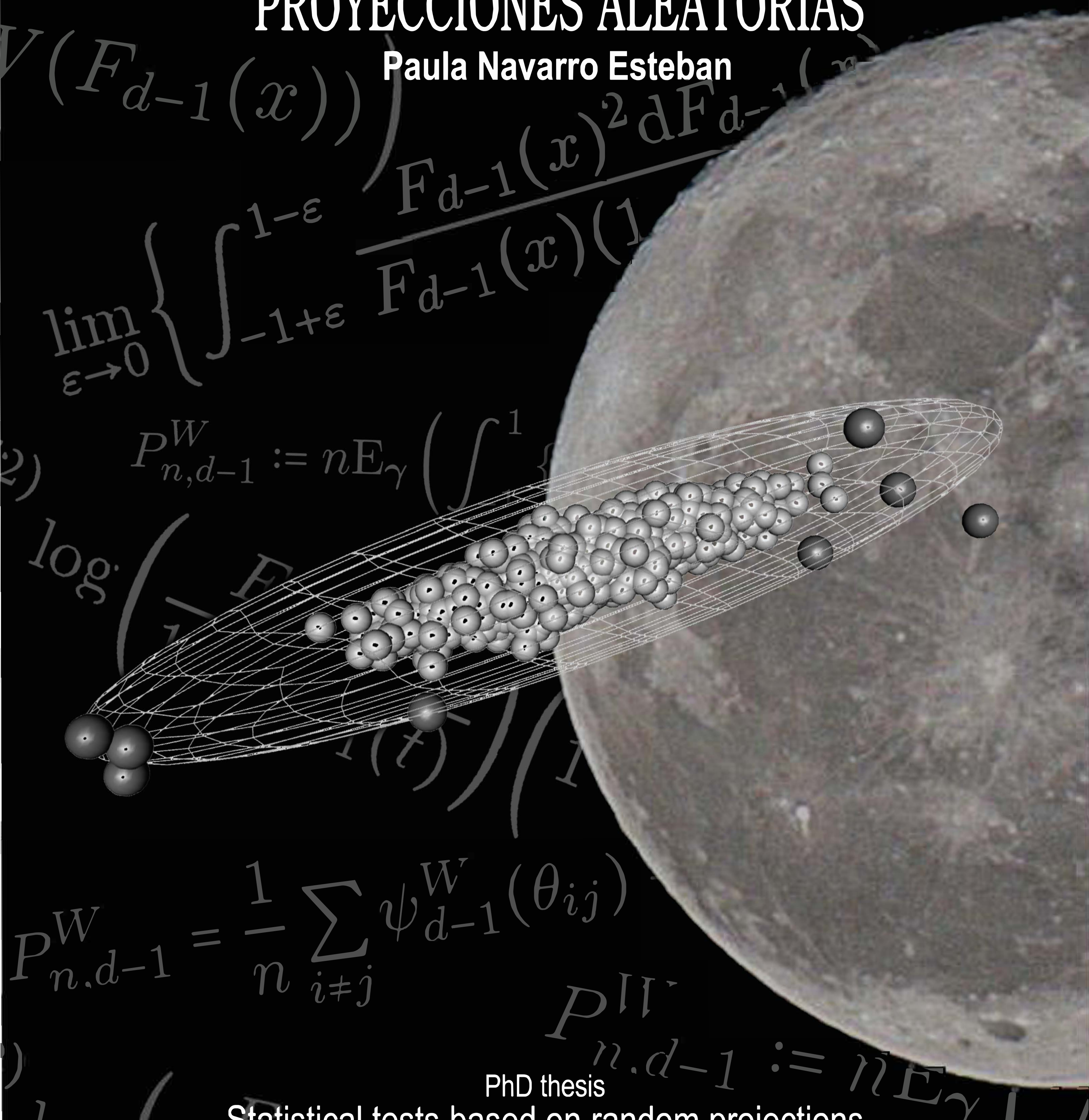
Programa de Doctorado en Ciencias y Tecnología

