way, but swapping the two indexes k, j. In the next section, we further detail the derivation of this estimator.

C.2.1 Detailed Estimation

Let us start by observing that, for $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$, the variance (5.10) takes the form

$$\begin{aligned}\left(\sigma_{12}^{PF}\right)^2 = &\left(\mu_0^{(1)}\right)^2\sigma_1^2\left(\mu_0^{(1)},\mu_0^{(1)};\mathbf{R}\mathbf{Q}_2\left(\mu_0^{(2)}\right),\mathbf{R}\mathbf{Q}_2\left(\mu_0^{(2)}\right)\right) \\ &+ \left(\mu_0^{(2)}\right)^2\sigma_2^2\left(\mu_0^{(2)},\mu_0^{(2)};\mathbf{R}\mathbf{Q}_1\left(\mu_0^{(1)}\right),\mathbf{R}\mathbf{Q}_1\left(\mu_0^{(1)}\right)\right) \\ &+ \left(\mu_0^{(1)}\mu_0^{(2)}\right)^2\frac{\mathrm{tr}^2\left[\mathbf{R}^2\mathbf{Q}_1^2(\mu_0^{(1)})\mathbf{Q}_2^2(\mu_0^{(2)})\right]}{N_1N_2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}\end{aligned} \tag{C.5}$$

and that the second order moment is a combination of two elements $\mu_0^{(j)}\sigma_j^2 = \sigma_j^2(\mu_0^{(j)},\mu_0^{(j)};\mathbf{A},\mathbf{A}), j = 1,2$ and

$$\left(\mu_0^{(1)}\mu_0^{(2)}\right)^2\frac{\mathrm{tr}^2\left[\mathbf{R}^2\mathbf{Q}_1^2(\mu_0^{(1)})\mathbf{Q}_2^2(\mu_0^{(2)})\right]}{N_1N_2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}. \tag{C.6}$$

Let us first focus on the estimation of $\sigma_j^2 = \sigma_j^2(\mu_0^{(j)}, \mu_0^{(j)}; \mathbf{A}, \mathbf{A}), j = 1, 2$. By plugging in the definition of Ω_j we obtain

$$\sigma_j^2(\mathbf{A}) = \frac{1}{1-\Gamma_j}\frac{1}{N_j}\mathrm{tr}\left[\mathbf{R}_j\mathbf{Q}_j^2\left(\mathbf{A}+\phi(\omega;\mathbf{A})\mathbf{I}_M\right)\mathbf{R}_j\mathbf{Q}_j^2\left(\mathbf{A}+\phi(\omega;\mathbf{A})\mathbf{I}_M\right)\right] + \\ + \frac{1}{\left(1-\Gamma(\omega_j)\right)^2}\frac{1}{N_j}\mathrm{tr}^2\left[\mathbf{R}_j^2\mathbf{Q}_j^3(\omega_j)\left(\mathbf{A}+\phi(\omega;\mathbf{A})\mathbf{I}_M\right)\right] \tag{C.7}$$

and by defining, for $m \in \mathbb{N}$,

$$\kappa_j(m) = \frac{1}{N}\mathrm{tr}\left[\mathbf{R}_j^2\mathcal{Q}_k^m\right],$$

we can further expand the above into

$$\sigma_j^2(\mathbf{A}) = \left(\frac{\kappa_j^2(3)}{1-\kappa_j^2(2)} + \frac{\kappa_j(4)}{1-\kappa_j^2(2)}\right)\left(\frac{\mu_j}{1-\kappa_j^2(2)}\frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_j\mathcal{Q}_j^2\right]\right)^2 \\ + 2\frac{\mu_j}{1-\kappa_j^2(2)}\frac{1}{N_j}\mathrm{tr}[\mathbf{A}\mathbf{R}_j\mathcal{Q}_j^2]\left(\frac{\kappa_j(3)}{1-\kappa_j^2(2)}\frac{1}{N_j}\mathrm{tr}[\mathbf{A}\mathbf{R}_j^2\mathcal{Q}_j^3] + \frac{1}{N_j}\mathrm{tr}[\mathbf{A}\mathbf{R}_j^2\mathcal{Q}_j^4]\right) \\ + \frac{1}{1-\kappa_j^2(2)}\left(\frac{1}{N_j}\mathrm{tr}[\mathbf{A}\mathbf{R}_j^2\mathcal{Q}_j^3]\right)^2 + \frac{\mu_j^2}{1-\kappa_j^2(2)}\frac{1}{N_j}\mathrm{tr}\left[\left(\mathbf{A}\mathbf{R}_j\mathcal{Q}_j^2\right)^2\right]. \tag{C.8}$$

In what follows, we will derive estimators for each one of these quantities.

To do so, we first recall that

$$\frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_j\left(\mathbf{R}_j-\omega_j(z_j)\mathbf{I}_M\right)^{-1}\right] \asymp \frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{R}}_j\left(\hat{\mathbf{R}}_j - z_j\mathbf{I}_M\right)^{-1}\right]. \tag{C.9}$$

and then notice that its sequential derivatives (up to the fifth order) with respect to z_j are closely related to the quantities of the type $N_j^{-1}\mathrm{tr}[\mathbf{A}\mathbf{R}_j^n\mathcal{Q}_j^m]$, for $n, m \in \mathbb{N}$. For instance, by taking the first order derivative of (C.9), we directly obtain

$$\frac{1}{N_j}\frac{1}{1-\Gamma_j(\omega_j)}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_j\left(\mathbf{R}_j-\omega_j\mathbf{I}_M\right)^{-2}\right] \asymp \frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{R}}_j\left(\hat{\mathbf{R}}_j - z_j\mathbf{I}_M\right)^{-2}\right] \tag{C.10}$$

and, by taking the limit as $z_j \to 0$, we see that

$$\frac{1}{N_j}\frac{1}{1-\Gamma_j(\omega_j)}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_j\left(\mathbf{R}_j-\mu_j\mathbf{I}_M\right)^{-2}\right] \asymp \frac{1}{N_j}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{R}}_j^{\#}\right].$$

Similarly, the second order derivative (together with $z_j \to 0$) leads to

$$\frac{1}{N_\ell}\frac{1}{\left(1-\theta_\ell(2)\right)^2}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathcal{Q}_\ell^3\right] + \frac{1}{N_\ell}\frac{\theta_\ell(3)}{\left(1-\theta_\ell(2)\right)^3}\mathrm{tr}\left[\mathbf{A}\mathbf{R}_\ell\mathcal{Q}_\ell^2\right] \asymp \frac{1}{N_\ell}\mathrm{tr}[\mathbf{A}(\hat{\mathbf{R}}_\ell^{\#})^2] \tag{C.11}$$

Moreover, for the particular case where $\mathbf{A} = \mathbf{I}$, we have the estimators

$$\hat{\kappa}_j(1) = \frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_j^{\#}\right] \asymp -\frac{1}{\mu_j} \tag{C.12}$$

$$\hat{\kappa}_j(2) = \frac{1}{N_j}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^2\right] \asymp \frac{1}{\mu_j^2}\frac{1}{1-\kappa_j(2)} \tag{C.13}$$

$$\hat{\kappa}_j(3) = \frac{1}{N_j}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^3\right] \asymp \frac{1}{\mu_j^2}\frac{\kappa_j(3)}{(1-\kappa_j(2))^3} - \frac{1}{\mu_j^3}\frac{\kappa_j(3)}{(1-\kappa_j(2))^2} \tag{C.14}$$

$$\hat{\kappa}_j(4) = \frac{1}{N_j}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^4\right] \asymp \frac{1}{\mu_j^4}\frac{1}{(1-\kappa_j(2))^3} + \frac{1}{\mu_j^2}\frac{\kappa_j(4)}{(1-\kappa_j(2))^4} \tag{C.15}$$

$$- 2\frac{1}{\mu^3}\frac{\kappa_j(3)}{(1-\kappa_j(2))^4} + 2\frac{1}{\mu_j^2}\frac{\kappa_j^2(3)}{(1-\kappa_j(2))^5} \tag{C.16}$$

where $\hat{\kappa}_\ell(3)$ and $\hat{\kappa}_\ell(4)$ are obtained as above using the third and fourth order derivatives of (C.9), respectively. Then, by manipulating these quantities, one can also obtain estimators for the entries of $\kappa_j(m), m \leq 4$. With this we have an estimator for almost all the quantities of $\sigma_j^2(\mathbf{A})$, except for the terms that contain the deterministic matrix $\mathbf{A}$ inside a trace which can be obtained by plugging in (C.12)-(C.16) into the sequential (first to third order) derivatives of (C.10) for some deterministic matrix $\mathbf{A}$.