Tesis Doctoral

# TESTS ESTADÍSTICOS BASADOS EN PROYECCIONES ALEATORIAS

**Paula Navarro Esteban**



$\lim_{\varepsilon\to 0}\left\{\int_{-1+\varepsilon}^{1-\varepsilon}\frac{F_{d-1}(x)^2 dF_{d-1}}{F_{d-1}(x)(1}\right.$
$P^W_{n,d-1} := nE_\gamma\left(\int^1\right.$
$P^W_{n,d-1} = \frac{1}{n}\sum_{i\neq j}\psi^W_{d-1}(\theta_{ij})$
$P^{II}_{n,d-1} := nE_\gamma\left(\int\right.$

PhD thesis

Statistical tests based on random projections

Dirigida por Juan A. Cuesta Albertos y Alicia Nieto Reyes

Escuela de Doctorado de la Universidad de Cantabria
Santander 2020 - Universidad de Cantabria

# Statistical tests based on random projections

**Paula Navarro Esteban**

Supervised by: *Juan A. Cuesta Albertos* and *Alicia Nieto Reyes*

**Paula Navarro Esteban**

*Statistical tests based on random projections*

Supervised by: Juan A. Cuesta Albertos and Alicia Nieto Reyes

Universidad de Cantabria
Santander, 2020

*A mi madre y a mi padre*

# Agradecimientos

Durante el desarrollo de esta tesis [1] han sido muchas las personas que de una forma u otra me han ayudado a seguir con este trabajo hasta su culminación. A todas ellas (son demasiadas como para ponerlas todas aquí), muchas gracias.

En primer lugar me gustaría agradecer a mi director, Juan, la oportunidad de realizar la tesis con él, ha sido el mayor aprendizaje que he tenido hasta ahora. Su dedicación, su pasión y la motivación por enseñar estadística es de admirar. Agradecerle además, su ánimo cuando no salían las cuentas como esperábamos. Gracias también a mi codirectora, Alicia, por sus comentarios, sus ideas y su tiempo.

I would like to thank Professor Heather Battey for making my stay in the University of Bristol more comfortable. Thank Xiaoli for sharing such a nice time in Bristol and her amazing culture.

En el ámbito académico, debo dar las gracias a Araceli, Carlos, Luis Alberto y Nuria por ayudarme a resolver desinteresadamente las dudas que les he planteado durante este tiempo. Mi más sincero agradecimiento a Araceli y Asun por facilitarme los temidos momentos burocráticos, también a mis compañeras de pasillo y de despacho por los momentos compartidos. De mi departamento "de adopción", agradezo, a Alexandra y a Juan, la confianza y las oportunidades que me han dado, tan importantes estos últimos años.

Una mención especial merece Eduardo, porque no está escrito las horas que hemos trabajado (sigo alucinando con tu motivación constante). Lo que iba a ser un trabajo sencillito acabó siendo un currazo tremendo, del que aprendí muchísimo, eso sí. Junto con él; agradecer a Antonio, Bea, Javi, José Luis y Luis por sus ideas, su compañía, su trabajo y su apoyo en las actividades del grupo de FDA (y en el resto de congresos). También a Carmen por su confianza desde el principio y darme la oportunidad de coordinar el grupo.

Dentro del mundillo matemático, debo dar las gracias a mis compis de carrera: Adela, Juan, Luciano, Noemí y Santi, porque a pesar de la distancia han sido un apoyo moral clave todos estos años.

[1] This thesis (except the title page) uses the *Clean Thesis* style developed by Ricardo Langner. La portada está diseñada con la ayuda de Leo, a quien agradezco sus ideas y su tiempo.

En el terreno personal, no imagino mi vida sin mi familia elegida: Ali, Charlie, Chelu, Dahir, Edu, Jorge, Juan, Rana, Santur y Teja. Son los *fueguitos* que encienden mi día a día y que hacen que este mundo sea un poquito mejor. Agredecer, asimismo, a Ángela, a Marina y a Usoa la sororidad creada.