At this point, we need to find the asymptotic equivalent for $(\mathbf{R}_k\mathcal{Q}_k)^2, k \neq j$ which will replace the matrix $\mathbf{A}^2$ in the last term of (C.8). Notice that this quantity cannot be directly estimated from the above estimators. Hence, to provide consistent estimators that can be used in $\sigma_j^2(\mathbf{A})$, we first observe that, for any $\alpha \in \mathbb{R}$ we have $\sigma_j^2(\mathbf{A}+\alpha\mathbf{I}) = \sigma_j^2(\mathbf{A})$ and $\sigma_j^2(\alpha\mathbf{A}) = \alpha^2\sigma_j^2(\mathbf{A})$, so that, by considering $\mathbf{R}_k\,\mathcal{Q}_k = \mathbf{I}_M + \mu_k\mathcal{Q}_k$, we end up with

$$\sigma_j^2(\mathbf{R}_k\mathcal{Q}_k) = \sigma_j^2(\mathbf{I}_M + \mu_k\mathcal{Q}_k) = \mu_k^2\sigma_j^2(\mathcal{Q}_k). \tag{C.17}$$

Therefore, it is sufficient to estimate for $\mathbf{A} = \mu_k\mathcal{Q}_k$ and $\mathbf{A}^2 = \mu_k^2\mathcal{Q}_k^2$. To that effect, we consider the identity

$$\frac{1}{N_k}\mathrm{tr}[\mathbf{A}(\hat{\mathbf{R}}_k - z_k\mathbf{I}_M)^{-1}] \asymp \frac{\omega_k(z_k)}{z_k}\frac{\mathrm{tr}[\mathbf{A}\mathcal{Q}_k\,(\omega_k(z_k))]}{N_k} \tag{C.18}$$

valid for any deterministic $\mathbf{B}$. Again, by letting $z_k \to 0$ we have

$$\frac{1}{N_k}\mathrm{tr}\left[\mathbf{B}\hat{\mathbf{P}}_k^{\perp}\right] \asymp -\mu_k\frac{1}{N_k}\mathrm{tr}\left[\mathbf{B}\mathcal{Q}_k\right] \tag{C.19}$$

or equivalently, we can express $\mathbf{A} = \mu_k \mathcal{Q}_k \asymp -\hat{\mathbf{P}}_k^{\perp}$. Similarly, taking the derivative of (C.18) and forcing again $z_k \to 0$ leads to

$$\frac{1}{N_k}\text{tr}\left[\mathbf{B}\hat{\mathbf{R}}_k^{\#}\right] \asymp \frac{1}{N_k}\frac{\text{tr}\left[\mathbf{B}\mathbf{R}_k\mathcal{Q}_k^2\right]}{1-\theta_k(2)} \tag{C.20}$$

or, after some algebra,

$$\frac{1}{N_k}\text{tr}\left[\mathbf{B}\mu_k^2\mathcal{Q}_k^2\right] \asymp \frac{1}{N_k}\text{tr}\left[\mathbf{B}\left(\hat{\mathbf{P}}_k^{\perp} - \frac{\hat{\kappa}(1)}{\hat{\kappa}(2)}\hat{\mathbf{R}}_k^{\#}\right)\right] \tag{C.21}$$

or, equivalently,

$$\mu_k\mathcal{Q}_k \asymp -\hat{\mathbf{P}}_k^{\perp} \tag{C.22}$$

$$\mu_k^2\mathcal{Q}_k^2 \asymp \hat{\mathbf{P}}_k^{\perp} - \frac{\hat{\kappa}_k(1)}{\hat{\kappa}_k(2)}\hat{\mathbf{R}}_k^{\#}. \tag{C.23}$$

Finally, the estimator for (C.6) is the direct combination of the elements already presented above, so that by rearranging the terms we have

$$\left(\mu_0^{(1)}\mu_0^{(2)}\right)^2 \frac{1}{N_1N_2}\frac{\text{tr}^2\left[\mathbf{R}^2\mathbf{Q}_1^2(\mu_0^{(1)})\mathbf{Q}_2^2(\mu_0^{(2)})\right]}{\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)} \asymp \frac{1}{N_1N_2}\frac{\text{tr}^2\left[\hat{\mathbf{R}}_2^{\#}\hat{\mathbf{R}}_1^{\#}\right]}{\hat{\kappa}_2(2)\hat{\kappa}_1(2)} \tag{C.24}$$

where for the trace we have applied (C.10) twice, first considering $\mathbf{A} = \mathbf{R}_j\mathcal{Q}_j^2$ and then $\mathbf{A} = \mathbf{R}_j\mathcal{Q}_j^2$.

Appendix D

Auxiliary Lemmas

D.1 Some useful lemmas

Lemma D.1. *Let $j \in \{0,1\}$. Under **(As1)-(As3)** we have*

$$\begin{gathered}\sup_{M}\sup_{z\in C_j}\|\mathbf{Q}_j(\omega_j(z))\|<+\infty\\ 0<\inf_{M}\inf_{z\in C_j}|\omega_j(z)|\leq\sup_{M}\sup_{z\in C_j}|\omega_j(z)|<+\infty\\ \inf_{M}\operatorname{dist}\left\{\omega_j(z),(-\infty,\mu_{\text{inf}}^{(j)}]\right\}>0\end{gathered}\tag{D.1}$$

where $\mu_{\text{inf}}^{(j)}=\inf_M \mu_0^{(j)}$. Furthermore,

$$\begin{gathered}\sup_{M\geq M_0}\sup_{z\in C_j}\left\|\hat{\mathbf{Q}}_j(z)\right\|<+\infty\\ 0<\inf_{M\geq M_0}\inf_{z\in C_j}|\hat{\omega}_j(z)|\leq\sup_{M\geq M_0}\sup_{z\in C_j}|\hat{\omega}_j(z)|<+\infty\\ \inf_{M\geq M_0}\operatorname{dist}\left\{\hat{\omega}_j(z),(-\infty,\mu_{\text{inf}}^{(j)}]\right\}>0\end{gathered}\tag{D.2}$$

with probability one for some M_0 sufficiently high.

Proof. It is well known that all the eigenvalues of $\hat{\mathbf{R}}_j$ are located inside $\mathcal{S}_j$ (plus $\{0\}$ if $N_j \leq M$) almost surely for all large M [58]. This implies that

$$\sup_{z\in C_j}\sup_{M\geq M_0}\|\hat{\mathbf{Q}}_j(z)\|\leq\sup_{z\in C_j}\operatorname{dist}^{-1}(z,\mathcal{S}_j\cup\{0\})<+\infty$$

almost surely for sufficiently large M_0. Next, observe that

$$\operatorname{Im}(z)=\operatorname{Im}[\omega_j(z)]\left(1-\overline{\Gamma}_j(\omega_j(z))\right)\tag{D.3}$$

where we have defined

$$\overline{\Gamma}_j(\omega)=\Gamma_j(\omega,\omega^*)=\frac{1}{N_j}\sum_{m=1}^{\bar{M}_j}K_m^{(j)}\frac{\left(\gamma_m^{(j)}\right)^2}{|\gamma_m^{(j)}-\omega|^2}.\tag{D.4}$$

Note that, by the definition of $\omega_j(z)$, we always have

$$\sup_{z\in C_j} \overline{\Gamma}_j(\omega_j(z)) < 1. \tag{D.5}$$

Indeed, the above inequality holds uniformly on $\mathbb{C}\backslash\mathbb{R}$ and follows from (D.3) and the fact that $\operatorname{Im}(z)\operatorname{Im}(\omega_j(z)) > 0$. On $\mathbb{R}$ it follows from the fact that C_j does not intersect $\mathcal{S}_j$ so that $\omega_j(z)$ is real valued and is chosen as the only root for which this inequality holds. Now, if we take the supremum in M on the left hand side of (D.5) the same identity holds, although the strict inequality may become equality. In any case, we see that

$$\inf_{m,M}\inf_{z\in C_j} |\gamma_m^{(j)} - \omega_j(z)|^2 > 0.$$

This implies that

$$\|\mathbf{Q}_j(\omega_j(z))\| \leq (\min_{m=1,\dots,\bar{M}_j}\inf_{m,M}\inf_{z\in C_j} |\gamma_m^{(j)} - \omega_j(z)|^2)^{-1},$$

which is therefore uniformly bounded as we wanted to prove.

Regarding (D.1), we reason by contradiction. Suppose we can select a sequence of points $z_{(M)} \in C_j$ such that $\omega_j(z_{(M)}) \to 0$ as $M \to \infty$. From (2.1) and because of (D.5) this would imply $z_{(M)} \to 0$, contradicting the fact that the contour C_j does not cross $\{0\}$. Similarly, assume that there exists a sequence such that $|\omega_j(z_{(M)})| \to \infty$. From (2.1), this would imply $|z_{(M)}| \to \infty$, contradicting the fact that $z_{(M)} \in C_j$.

To show (D.2), we first observe that

$$\begin{aligned}\left|1 - \frac{1}{N_j}\operatorname{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right| &\geq \left|1 - \frac{1}{N_j}\operatorname{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))\right]\right| - \frac{1}{N_j}\left|\operatorname{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))\right] - \operatorname{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right| \\ &= \left|\frac{z}{\omega_j(z)}\right| - \left|\frac{1}{N_j}\operatorname{tr}\left[\mathbf{R}_j\mathbf{Q}_j(\omega_j(z))\right] - \frac{1}{N_j}\operatorname{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right|.\end{aligned}$$

The second term converges almost surely to zero whereas the first term is uniformly bounded in M thanks to (D.1). This implies that

$$\inf_{M\geq M_0}\inf_{z\in C}\left|1 - \frac{1}{N_j}\operatorname{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right| > 0 \tag{D.6}$$

almost surely for sufficiently large M_0. On the other hand, it is obvious that

$$|1 - N_j^{-1}\operatorname{tr}[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)]| < 1 + |\frac{1}{N_j}\operatorname{tr}[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)]| \leq 1 + \frac{M}{N_j}\|\hat{\mathbf{R}}_j\|\|\hat{\mathbf{Q}}_j(z)\|$$

so that

$$\sup_{M\geq M_0}\sup_{z\in C}\left|1 - \frac{1}{N_j}\operatorname{tr}\left[\hat{\mathbf{R}}_j\hat{\mathbf{Q}}_j(z)\right]\right| < +\infty.$$

This directly shows (D.2).

Next, if we let $\mu_0^{(j)} \leq 0$ denote the smallest root of

$$\mu_0^{(j)}(1 - \frac{1}{N_j}\text{tr}[\mathbf{R}_j\mathbf{Q}_j(\mu_0^{(j)})]) = 0$$

we can write (by subtracting this equation and (2.1))

$$z = \left(\omega_j\left(z\right) - \mu_0^{(j)}\right)\left(1 - \Gamma_j\left(\omega_j\left(z\right), \mu_0^{(j)}\right)\right).$$

By taking the real parts on both sides of the equation we obtain

$$\text{Re}(z) = \left(1 - \overline{\Gamma}(\omega_j(z))\right)\text{Re}\left[\omega_j\left(z\right) - \mu_0^{(j)}\right] + \frac{1}{N_j}\sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\left(\gamma_m^{(j)}\right)^2\left|\omega_j\left(z\right) - \mu_0^{(j)}\right|^2}{\left|\gamma_m^{(j)} - \omega_j\left(z\right)\right|^2\left(\gamma_m^{(j)} - \mu_0^{(j)}\right)}.$$

Now, assume that there exists a sequence of points $z_{(M)}$ in C_j such that $\omega_j(z_{(M)}) \to \theta$, where $\theta \in (-\infty, \mu_{\text{inf}}^{(j)}]$. From the above equation we can conclude that, along that subsequence,

$$\frac{1}{N_j}\sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\left(\gamma_m^{(j)}\right)^2\left(\theta - \mu_0^{(j)}\right)^2}{\left(\gamma_m^{(j)} - \theta\right)^2\left(\gamma_m^{(j)} - \mu_0^{(j)}\right)} + \left(1 - \Gamma(\theta)\right)\left(\theta - \mu_0^{(j)}\right) - \text{Re}(z_{(M)}) \to 0$$

or, alternatively

$$\left(1 - \Gamma(\theta, \mu_0^{(j)})\right)\left(\theta - \mu_0^{(j)}\right) - \text{Re}(z_{(M)}) \to 0$$

which is a contradiciton since (by Cauchy-Schwarz)

$$\left|\Gamma_j\left(\theta, \mu_0^{(j)}\right)\right|^2 \leq \Gamma_j\left(\theta\right)\Gamma_j\left(\mu_0^{(j)}\right) \leq 1$$

whereas we always have $(\theta - \mu_0^{(j)})\text{Re}\left[z_{(M)}\right] \leq 0$ by the construction of the contour. The last inequality in the above equation follows from the fact that both $\mu_0^{(j)}$ and θ correspond to values of z for which the root $\omega_j\left(z\right)$ is real-valued. The result for $\hat{\omega}_j\left(z\right)$ follows from the above and the fact that $\sup_{z\in C}\left|\hat{\omega}_j\left(z\right) - \omega_j\left(z\right)\right| \to 0$ almost surely as $M \to \infty$. □

Lemma D.2. *Let $j \in \{0, 1\}$. Under **(As1)-(As3)** we have*

$$\sup_M \sup_{z\in C_j} \overline{\Gamma}_j(\omega_j(z)) < 1$$

Proof. We first observe that for each $z \in C$ we have $\overline{\Gamma}_j(\omega_j(z)) < 1$. Indeed, if $\mathrm{Im}(z) \neq 0$ we can take imaginary parts on both sides of (2.1) and obtain

$$\mathrm{Im}(z) = \left(1 - \overline{\Gamma}_j(\omega_j(z))\right) \mathrm{Im}\left[\omega_j(z)\right] \tag{D.7}$$

Since $\mathrm{Im}[z]$ and $\mathrm{Im}[\omega_j(z)]$ have, by definition, the same sign, we see that $\overline{\Gamma}_j(\omega_j(z)) < 1$ for each $z \in C_j$. When $\mathrm{Im}(z) \neq 0$ the same property follows directly from the fact that $\omega_j(z)$ is chosen as the only real-valued root of (2.1) such that (2.2) holds, which is equivalent to $\overline{\Gamma}_j(\omega_j(z)) < 1$ because $\omega_j(z)$ is real valued.

So, it remains to prove that the supremum over M also holds. To that effect, we reason by contradiction. Assume that there exists a sequence of points $z_{(M)}$ in C_j such that

$$\overline{\Gamma}_j(\omega_j(z_{(M)})) \to 1. \tag{D.8}$$

Since $z_{(M)}$ is bounded, we can find a convergent subsequence (say $z_{(M')}$) such that $z_{(M')} \to z_*$, where $z_* \in C_j$ by compactness. Now, assume first that $\mathrm{Im}(z_*) \neq 0$. Clearly, from (D.7) we must have $|\mathrm{Im}[\omega_j(z_{(M')})]| \to \infty$ along that subsequence. However, since $|\gamma_m^{(j)} - \omega_j(z)| > |\mathrm{Im}[\omega_j(z)]|$ we have

$$\overline{\Gamma}_j(\omega_j(z_{(M')})) < \frac{1}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{\left(\gamma_m^{(j)}\right)^2}{\mathrm{Im}^2\left[\omega_j\left(z_{(M')}\right)\right]} \to 0$$

which clearly contradicts (D.8). Assume therefore that $\mathrm{Im}(z_*) = 0$, so that we either have $z_* < \theta_j^-$ or $z > \theta_j^+$, where θ_j^- and θ_j^+ respectively denote the lower and upper limits of the interval in (2.5). Assume that $z_* < \theta_j^-$ (the reasoning for $z_* > \theta_+$ being completely equivalent). By subtracting two instances of the equation in (2.1) when evaluating it at two different points $z = z_{(M')}$ and at $z = \theta_j^-$ in C_j, we see that

$$\theta_j^- - z_{(M')} = \left(1 - \Gamma_j(z_{(M')}, \theta_j^-)\right).$$

Observe that by construction of the contour we have

$$\lim_{M \to \infty} \left(\theta_j^- - z_{(M')}\right) = \theta_j^- - z_{(M')} > 0,$$

so that it follows from the above equation that $\liminf_{M \to 0}(\omega_j(\theta_j^-)) - \omega_j(z_{(M')})) > 0$. Now, consider the real-valued function $\Gamma_j(\omega)$ on the interval $(0, \omega_j\left(\theta_j^-\right))$ This is a strongly convex function since the second order derviative is bounded away from zero, that is

$$\inf_{\omega \in (0, \omega_j\left(\theta_j^-\right))} \Gamma_j''(\omega) = \frac{6}{N_j} \sum_{m=1}^{\bar{M}_j} \frac{K_m^{(j)} \left(\gamma_m^{(j)}\right)^2}{\left(\gamma_m^{(j)} - \omega_j\left(\theta_j^-\right)\right)^4} \equiv \varsigma_M > 0.$$

Furthermore, $\inf_M \varsigma_M > 0$ because $\gamma_m^{(j)} - \omega_j(\theta_j^-) \leq \sup_M(\gamma_{\bar{M}_j}^{(j)} + |\omega_j\left(\theta_j^-\right)|) < \infty$ and

$$\inf_M \varsigma_M \geq \inf_M \frac{6}{N_j} \frac{\sum_{m=1}^{\bar{M}_j} K_m^{(j)} \left(\gamma_m^{(j)}\right)^2}{\sup_M \left(\gamma_{\bar{M}_j}^{(j)} + \left|\omega_j\left(\theta_j^-\right)\right|\right)^4} > 0.$$

By strong convexity, we have

$$\begin{aligned} 1 - \Gamma_j(\omega_j(z_{(M')})) &\geq \Gamma_j(\omega_j(\theta_j^-)) - \Gamma_j(\omega_j(z_{(M')})) \geq \\ &\geq \frac{2}{N_j} \sum_{m=1}^{\bar{M}_j} K_m^{(j)} \frac{(\gamma_m^{(j)})^2}{(\gamma_m^{(j)} - \omega_j(z_{(M')}))^3} (\omega_j\left(\theta_j^-\right) - \omega_j(z_{(M')})) + \\ &\quad + \frac{\varsigma_M}{2} (\omega_j(\theta_j^-) - \omega_j(z_{(M')}))^2. \end{aligned}$$

By the definition of $z_{(M')}$ the left hand side of the above inequality converges to zero, leading to a contradiction. The same reasoning can be applied to the case where $z_* > \theta_j^+$. □

D.2 Auxiliary Lemmas for Theorem 2.2

In this appendix, we provide some bounds on expectations and variances of different random functions of complex variable. We will assume that assumptions (As1)-(As3) hold, and that the observations are Gaussian distributed. We will also keep the shorthand notation introduced in the previous section.

Lemma D.3. *Let* $\mathbf{A}$ *denote an* $M \times M$ *deterministic matrix with bounded spectral norm. Then, we can write*

$$\begin{aligned} \frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\phi_1\right] &= \frac{1}{N_1}\mathrm{tr}[\mathbf{A}\bar{\mathbf{Q}}_1] + \mathcal{O}(M^{-1}) \\ \frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\phi_1\right] &= \frac{z_1}{\omega_1}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right] + \mathcal{O}(M^{-1}) \end{aligned}$$

and also

$$\begin{aligned} \mathrm{var}\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\right]\phi_1\right) &= \mathcal{O}(M^{-2}) \\ \mathrm{var}\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\right]\phi_1\right) &= \mathcal{O}(M^{-2}). \end{aligned}$$

Lemma D.4. *Let* $\mathbf{A}, \mathbf{B}$ *denote two* $M \times M$ *deterministic matrices with bounded spectral norm. Then, we can write*

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right] + \frac{z_1 z_1'}{\omega_1\omega_1'}\frac{\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}(z_1)\mathbf{R}\bar{\mathbf{Q}}_{1'}\right]\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right]}{N_1^2(1-\gamma(z_1,z_1'))} + \mathcal{O}(M^{-1}) \quad \text{(D.9)}$$

together with

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] = \frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{B}\right] + z_1'\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right] + z_1'\frac{z_1 z_1'}{\omega_1\omega_1'}\frac{\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}(z_1)\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right]}{N_1^2(1-\gamma(z_1,z_1'))} + \mathcal{O}(M^{-1}) \quad \text{(D.10)}$$

and also

$$\mathrm{var}\,\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \mathcal{O}(M^{-2})$$
$$\mathrm{var}\,\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] = \mathcal{O}(M^{-2}).$$