Gracias también a mi familia y a mis amigas/os de Almazán (a las/os que se han quedado a pesar de las condiciones y las/os que se han visto forzadas/os a marcharse). En especial, a mis amigas de la infancia Alejandra y Tania por todo lo vivido. Otra mención especial es para mi hermana Elena, mi sobrina Hanna y mi sobrino Izan, ¡gracias por darme los momentos más bonitos y entrañables!

Llegar hasta aquí no hubiera sido posible sin el amor y el apoyo incondicional de mi madre, Pilar, y de mi padre, José Luis. No sé cómo agradeceros todo lo que me habéis dado y enseñado sin esperar nada a cambio. Todo ha sido más fácil gracias a los dos. Os quiero mucho.

Por último me gustaría dedicar estas líneas a varias mujeres que me han servido de inspiración y de referencia: Alexandra, Amparo, Emma, Euphemia, Frida, Lola, Simone, Teresa, Voltairine, etc. Ellas junto con muchas otras (la mayoría invisibilizadas) son grandes ejemplos a seguir.

# Contents

List of Figures xiii

List of Tables xv

Resumen en español xvii

1 Introduction 1

1.1 Outlier detection . . . . . 4

1.2 Uniformity tests on the hypersphere . . . . . 6

1.3 Analysis of simulated and real datasets . . . . . 7

2 Background 9

2.1 General notation . . . . . 9

2.1.1 Special functions and orthogonal polynomials . . . . . 10

2.2 Random projections . . . . . 11

2.3 What is an outlier? . . . . . 13

2.3.1 When is a point far away from the rest? . . . . . 14

2.4 Methods to detect outliers . . . . . 16

2.4.1 Methods that estimate the covariance matrix . . . . . 16

2.4.2 Methods to handle high-dimensional cases . . . . . 17

2.5 Spherical Distributions . . . . . 20

2.5.1 Uniform distribution on the hypersphere . . . . . 21

2.6 Uniformity tests on the hypersphere . . . . . 22

2.6.1 Sobolev tests of uniformity . . . . . 22

2.6.2 A test based on random projections . . . . . 27

2.7 Required results about integral equations . . . . . 27

2.8 Sequential analysis . . . . . 28

3 Outliers detection: known parameters 31

3.1 Scenario of the method and notation . . . . . 31

3.2 The distribution of the projections $Y^{\mathbf{V}}$ . . . . . 33

3.2.1 Several projections . . . . . 41

3.2.2 The testing problem revisited . . . 42

3.3 Moments of $K$ . . . 44

3.4 Asymptotic properties of the threshold $C_n^d$ . . . 46

3.5 Computation of the constants $a$ and $b$ . . . 50

3.5.1 Computation of $(a, b)$ when $\Sigma = I_d$ . . . 52

3.5.2 Computation of $(a, b)$ when $\Sigma \neq I_d$ . . . 55

3.6 On the error of type II . . . 57

3.7 Numerical studies . . . 60

**4 Outliers detection: unknown parameters 65**

4.1 Introduction to the method and additional notation . . . 65

4.2 Some properties of the distribution of $Y_n^{\mathbf{V}}$ . . . 66

4.3 Moments of $K_n$ . . . 71

4.4 Robust versions of $K_n$ and $Y_n^{\mathbf{V}}$. Additional notation . . . 71

4.5 Computation of the constants $a$ and $b$ . . . 79

4.5.1 Computation of $(a, b)$ when $\Sigma = I_d$ . . . 81

4.5.2 Computation of $(a, b)$ when $\Sigma \neq I_d$ . . . 82

4.6 Practical implementation . . . 82

4.6.1 Which value should we choose for $l_\Sigma^r$? . . . 83

4.6.2 How many simulated values of $\tilde{Y}^{\mathbf{V}}$ are required to obtain $a$ and $b$? 83

4.6.3 Algorithm to analyse a sample . . . 83