Proof. The proof follows the same steps as the proof of [75, Lemma 12], so we provide here just a sketch of the main steps. The proof of the variances follows directly from the Nash-Poincaré variance inequality in (A.5), so that we will only prove the first two identities. Let us first consider the first and second identities. Developing with respect to $\mathbf{X}$ and applying the integration by parts formula in (A.4), we obtain

$$\mathbb{E}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] = \frac{z_1'}{\omega_1'}\mathbb{E}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\phi_1\right] - \frac{z_1'}{\omega_1'}\mathbb{E}\left[\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\frac{1}{N_1}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\right]\phi_1\right] - \frac{z_1'}{\omega_1'}\mathbb{E}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\alpha_1(z_1')\right] + \mathcal{O}(M^{-\mathbb{N}}) \quad \text{(D.11)}$$

where we recall that $\alpha_1(z_1) = N_1^{-1}\mathrm{tr}[\mathbf{R}_1\hat{\mathbf{Q}}_1\phi_1] - N_1^{-1}\mathrm{tr}[\mathbf{R}_1\bar{\mathbf{Q}}_1]$ and where we have used the identity

$$\frac{\partial}{\partial X_{ij}^*}\hat{\mathbf{Q}}_1 = -\hat{\mathbf{Q}}_1\mathbf{Y}_1\frac{\mathbf{e}_j\mathbf{e}_i^H}{N_1}\mathbf{R}_1^{1/2}\hat{\mathbf{Q}}_1$$

as well as the fact that $1 + N_1^{-1}\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\right] = \omega_1/z_1$. Usign the resolvent identity

$(\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1 = z_1'\hat{\mathbf{Q}}_{1'} + \mathbf{I}_M)$, this leads to

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\phi_1\right] + \frac{z_1'}{\omega_1'}\mathbb{E}\left[\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_{1'}\right]\frac{1}{N_1}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\right]\phi_1\right] + \frac{z_1'}{\omega_1'}\mathbb{E}\left[\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\bar{\mathbf{Q}}_{1'}\phi\right]\alpha_1(z_1')\right] + \mathcal{O}(M^{-\mathbb{N}}).$$

Hence, applying Lemma D.3 we see that

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right] + \frac{z_1z_1'}{\omega_1\omega_1'}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}(z_1)\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\phi_1\right] + \mathcal{O}(M^{-1})$$

where we have used the fact that $\mathbb{E}\left[|\alpha_1(z_1')|^2\right] = \mathcal{O}(M^{-2})$. Setting $\mathbf{A} = \mathbf{R}_1$ we readily see that

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\mathbf{B}\bar{\mathbf{Q}}_{1'}\right]}{N_1(1-\gamma_{11'})} + \mathcal{O}(M^{-1})$$

where we recall that

$$\gamma_{11'} = \frac{z_1}{\omega_1}\frac{z_1'}{\omega_1'}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right].$$

Inserting this into the above function, we arrive at the result in (D.9) and inserting it into (D.11) we obtain (D.10). □

Lemma D.5. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *denote two* $M \times M$ *deterministic matrices with bounded spectral norm. Then, we can write*

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{1}{1-\gamma_{11'}}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\Omega_1\left(\omega_1';\mathbf{B}\right)\bar{\mathbf{Q}}_{1'}\right] + \frac{z_1z_1'}{\omega_1\omega_1'}\frac{\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]\mathrm{tr}\left[\mathbf{R}_1^2\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_{1'}\Omega_1\left(\omega_1';\mathbf{B}\right)\bar{\mathbf{Q}}_{1'}\right]}{N_1^2\left(1-\gamma_{11'}\right)^2} + \frac{1}{z_1'}\frac{\phi_1\left(\omega_1';\mathbf{B}\right)}{1-\gamma_{11'}}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right] + \mathcal{O}(M^{-1}) \quad \text{(D.12)}$$

and

$$\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] = \frac{z_1'}{\omega_1'}\frac{\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\Omega_1\left(\omega_1';\mathbf{B}\right)\bar{\mathbf{Q}}_{1'}\mathbf{R}_1\right]}{N_1\left(1-\gamma_{11'}\right)} + \frac{z_1}{\omega_1}\left(\frac{z_1'}{\omega_1'}\right)^2\frac{\mathrm{tr}\left[\mathbf{A}\mathbf{R}_1^2\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_{1'}\right]\mathrm{tr}\left[\mathbf{R}_1^2\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_{1'}^2\Omega_1\left(\omega_1';\mathbf{B}\right)\right]}{N_1^2\left(1-\gamma_{11'}\right)^2} - \frac{z_1z_1'}{\omega_1\omega_1'}\frac{\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right]\mathrm{tr}\left[\mathbf{R}_1^2\bar{\mathbf{Q}}_1\bar{\mathbf{Q}}_{1'}^2\Omega_1\left(\omega_1';\mathbf{B}\right)\right]}{N_1^2\left(1-\gamma_{11'}\right)^2} + \mathcal{O}(M^{-1}). \quad \text{(D.13)}$$

On the other hand, we have

$$\operatorname{var}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \mathcal{O}(M^{-2})$$
$$\operatorname{var}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] = \mathcal{O}(M^{-2}).$$

Proof. The proof that the variance decays as $\mathcal{O}(M^{-2})$ follows the conventional approach from the Nash-Poincaré inequality, and is therefore omitted. To proof the first two identities, we proceed as in the proof of Lemma D.4. Using first the resolvent identity on $\hat{\mathbf{Q}}_{1'} = (z_1')^{-1}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1 - (z_1')^{-1}\mathbf{I}_M$ together with the integration by parts formula and Lemmas D.3 and D.4, we can write

$$\begin{aligned}\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] &= \frac{1}{z_1'}\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\phi_1\right]\\
&\quad - \frac{1}{z_1'}\frac{z_1}{\omega_1'}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right]\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right]\\
&\quad + \frac{1}{z_1'}\left(1-\frac{\omega_1'}{z_1'}\right)\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right]\\
&\quad\quad - \frac{1}{z_1'}\frac{1}{1-\gamma_{11'}}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{B}\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]\\
&\quad\quad - \frac{\omega_1'}{(z_1')^2}\frac{\phi_1(\omega_1';\mathbf{B})}{1-\gamma_{11'}}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]\\
&\quad\quad\quad - \frac{\omega_1'}{(z_1')^3}\phi_1(\omega_1';\mathbf{B})\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right] + \mathcal{O}(M^{-1}) \qquad \text{(D.14)}\end{aligned}$$

where we have decorrelated the double terms using the fact that all variances decay as $\mathcal{O}(M^{-2})$ and we have used the identity $1 + N_1^{-1}\operatorname{tr}\left[\mathbf{R}_1\bar{\mathbf{Q}}_1\right] = \omega_1'/z_1$. In a similar way, we can develop the term

$$\begin{aligned}\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\hat{\mathbf{R}}_1\phi_1\right] &=\\
&= \frac{z_1'}{\omega_1'}\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\mathbf{R}_1\phi_1\right]\\
&\quad - \frac{z_1'}{\omega_1'}\frac{z_1}{\omega_1}\frac{1}{N_1}\mathbb{E}\operatorname{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right]\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right]\\
&\quad\quad - \frac{1}{\omega_1'}\frac{\phi_1(\omega_1';\mathbf{B})}{1-\gamma_{11'}}\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\mathbf{R}_1\right]\\
&\quad\quad + \frac{1}{z_1'}\frac{\gamma_{11'}}{1-\gamma_{11'}}\phi_1(\omega_1';\mathbf{B})\frac{1}{N_1}\operatorname{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right] + \mathcal{O}(M^{-1}) \qquad \text{(D.15)}\end{aligned}$$

Inserting this back into the first equation and replacing $\mathbf{A}$ with $\bar{\mathbf{Q}}_{1'}\mathbf{A}$ we obtain

$$\begin{aligned}\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{A}\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \\ = \frac{1}{1-\gamma_{11'}}\frac{1}{N_1}\mathrm{tr}\left[\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\Omega_1\left(\omega_1';\mathbf{B}\right)\bar{\mathbf{Q}}_{1'}\right] \\ + \left(\frac{z_1}{\omega_1}\frac{z_1'}{\omega_1'}\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] + \frac{\phi_1\left(\omega_1';\mathbf{B}\right)}{z_1'}\right)\times \\ \times\frac{1}{N_1}\mathrm{tr}\left[\bar{\mathbf{Q}}_{1'}\mathbf{A}\bar{\mathbf{Q}}_1\mathbf{R}_1\right]+\mathcal{O}(M^{-1}) \quad \text{(D.16)}\end{aligned}$$

Particularizing this expression for $\mathbf{A}=\mathbf{R}_1$ we see that

$$\begin{aligned}\frac{1}{N_1}\mathbb{E}\mathrm{tr}\left[\mathbf{R}_1\hat{\mathbf{Q}}_1\mathbf{R}_1\hat{\mathbf{Q}}_{1'}\mathbf{B}\hat{\mathbf{Q}}_{1'}\phi_1\right] = \frac{\omega_1'}{\left(z_1'\right)^2}\frac{\omega_1}{z_1}\frac{\gamma_{11'}\phi_1\left(\omega_1';\mathbf{B}\right)}{1-\gamma_{11'}} \\ + \frac{1}{\left(1-\gamma_{11'}\right)^2}\frac{1}{N_1}\mathrm{tr}\left[\Omega_1\left(\omega_1';\mathbf{B}\right)\bar{\mathbf{Q}}_{1'}\mathbf{R}_1\bar{\mathbf{Q}}_1\mathbf{R}_1\bar{\mathbf{Q}}_{1'}\right]+\mathcal{O}(M^{-1})\end{aligned}$$

and inserting this back into (D.16), together with (D.15), we obtain (D.13). □

Appendix E

Synthesis of Developed Methods

In this appendix, we provide a concise summary of all the methods proposed in this thesis, encompassing their general and closed form solutions alongside with corresponding assumptions and remarks on their applicability. The purpose of this appendix is to serve as a convenient reference for readers, rather than introducing new material. Also, we often refer the reader to different chapters of this thesis for a detailed explanation of the different results presented throughout this thesis.

Let us start by recalling that the main focus of this thesis lies around the study of distances between sample covariance matrices that can be written as

$$\hat{d}_M = \sum_{l=1}^{L} \frac{1}{M} \mathrm{tr}\left[f_1^{(l)}\left(\hat{\mathbf{R}}_1\right) f_2^{(l)}\left(\hat{\mathbf{R}}_2\right)\right]$$

for certain functions $f_1^{(l)}, f_2^{(l)} : \mathbb{C}^{M\times M} \to \mathbb{C}^{M\times M}$, $l = 1, \ldots, L$. Typically, these functions are understood to be the result of applying scalar analytic functions to the real eigenvalues of the Hermitian matrices $\mathbf{R}_j$, $j \in \{1, 2\}$. With some abuse of notation, $f_j^{(l)}, l = 1, \ldots, L$ will also denote these scalar functions. Throughout this thesis, we have mainly considered the following assumptions (see Section 2.1 for details):

Assumption 1 (As1): For $j \in \{1, 2\}$ and $k = 1, \ldots, N_j$ the observations $\mathbf{y}_j(k)$ (see Remark 1.1) are all independent and can be expressed as

$$\mathbf{y}_j(k) = \mathbf{R}_j^{\frac{1}{2}} \mathbf{x}_j(k)$$

where $\mathbf{R}_j$ is an Hermitian positive definite matrix and $\mathbf{x}_j(k)$ is a vector of i.i.d. random entries with zero mean and unit variance.

Assumption 2 (As2): The different eigenvalues of $\mathbf{R}_j$, denoted by $0 < \gamma_1^{(j)} < \ldots < \gamma_{\bar{M}_j}^{(j)}$ ($j \in \{1, 2\}$), may vary with M but always have but we always have $\inf_M \gamma_1^{(j)} > 0$ and $\sup_M \gamma_{\bar{M}}^{(j)} < \infty$, where $\bar{M}_j$ is the total number of distinct eigenvalues.

Assumption 3 (As3): The quantities N_1 and N_2 depend on M, that is $N_1 = N_1(M)$ and $N_2 = N_2(M)$. Furthermore, when $M \to \infty$ we have, for $j \in \{1, 2\}$, $N_j(M) \to \infty$ in a way that $M/N_j \to c_j$ for some constant $0 < c_j < \infty$ such that $c_j \neq 1$.

Assumption 4 (As4): For $j \in \{1, 2\}$ and $l = 1, \ldots, L$, the quantity $f_j^{(l)}(\hat{\mathbf{R}}_j)$ can be expressed as

$$f_j^{(l)}\left(\hat{\mathbf{R}}_j\right) = \frac{1}{2\pi \mathrm{j}} \oint_{C_j^-} f_j^{(l)}(z)\hat{\mathbf{Q}}_j(z)dz \tag{E.1}$$

with probability one for all large M, where C_j^- is a negatively oriented simple closed contour enclosing

$$\mathcal{S}_j = \left[\inf_M \left[\gamma_1^{(j)} \times \left(1 - \sqrt{M/N_j}\right)^2\right] \quad , \quad \sup_M \left[\gamma_M^{(j)} \times \left(1 + \sqrt{M/N_j}\right)^2\right]\right].$$

and not crossing zero and where, with some abuse of notation, $f_j^{(l)}(z)$ denotes a complex function analytic on an open set including C_j.

E.1 *Plug-in* Distance

We recall the general results from Corollary 2.1 and Theorem 2.2, so that assuming that **(As1)-(As4)** hold and that the observations are Gaussian distributed. If $\liminf_{M\to\infty} \sigma_M^2 > 0$ we have

$$\frac{M(\hat{d}_M - \bar{d}_M) - \mathfrak{m}_M}{\sigma_M} \to \mathcal{N}(0, 1).$$

Throughout this thesis we have particularized each of these quantities, namely $\bar{d}_M, \mathfrak{m}_M$ and σ_M, for the Euclidean, symmetrized KL and Subspace distances. When not mentioned, the results are understood to hold both for the undersampled and oversampled regimes.

E.1.1 Euclidean Distance

- *Distance*:

$$\hat{d}_M^E = \frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_1 - \mathbf{R_2})^2\right]$$

- *Asymptotic equivalent*:

$$\bar{d}_M^E = \frac{1}{M}\mathrm{tr}\left[(\mathbf{R}_1 - \mathbf{R}_2)^2\right] + \frac{1}{MN_1}\mathrm{tr}^2\left[\mathbf{R}_1\right] + \frac{1}{MN_2}\mathrm{tr}^2\left[\mathbf{R}_2\right]$$

- *Asymptotic (second order) mean of* $\hat{\varsigma}_M^E$:

$$\mathfrak{m}_M^E = \varsigma \left(\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1^2\right] + \frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right] \right)$$

- *Asymptotic variance of* $\hat{\varsigma}_M^E$:

$$\begin{aligned} \frac{\left(\sigma_M^E\right)^2}{1+\varsigma} = & 2\left(\frac{1}{N_1} \mathrm{tr}\left[\mathbf{R}_1^2\right] \right)^2 + 2\left(\frac{1}{N_2} \mathrm{tr}\left[\mathbf{R}_2^2\right] \right)^2 + \frac{4}{N_1 N_2} \mathrm{tr}^2[\mathbf{R}_1 \mathbf{R}_2] \\ & + \frac{4}{N_1} \mathrm{tr}\left[(\mathbf{R}_1 \Delta_1)^2\right] + \frac{4}{N_2} \mathrm{tr}\left[(\mathbf{R}_2 \Delta_2)^2\right] \end{aligned}$$

E.1.2 Symmetrized KL Distance

- *Distance*:

$$\hat{d}_M^{KL} = \frac{1}{2M} \mathrm{tr}\left[\mathbf{R}_1 \mathbf{R}_2^{-1} + \mathbf{R}_1^{-1} \mathbf{R}_2\right] - 1$$

where, in the undersampled regime, one can replace $\hat{\mathbf{R}}_j^{-1}$ *by* $\mathbf{R}_j^{\#}$*, for* $j = 1, 2$.

- *Asymptotic equivalent*:

$$\bar{d}_M^{KL} = \frac{\mathrm{tr}\left[\mathbf{R}_1 \mathbf{Q}_1^2(\mu_0^{(1)}) \mathbf{R}_2\right]}{2M\left(1 - \Gamma_1\left(\mu_0^{(1)}\right)\right)} + \frac{\mathrm{tr}\left[\mathbf{R}_2 \mathbf{Q}_2^2(\mu_0^{(2)}) \mathbf{R}_1\right]}{2M\left(1 - \Gamma_2\left(\mu_0^{(2)}\right)\right)} - 1,$$

where

$$\Gamma_j(\omega) = \frac{1}{N_j} \mathrm{tr}\left[\mathbf{R}_j^2 \mathbf{Q}_j^2(\omega)\right]$$

and $\mu_0^{(j)}, j = 1, 2$ *is the smallest solution to (2.3).*

- *Asymptotic equivalent tailored to the oversampled regime*:

$$\bar{d}_M^{KL} = \frac{1}{2M} \left(\frac{N_1 \mathrm{tr}\left[\mathbf{R}_1^{-1} \mathbf{R}_2\right]}{N_1 - M} + \frac{N_2 \mathrm{tr}\left[\mathbf{R}_2^{-1} \mathbf{R}_1\right]}{N_2 - M} \right) - 1.$$

- *Asymptotic (second order) mean of* $\hat{\varsigma}_M^{KL}$:

$$\mathfrak{m}_M^{KL} = \varsigma \sum_{\substack{i,j \in \{1,2\} \\ i \neq j}} \frac{d[\omega_i \mathfrak{m}_i(\omega_i, \mathbf{R}_j)]/d\omega_i \big|_{\omega_i = \mu_0^{(i)}}}{2\left(1 - \Gamma_i(\mu_0^{(i)})\right)}$$

where we recall that $\mathfrak{m}_j(\omega, \mathbf{A})$ *is defined as in (2.18).*

- *Asymptotic (second order) mean of* $\hat{\varsigma}_M^{KL}$ tailored to the oversampled regime:

$$\mathfrak{m}_M^{KL} = \frac{\varsigma}{2}\left[\frac{N_1 \mathrm{tr}[\mathbf{R}_2\mathbf{R}_1^{-1}]}{(N_1-M)^2} + \frac{N_2 \mathrm{tr}[\mathbf{R}_1\mathbf{R}_2^{-1}]}{(N_2-M)^2}\right].$$

- *Asymptotic variance of* $\hat{\varsigma}_M^{KL}$:

$$\frac{\left(\sigma_M^{KL}\right)^2}{1+\varsigma} = \frac{\partial^2\left[\omega_1\omega_1'\Upsilon_{11}\left(\omega_1,\omega_1'\right)\right]/\partial\omega_1\partial\omega_1'\big|_{\omega_1=\omega_1'=\mu_0^{(1)}}}{4\left(1-\Gamma_1(\mu_0^{(1)})\right)^2} + \frac{\partial^2\left[\omega_2\omega_2'\Upsilon_{22}\left(\omega_2,\omega_2'\right)\right]/\partial\omega_2\partial\omega_2'\big|_{\omega_2=\omega_2'=\mu_0^{(2)}}}{4\left(1-\Gamma_2(\mu_0^{(2)})\right)^2} + \frac{\partial^2\left[\omega_1\omega_2\Upsilon_{12}\left(\omega_1,\omega_2\right)\right]/\partial\omega_1\partial\omega_2\big|_{\omega_1=\mu_0^{(1)},\omega_2=\mu_0^{(2)}}}{2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}$$

where we have defined

$$\Upsilon_{11}\left(\omega_1,\omega_1'\right) = \frac{\mathrm{tr}^2\left[\mathbf{R}_2\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{Q}_1\left(\omega_1'\right)\right]}{N_1N_2\left(1-\Gamma_1(\omega_1,\omega_1')\right)} + \sigma_1^2\left(\omega_1,\omega_1';\mathbf{R}_2,\mathbf{R}_2\right) + \frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_1\left(\omega_1'\right)\right]$$

where $\Upsilon_{22}\left(\omega_2,\omega_2'\right)$ *is defined equivalently and*

$$\Upsilon_{12}\left(\omega_1,\omega_2\right) = \frac{1}{N_1N_2}\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\right] - \frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\mathbf{Q}_2\left(\omega_2\right)\mathbf{R}_1\mathbf{Q}_1\left(\omega_1\right)\Omega_1\left(\omega_1;\mathbf{R}_2\right)\right] - \frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\mathbf{Q}_1\left(\omega_1\right)\mathbf{R}_2\mathbf{Q}_2\left(\omega_2\right)\Omega_2\left(\omega_2;\mathbf{R}_1\right)\right].$$

Finally, $\sigma_j^2(\omega_j,\omega_j,\mathbf{R}_k,\mathbf{R}_k)$, for $j,k=1,2$ and $j\neq k$ is defined as in (2.22).