4.6.4 How to reduce the role of the randomness in deciding if a point is outlier or not? . . . 84

4.7 Numerical studies . . . 85

4.7.1 Simulations . . . 85

4.7.2 Comparison with other procedures . . . 86

4.7.3 The procedure in practice: Two real data examples . . . 90

**5 On a class of uniformity tests on the hypersphere 95**

5.1 Projected statistics: A new approach to testing uniformity . . . 95

5.1.1 Genesis . . . 95

5.1.2 $U$-statistic form of $P_{n,d-1}^W$ . . . 98

5.1.3 Extending the Watson test . . . 103

5.1.4 Extending the Ajne and Rothman test . . . 105

5.1.5 An Anderson–Darling-like test . . . 109

5.1.6 Summary . . . 113

5.2 Connection with the Sobolev class . . . 114

5.2.1 Relation between Sobolev and projected classes . . . 114

5.2.2 Asymptotic null distributions and local optimality . . . 122

5.2.3 Computation of asymptotic tail probabilities . . . . . 131
5.3 Simulation study . . . . . 132
5.3.1 Asymptotic null distribution accuracy . . . . . 132
5.3.2 Empirical power investigation . . . . . 133
5.4 Real data applications . . . . . 138
5.4.1 Sunspots . . . . . 138
5.4.2 Long-period comets . . . . . 139
5.4.3 Craters on Rhea . . . . . 140

**6 Future work** **143**

**Bibliography** **145**

**A Appendix** **155**
A.1 Additional tables of Chapter 3 . . . . . 155
A.2 Additional tables of Chapter 4 . . . . . 159
A.3 Additional tables and further simulations of Chapter 5 . . . . . 163

# List of Figures

1.1 Representation of the set $\{\mathbf{x} : \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma = C_n^d(\delta)\}$ for $d = 3$, $n = 100$, $\delta = 0.05$ . 5

2.1 Outlier with a different distribution from the sample . . . . . . . . . . . . . . . 13
2.2 Importance of a sample size dependent definition of outlier . . . . . . . . . . . 15

3.1 Representation of $\mathbf{P}\left(\mathbf{X}'\mathbf{V}/\|\mathbf{V}\| < s \,|\|\mathbf{X}\| = C_n^d\right)$ . . . . . . . . . . . . . . . . . . 35
3.2 Plot of the curves $\mathrm{E}(K|\,\|\mathbf{X}\| = t)$ and $\mathrm{Var}(K|\,\|\mathbf{X}\| = t)^{1/2}/\mathrm{E}(K|\,\|\mathbf{X}\| = t)$ . . . 47
3.3 Values of $C_n^d(0.05)$ and of $C_n^d(0.007)$ . . . . . . . . . . . . . . . . . . . . . . . . 50
3.4 Probability of declaring a point as outlier given its Mahalanobis norm . . . . 58

4.1 Plot of wines identified as outliers . . . . . . . . . . . . . . . . . . . . . . . . . . 91
4.2 Boxplot of the squared of the differences among the components of each point in the wine data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
4.3 Particular outliers in wine data with RP and PCOut methods . . . . . . . . . 92
4.4 Particular outliers detected in octane data with RP and PCOut methods . . . 93

5.1 Depiction of $\tilde{\psi}_{d-1}^{\mathrm{CvM}}(\theta)$, $\tilde{\psi}_{d-1}^{\mathrm{R}_{1/3}}(\theta)$, and $\tilde{\psi}_{d-1}^{\mathrm{AD}}(\theta)$ . . . . . . . . . . . . . . . . . . 113
5.2 Depiction of the projected densities generated by $f^{\mathrm{CvM}}$, $f^{\mathrm{R}_{1/3}}$, and $f^{\mathrm{AD}}$ . . 135
5.3 Craters on Rhea . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 142

A.1 Difference of empirical powers with respect to the ILRT for different deviations 165
A.2 Empirical powers for different deviations . . . . . . . . . . . . . . . . . . . . . 165