- *Asymptotic variance of* $\hat{\varsigma}_M^{KL}$ tailored to the oversampled regime:

$$\frac{(\sigma_M^{KL})^2}{1+\varsigma} = \frac{N_1^2\Upsilon_{11}(0,0)}{4(N_1-M)^2} + \frac{N_2^2\Upsilon_{22}(0,0)}{4(N_2-M)^2} + \frac{N_1N_2\Upsilon_{12}(0,0)}{2(N_1-M)(N_2-M)}$$

where

$$\Upsilon_{11}(0,0) = \frac{N_1+N_2-M}{N_2(N_1-M)}\left[\mathrm{tr}[(\mathbf{R}_1^{-1}\mathbf{R}_2)^2] + \frac{\mathrm{tr}^2[\mathbf{R}_1^{-1}\mathbf{R}_2]}{N_1-M}\right]$$

with $\Upsilon_{22}(0,0)$ *equivalently defined by swapping indexes* $(1\leftrightarrow 2)$*, and where*

$$\Upsilon_{12}(0,0) = \frac{M^2}{N_1N_2} - \frac{M}{N_1} - \frac{M}{N_2}.$$

E.1.3 Subspace Similarity

This distance and all their related results are only valid in the undersampled regime (otherwise, one cannot possibly define the original subspaces).

- *Distance*:

$$\hat{d}_M^{SS} = \frac{1}{M}\mathrm{tr}\left[(\mathbf{P}_1 - \mathbf{P}_2)^2\right]$$

where $\mathbf{P}_i = \mathbf{Y}_i \left(\mathbf{Y}_i^H \mathbf{Y}_i\right)^{-1} \mathbf{Y}_i^H$ *is the projection matrix onto the column space.*

- *Asymptotic equivalent in undersampled regime*:

$$\bar{d}_M^{SS} = \frac{N_1}{M} + \frac{N_2}{M} - \frac{2}{M}\mathrm{tr}\left[\mathbf{R}_1\mathbf{Q}_1(\mu_0^{(1)})\mathbf{R}_2\mathbf{Q}_2(\mu_0^{(2)})\right]$$

- *Asymptotic (second order) mean of* $\hat{\varsigma}_M^{SS}$ *in the undersampled regime*:

$$\mathfrak{m}_M^{SS} = -2\mu_0^{(1)}\mathfrak{m}_1(\mu_0^{(1)}, \mathbf{R}_2\mathbf{Q}_2(\mu_0^{(2)})) - 2\mu_0^{(2)}\mathfrak{m}_2(\mu_0^{(2)}, \mathbf{R}_1\mathbf{Q}_1(\mu_0^{(1)})),$$

see (2.18) for the definition of $\mathfrak{m}_j(\omega, \mathbf{A})$.

- *Asymptotic variance of* $\hat{\varsigma}_M^{SS}$ in the undersampled regime:

$$\begin{aligned}\frac{\left(\sigma_M^{SS}\right)^2}{1+\varsigma} &= 4\left(\mu_0^{(1)}\right)^2 \sigma_1^2\left(\mu_0^{(1)}, \mu_0^{(1)}; \mathbf{R}_2\mathbf{Q}_2\left(\mu_0^{(2)}\right), \mathbf{R}_2\mathbf{Q}_2\left(\mu_0^{(2)}\right)\right) \\ &+ 4\left(\mu_0^{(2)}\right)^2 \sigma_2^2\left(\mu_0^{(2)}, \mu_0^{(2)}; \mathbf{R}_1\mathbf{Q}_1\left(\mu_0^{(1)}\right), \mathbf{R}_1\mathbf{Q}_1\left(\mu_0^{(1)}\right)\right) \\ &+ 4\left(\mu_0^{(1)}\mu_0^{(2)}\right)^2 \frac{\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{Q}_1^2(\mu_0^{(1)})\mathbf{R}_2\mathbf{Q}_2^2(\mu_0^{(2)})\right]}{N_1 N_2\left(1-\Gamma_1(\mu_0^{(1)})\right)\left(1-\Gamma_2(\mu_0^{(2)})\right)}.\end{aligned}$$

E.1.4 *Plug-in* Correction Terms Tailored to $\mathbf{R}_1 = \mathbf{R}_2$

We recall the correction terms (see Chapter 5 for a detailed explanation) designed for the *plug-in* distances. These correction terms are designed to correctly approximate their asymptotic equivalents under the null hypothesis (i.e., $\mathbf{R}_1 = \mathbf{R}_2$) while penalizing comparisons under the alternative hypothesis (i.e., $\mathbf{R}_1 \neq \mathbf{R}_2$). These results are inspired by the clustering of wireless devices, hence are tailored for complex observations ($\varsigma = 0$).

Correction Term of Euclidean Distance

$$\hat{\tilde{d}}_M^E = \frac{1}{MN_k}\mathrm{tr}^2[\hat{\mathbf{R}}_k] + \frac{1}{MN_j}\mathrm{tr}^2[\hat{\mathbf{R}}_j].$$

Correction Term of Symmetrized KL Distance

$$\hat{\tilde{d}}_M^{KL} = \frac{N_k}{2M}\left(\frac{N_j^{-1}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_j^{\#}\right)^2\right]}{\left(N_j^{-1}\mathrm{tr}\left[\hat{\mathbf{R}}_j^{\#}\right]\right)^2} - 1\right) + \frac{N_j}{2M}\left(\frac{N_k^{-1}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_k^{\#}\right)^2\right]}{\left(N_k^{-1}\mathrm{tr}\left[\hat{\mathbf{R}}_k^{\#}\right]\right)^2} - 1\right) - 1.$$

Correction Term of Subspace Similarity

$$\hat{\tilde{s}}_{kj}^{PF} = \begin{cases} \frac{N_k}{2M}\left(1 - \frac{\hat{\kappa}_k^2(1)}{\hat{\kappa}_k(2)}\right) + \frac{N_j}{2M}\left(1 - \frac{\hat{\kappa}_j^2(1)}{\hat{\kappa}_j(2)}\right) & N_k = N_j \\ \frac{1}{2M}\left(\frac{N_k\hat{v}_k(k) - N_j\hat{v}_j(k)}{\hat{v}_k(k) - \hat{v}_j(k)} + \frac{N_k\hat{v}_k(j) - N_j\hat{v}_j(j)}{\hat{v}_k(j) - \hat{v}_j(j)}\right) & N_k \neq N_j \end{cases}$$

where we have defined, for $l \in \{k, j\}$,

$$\hat{\kappa}_l(m) = \frac{1}{N_l}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_l^{\#}\right)^m\right],\ m \in \mathbb{N}.$$

Furthermore, for $N_k = N_j$, we have $\hat{v}_\ell(l) = -\hat{\kappa}_\ell^{-1}(1)$, whereas for $N_k \neq N_j$ we take

$$\hat{v}_j(k) = \gamma\left(1 - \frac{N_j}{N_k}\right)$$

where γ is the smallest solution to

$$\frac{1}{N_j}\mathrm{tr}\left[\hat{\mathbf{R}}_k\left(\hat{\mathbf{R}}_k - \gamma\mathbf{I}_M\right)^{-1}\right] = 1.$$

E.2 Consistent Estimators

We briefly summary the expressions related to the consistent estimator of a distance between covariance matrices, namely $\tilde{d}_M$. The main idea is to design consistent estimators that approximate the true distance between covariance matrices, i.e., $d_M - \tilde{d}_M \to 0$, almost surely. In (3.8) we have proposed a general expression of $\tilde{d}_M$. Note that this is different from correction terms presented above for the *plug-in* distance. Those were designed to penalize the scenario where $\mathbf{R}_1\mathbf{R}_2$ and are only valid in for complex-valued observations. In this appendix, we recall some of the results presented in Chapter 3, including the closed form solutions of these consistent estimators for the Euclidean distance, symmetrized KL distance and log-Euclidean. Finally, we also summary the CLT of these quantities.

E.2.1 Euclidean Distance

- *Consistent estimator*:

$$\tilde{d}_M^E = \frac{1}{M}\mathrm{tr}\left[\left(\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2\right)^2\right] - \frac{1}{MN_1}\mathrm{tr}^2\left[\hat{\mathbf{R}}_j\right] - \frac{1}{MN_2}\mathrm{tr}^2\left[\hat{\mathbf{R}}_2\right].$$

- *Asymptotic (second order) mean of $\tilde{\varsigma}_M$:*

$$\tilde{\mathfrak{m}}_M^E = \varsigma\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\right] + \frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2^2\right]\right).$$

- *Asymptotic variance of $\tilde{\varsigma}_M^E$*:

$$\begin{aligned}\frac{\tilde{\sigma}_M^2}{1+\varsigma} &= 2\left(\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1^2\right]\right)^2 + 4\frac{1}{N_1}\mathrm{tr}\left[\mathbf{R}_1\mathbf{\Delta}\mathbf{R}_1\mathbf{\Delta}\right] \\ &\quad + 2\left(\frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2^2\right]\right)^2 + 4\frac{1}{N_2}\mathrm{tr}\left[\mathbf{R}_2\mathbf{\Delta}\mathbf{R}_2\mathbf{\Delta}\right] \\ &\quad + 4\frac{1}{N_1N_2}\mathrm{tr}^2\left[\mathbf{R}_1\mathbf{R}_2\right]\end{aligned}$$

where, now, $\mathbf{\Delta} = \mathbf{R}_1 - \mathbf{R}_2$.

E.2.2 Symmetrized KL Distance

We recall that the function ω^{-1} is not holomorphic at the origin, which implies that we can only tolerate $\mu_{inf} = 0$. In particular, this implies that we can only obtain a consistent estimator for the oversampled case (namely $N_1 > M$ and $N_2 > 0$).

- *Consistent estimator*:

$$\tilde{d}_M^{KL} = \left(1 - \frac{M}{N_1}\right)\frac{1}{2M}\mathrm{tr}\left[\hat{\mathbf{R}}_1^{-1}\hat{\mathbf{R}}_2\right] + \left(1 - \frac{M}{N_2}\right)\frac{1}{2M}\mathrm{tr}\left[\hat{\mathbf{R}}_2^{-1}\hat{\mathbf{R}}_1\right] - 1.$$

- *Asymptotic (second order) mean of $\tilde{\varsigma}_M^{KL}$*:

$$\tilde{\mathfrak{m}}_M^{KL} = \frac{\varsigma}{2}\left(\frac{1}{N_1 - M}\mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\right] + \frac{1}{N_2 - M}\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right]\right).$$

- *Asymptotic variance $\tilde{\varsigma}_M^{KL}$*:

$$\begin{aligned}\frac{\tilde{\sigma}_M^2}{(1+\varsigma)(N_1+N_2-M)} &= -\frac{M}{2N_1N_2} + \frac{\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\mathbf{R}_1\mathbf{R}_2^{-1}\right]}{4\left(N_2 - M\right)N_1} + \frac{\mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\mathbf{R}_1^{-1}\mathbf{R}_2\right]}{4\left(N_1 - M\right)N_2} \\ &\quad + \frac{1}{4N_1}\left(\frac{\mathrm{tr}\left[\mathbf{R}_1\mathbf{R}_2^{-1}\right]}{N_2 - M}\right)^2 + \frac{1}{4N_2}\left(\frac{\mathrm{tr}\left[\mathbf{R}_1^{-1}\mathbf{R}_2\right]}{N_1 - M}\right)^2.\end{aligned}$$

E.2.3 Log-Euclidean Distance

- *Consistent estimator*:

$$\tilde{d}_M^{LE} = \alpha^{(1)} + \alpha^{(2)} - \frac{2}{M}\sum_{k=1}^{M}\sum_{m=1}^{M}\beta_k^{(1)}\beta_m^{(2)}\left|\left(\hat{\mathbf{e}}_k^{(1)}\right)^H\hat{\mathbf{e}}_m^{(2)}\right|^2$$

where

$$\begin{aligned}\alpha^{(j)} = &\left(\frac{N_j}{M}-1\right)\sum_{r=1}^{M}\left(1+\log\hat{\mu}_r^{(j)}\right)^2 - \left(1+\log\hat{\lambda}_k^{(j)}\right)^2 \\ &+\frac{1}{M}\sum_{k=1}^{M}\left(1+\log\hat{\lambda}_k^{(j)}\right)^2 - \left(\frac{N_j}{M}-1\right)\log^2\left(1-\frac{M}{N_j}\right) \\ &+1+\frac{2}{M}\sum_{k=1}^{M}\sum_{r=1}^{M}\left[\Phi_2\left(\frac{\hat{\mu}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\right)-\Phi_2\left(\frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\right)\right] \\ &+\frac{2}{M}\sum_{k=1}^{M}\left(\sum_{\substack{r=1\\r\neq k}}^{M}\log\frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\log\frac{\hat{\lambda}_k^{(j)}}{\left|\hat{\lambda}_k^{(j)}-\hat{\lambda}_r^{(j)}\right|}\right. \\ &\left.-\sum_{r=1}^{M}\log\frac{\hat{\mu}_r^{(j)}}{\hat{\lambda}_k^{(j)}}\log\frac{\hat{\lambda}_k^{(j)}}{\left|\hat{\lambda}_k^{(j)}-\hat{\mu}_r^{(j)}\right|}\right)\end{aligned}$$

where we have defined

$$\Phi_2(x) = \begin{cases}\mathrm{Li}_2(x) & x<1\\ \frac{\pi^2}{3}-\frac{1}{2}\log^2 x - \mathrm{Li}_2\left(x^{-1}\right) & x\geq 1\end{cases}$$

and where $\mathrm{Li}_2(x) = -\int_0^x y^{-1}\log(1-y)dy$ is the dilogarithm function.

The coefficients $\beta_k^{(j)}$, $k=1,\ldots,M$, are defined as

$$\begin{aligned}\beta_k^{(j)} = 1 &+ \left(1+\sum_{\substack{m=1\\m\neq k}}^{M}\frac{\hat{\lambda}_k^{(j)}}{\hat{\lambda}_m^{(j)}-\hat{\lambda}_k^{(j)}} - \sum_{m=1}^{M}\frac{\hat{\mu}_k^{(j)}}{\hat{\lambda}_m^{(j)}-\hat{\mu}_k^{(j)}}\right)\log\hat{\lambda}_k^{(j)} \\ &+\left(\sum_{\substack{r=1\\r\neq k}}^{M}\frac{\hat{\lambda}_r^{(j)}}{\hat{\lambda}_r^{(j)}-\hat{\lambda}_k^{(j)}}\log\hat{\lambda}_r^{(j)} - \sum_{r=1}^{M}\frac{\hat{\mu}_r^{(j)}}{\hat{\mu}_r^{(j)}-\hat{\lambda}_k^{(j)}}\log\hat{\mu}_r^{(j)}\right)\end{aligned}$$

where we have denoted by $\hat{\mu}_0^{(j)} < \ldots < \hat{\mu}_M^{(j)}$ the M solutions to the polynomial equation in (3.3) by interchanging the covariance matrix $\mathbf{R}_j$ by its estimator $\hat{\mathbf{R}}_j$.

- *Asymptotic (second order) mean of $\hat{\varsigma}_M^{LE}$*:

$$\begin{aligned}\tilde{\mathfrak{m}}_M^{LE} = &-\varsigma \sum_{m=1}^{\bar{M}_1} \frac{1}{K_m^{(1)}} \mathrm{tr}\left[\mathbf{E}_m^{(1)}(\mathbf{E}_m^{(1)})^H \left(\log \gamma_m^{(1)} \mathbf{I}_M - \log \mathbf{R}_2\right)^2\right] \\ &-\varsigma \sum_{m=1}^{\bar{M}_2} \frac{1}{K_m^{(2)}} \mathrm{tr}\left[\mathbf{E}_m^{(2)}(\mathbf{E}_m^{(2)})^H \left(\log \mathbf{R}_1 - \log \gamma_m^{(2)} \mathbf{I}_M\right)^2\right] \\ &+\frac{\varsigma}{2} \sum_{m=1}^{2\bar{M}_1} \frac{\mathrm{tr}\left[\mathbf{R}_1^2 \mathbf{Q}_1^3(\theta_m^{(1)}) \left(\log \theta_m^{(1)} \mathbf{I}_M - \log \mathbf{R}_2\right)^2\right]}{\mathrm{tr}\left[\mathbf{R}_1^2 \mathbf{Q}_1^3(\theta_m^{(1)})\right]} \\ &+\frac{\varsigma}{2} \sum_{m=1}^{2\bar{M}_2} \frac{\mathrm{tr}\left[\mathbf{R}_2^2 \mathbf{Q}_2^3(\theta_m^{(2)}) \left(\log \mathbf{R}_1 - \log \theta_m^{(2)} \mathbf{I}_M\right)^2\right]}{\mathrm{tr}\left[\mathbf{R}_2^2 \mathbf{Q}_2^3(\theta_m^{(2)})\right]}\end{aligned}$$

  where $\theta_m^{(j)}$, $m = 1, \ldots, 2\bar{M}_j$ are the solutions to $\Gamma_j(\omega_j) = 1, j = 1, 2$.

- *Asymptotic variance $\hat{\varsigma}_M^{LE}$*: We evaluate the variance using using numerical integration.

Bibliography

[1] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction," in *International work-conference on artificial neural networks*. Springer, 2005, pp. 758–770.

[2] R. Bellman and R. Kalaba, "On adaptive control processes," *IRE Transactions on Automatic Control*, vol. 4, no. 2, pp. 1–9, 1959.

[3] I. T. Jolliffe, "Principal component analysis," *Technometrics*, vol. 45, no. 3, p. 276, 2003.

[4] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[5] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.

[6] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, 2016, pp. 90–94.

[7] R. Pereira, X. Mestre, and D. Gregoratti, "Subspace Based Hierarchical Channel Clustering in Massive MIMO," in *2021 IEEE Globecom Workshops (GC Workshops)*. IEEE, 2021, pp. 1–6.

[8] C. Ye, K. Slavakis, P. V. Patil, J. Nakuci, S. F. Muldoon, and J. Medaglia, "Network clustering via kernel-ARMA modeling and the Grassmannian: The brain-network case," *Signal Processing*, vol. 179, p. 107834, 2021.

[9] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39 501–39 514, 2018.

[10] L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 90–105, jun 2004. [Online]. Available: https://doi.org/10.1145/1007730.1007731

[11] E. Elhamifar and R. Vidal, "Sparse manifold clustering and embedding," *Advances in neural information processing systems*, vol. 24, 2011.

[12] Z. Li, Y. Chen, Y. LeCun, and F. T. Sommer, "Neural manifold clustering and embedding," *arXiv preprint arXiv:2201.10000*, 2022.

[13] D. S. Satish and C. C. Sekhar, "Kernel based clustering and vector quantization for speech recognition," in *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.* IEEE, 2004, pp. 315–324.

[14] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001.

[15] V. M. Patel and R. Vidal, "Kernel sparse subspace clustering," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 2849–2853.

[16] X. Mestre, R. Pereira, and D. Gregoratti, "Asymptotic spectral behavior of kernel matrices in complex valued observations," in *2021 IEEE Data Science and Learning Workshop (DSLW)*. IEEE, 2021, pp. 1–6.

[17] Z. Liao and R. Couillet, "On Inner-product Kernels of High Dimensional Data," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 579–583.

[18] R. Wang, M. Shen, Y. He, and X. Liu, "Performance of cell-free massive MIMO with joint user clustering and access point selection," *IEEE Access*, vol. 9, pp. 40 860–40 870, 2021.

[19] N. Pang, J. Zhang, C. Zhang, and X. Qin, "Parallel hierarchical subspace clustering of categorical data," *IEEE Transactions on Computers*, vol. 68, no. 4, pp. 542–555, 2018.

[20] Y. Xu, G. Yue, and S. Mao, "User grouping for massive MIMO in FDD systems: New design methods and analysis," *IEEE Access*, vol. 2, pp. 947–959, 2014.

[21] M. Saideh, Y. Alsaba, I. Dayoub, and M. Berbineau, "Joint Interference Cancellation for Multi-Carrier Modulation-Based Non-Orthogonal Multiple Access," *IEEE Communications Letters*, vol. 23, no. 11, pp. 2114–2117, 2019.

[22] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing—The Large-Scale Array Regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[23] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4611–4624, 2016.

[24] J. Nam, A. Adhikary, J.-Y. Ahn, and G. Caire, "Joint Spatial Division and Multiplexing: Opportunistic Beamforming, User Grouping and Simplified Downlink Scheduling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 876–890, 2014.

[25] R. Pereira, X. Mestre, and D. Gregoratti, "Clustering complex subspaces in large dimensions," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5712–5716.

[26] Y.-C. Wong, "Differential geometry of Grassmann manifolds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 57, no. 3, p. 589, 1967.

[27] L.-H. Lim, K. S.-W. Wong, and K. Ye, "The Grassmannian of affine subspaces," *Foundations of Computational Mathematics*, vol. 21, pp. 537–574, 2021.

[28] R. Pereira, A. A. Deshpande, C. J. Vaca-Rubio, X. Mestre, A. Zanella, D. Gregoratti, E. De Carvalho, and P. Popovski, "User clustering for rate splitting using machine learning," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 722–726.

[29] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 215–223.

[30] F. Ieva, A. M. Paganoni, and N. Tarabelloni, "Covariance-based clustering in multivariate and functional data analysis," *Journal of Machine Learning Research*, vol. 17, no. 143, pp. 1–21, 2016.

[31] H. Lee, H.-J. Ahn, K.-R. Kim, P. T. Kimd, and J.-Y. Koo, "Geodesic clustering for covariance matrices," *Communications for Statistical Applications and Methods*, vol. 22, no. 4, pp. 321–331, 2015.

[32] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[33] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2011.

[34] ——, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, 2013.

[35] X. Sun, X. Gao, G. Y. Li, and W. Han, "Agglomerative user clustering and downlink group scheduling for FDD massive MIMO systems," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[36] N. Czink, B. Bandemer, C. Oestges, T. Zemen, and A. Paulraj, "Analytical Multi-User MIMO Channel Modeling: Subspace Alignment Matters," *IEEE Transactions on Wireless Communications*, vol. 11, no. 1, pp. 367–377, 2012.

[37] Y. Thanwerdas and X. Pennec, "O(n)-invariant Riemannian metrics on SPD matrices," *Linear Algebra and its Applications*, vol. 661, pp. 163–201, 2023.

[38] I. Horev, F. Yger, and M. Sugiyama, "Geometry-aware principal component analysis for symmetric positive definite matrices," in *Asian Conference on Machine Learning*. PMLR, 2016, pp. 1–16.

[39] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2496–2503.

[40] I. L. Dryden, A. Koloydenko, and D. Zhou, "Non-Euclidean Statistics for Covariance Matrices, with Applications to Diffusion Tensor Imaging," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 1102–1123, 2009.

[41] F. Barbaresco, "Innovative tools for radar signal processing based on Cartan's geometry of SPD matrices & Information Geometry," in *2008 IEEE Radar Conference*, 2008, pp. 1–6.

[42] K. M. Wong, J.-K. Zhang, J. Liang, and H. Jiang, "Mean and median of PSD matrices on a Riemannian manifold: Application to detection of narrow-band sonar signals," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6536–6550, 2017.

[43] M. Moakher and P. G. Batchelor, "Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization," in *Visualization and Processing of Tensor Fields*. Springer, 2006, pp. 285–298.

[44] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *International conference on machine learning*, 2015, pp. 720–729.

[45] I. L. Dryden, X. Pennec, and J.-M. Peyrat, "Power Euclidean metrics for covariance matrices with application to diffusion tensor imaging," *arXiv preprint arXiv:1009.3045*, 2010.

[46] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[47] J. Zhang, G. Zhu, R. W. Heath Jr, and K. Huang, "Grassmannian learning: Embedding geometry awareness in shallow and deep learning," *arXiv preprint arXiv:1808.02229*, 2018.

[48] X. Han, G. Pan, and Q. Yang, "A unified matrix model including both CCA and F matrices in multivariate analysis: The largest eigenvalue and its applications," 2018.

[49] Z. Bao, J. Hu, G. Pan, and W. Zhou, "Canonical correlation coefficients of high-dimensional Gaussian vectors," *The Annals of Statistics*, vol. 47, no. 1, pp. 612–640, 2019.

[50] F. Yang, "Limiting distribution of the sample canonical correlation coefficients of high-dimensional random vectors," *Electronic Journal of*

Probability, vol. 27, no. none, pp. 1 – 71, 2022. [Online]. Available: https://doi.org/10.1214/22-EJP814

[51] Y. Yang and G. Pan, "The convergence of the empirical distribution of canonical correlation coefficients," *Electron. J. Probab*, vol. 17, no. 64, pp. 1–13, 2012.

[52] M. Tiomoko, R. Couillet, E. Moisan, and S. Zozor, "Improved estimation of the distance between covariance matrices," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7445–7449.

[53] M. Tiomoko and R. Couillet, "Random matrix-improved estimation of the Wasserstein distance between two centered Gaussian distributions," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[54] R. Couillet, M. Tiomoko, S. Zozor, and E. Moisan, "Random matrix-improved estimation of covariance matrix distances," *Journal of Multivariate Analysis*, vol. 174, p. 104531, 2019.

[55] M. Tiomoko, R. Couillet, F. Bouchard, and G. Ginolhac, "Random matrix improved covariance estimation for a large class of metrics," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6254–6263.

[56] X. Mestre, "On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices," *IEEE Transactions on Signal Processing*, vol. 56, no. 11, pp. 5353–5368, 2008.

[57] J. Silverstein and Z. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 175–192, 1995.

[58] Z.-D. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices," *The Annals of Probability*, vol. 26, no. 1, pp. 316–345, 1998.

[59] J. B. Conway, *Compactness and Convergence in the Space of Analytic Functions*. New York, NY: Springer New York, 1978, pp. 142–194. [Online]. Available: https://doi.org/10.1007/978-1-4612-6313-5_7

[60] M. Tiomoko and R. Couillet, "Estimation of covariance matrix distances in the high dimension low sample size regime," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 341–345.

[61] X. Mestre, "Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5113–5129, 2008.

[62] X. Mestre, F. Rubio, and P. Vallet, "Improved estimation of the logarithm of the covariance matrix," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2012, pp. 377–380.

[63] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[64] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[65] J. E. Chacón and A. I. Rastrojo, "Minimum adjusted rand index for two clusterings of a given size," *Advances in Data Analysis and Classification*, vol. 17, no. 1, pp. 125–133, 2023.

[66] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.

[67] G. Zhang, L. Jiang, P. Ji, S. Zou, C. He, and D. He, "A Modified K-means User Grouping Design for HAP Massive MIMO Systems," in *2021 International Conference on Networking and Network Applications (NaNA)*, 2021, pp. 288–292.

[68] A. Edelman, T. A. Arias, and S. T. Smith, "The Geometry of Algorithms with Orthogonality Constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, p. 303–353, Apr. 1999. [Online]. Available: https://doi.org/10.1137/S0895479895290954

[69] E. Bosse, J. Roy, and D. Grenier, "Data fusion concepts applied to a suite of dissimilar sensors," in *Proceedings of 1996 Canadian Conference on Electrical and Computer Engineering*, vol. 2, 1996, pp. 692–695 vol.2.

[70] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP journal on wireless communications and networking*, vol. 2018, no. 1, pp. 1–54, 2018.

[71] M. Sadek, A. Tarighat, and A. H. Sayed, "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1711–1721, 2007.

[72] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, "A new approach for mutual information analysis of large dimensional multi-antenna channels," *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 3987–4004, 2008.

[73] L. Pastur and M.Shcherbina, *Eigenvalue Distribution of Large Random Matrices*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2011, vol. 171.

[74] W. Hachem, P. Loubaton, X. Mestre, J. Najim, and P. Vallet, "Large information plus noise random matrix models and consistent subspace estimation in large sensor networks," *Random Matrices: Theory and Applications*, vol. 1, no. 2, Apr. 2012.

[75] D. Schenck, X. Mestre, and M. Pesavento, "Probability of resolution of MUSIC and g-MUSIC: An asymptotic approach," *IEEE Transactions on Signal Processing*, vol. 70, pp. 3566–3581, 2022.

[76] X. Mestre and P. Vallet, "On the Resolution Probability of Conditional and Unconditional Maximum Likelihood DoA Estimation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4656–4671, 2020.

[77] R. Pereira, X. Mestre, and D. Gregoratti, "Asymptotics of Distances Between Sample Covariance Matrices," *Submitted to IEEE Transactions on Signal Processing*, 2023.