# List of Tables

3.1 Values of $C_n^d$ . . . 49

3.2 Values of $(a, b)$ for $\Sigma = I_d$ . . . 55

3.3 Approximated values of $b$ for $\Sigma = \Sigma_i^d$ with $i = 1, \dots, 4$ . . . 57

3.4 Values of $b_\Sigma$ giving the greatest difference $|b_I - b_\Sigma|$ for $\Sigma = \Sigma_i^d, i = 1, \dots, 4$ . . . 57

3.5 Values of $k$ and $k^h$ for $n = 100$ . . . 60

3.6 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ for $n = 50$ . . . 61

3.7 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ for $n = 50$ . . . 61

3.8 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ for $n = 50$ . . . 62

3.9 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ for $n = 50$ . . . 62

3.10 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ for $n = 50$ when we truncate . . . 63

3.11 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ for $n = 50$ when we truncate . . . 63

4.1 Obtained values of $(a, b)$ when $\Sigma = I_d$ . . . 81

4.2 Values of $b_\Sigma$ giving the greatest difference $|b_I - b_\Sigma|$ for $\Sigma = \Sigma_i^d, i = 1, \dots, 4$ . . . 82

4.3 Computation times of $a$, $b$ with $N = 10^5$ simulated values of $\tilde{Y}^{\mathbf{V}}$ . . . 83

4.4 Obtained values of $(a, b)$ when $\Sigma = I_d$ . . . 84

4.5 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\| = C_n^d$ when $\Sigma = I_d$ . . . 84

4.6 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ for $n = 50$ . . . 86

4.7 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ with $n = 50$ . . . 87

4.8 Proportion of outliers found in a clean data set for several covariance matrices. . . . 89

4.9 Proportion of outliers correctly identified in samples containing 10% of them 90

4.10 Wines identified as outliers . . . 91

4.11 Outliers in the gasolines . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93

5.1 Extensions and connections given by the new family of projected tests. . . . 113
5.2 Exact-$n$ and asymptotic critical values of the CvM, AD, and Rt uniformity tests 133
5.3 Rejection frequencies using asymptotic critical values of the CvM, AD, and Rt uniformity tests . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133
5.4 Empirical powers for the investigated uniformity tests on $\Omega_1$ for $n = 100$ and $\kappa = 0.50$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137
5.5 Asymptotic $p$-values of the CvM, AD, and Rt tests when applied to the crater locations of the planets and moons . . . . . . . . . . . . . . . . . . . . . . . . 140
5.6 Asymptotic $p$-values of the CvM, AD, and Rt tests when applied to [78]'s Rhea crater database. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 141

A.1 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155
A.2 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . 156
A.3 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ when we truncate . . . . . . . . . . . . . . . . . . . . . . . . . 157
A.4 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ when we truncate . . . . . . . . . . . . . . . . . . . . 158
A.5 Obtained values of $(a, b)$ when $\Sigma = I_d$ . . . . . . . . . . . . . . . . . . . . . . . 159
A.6 Approximated values of $b_\Sigma$ for $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$ . . . . . . . . . . . . . 159
A.7 Approximated values of $b_\Sigma$ for $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$ (non-robust) . . . . . 160
A.8 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ with the non-robust procedure . . . . . . . . . . . . . . . . . . . . 160
A.9 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 161
A.10 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ with the non-robust procedure . . . . . . . . . . . . . 161
A.11 Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$ with $r = 1.2, 2$ . . . . . . . . . . . . . . . . . . . . . . . . . . 162
A.12 Proportion of outliers found in a clean data set for several covariance matrices. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 163
A.13 Proportion of outliers correctly identified in samples containing 10% of them163
A.14 Empirical powers for the uniformity tests on $\Omega^1$ . . . . . . . . . . . . . . . . . . 166
A.15 Empirical powers for the uniformity tests on $\Omega^2$ . . . . . . . . . . . . . . . . . . 167
A.16 Empirical powers for the uniformity tests on $\Omega^3$ . . . . . . . . . . . . . . . . . . 168
A.17 Empirical powers for the uniformity tests on $\Omega^{10}$ . . . . . . . . . . . . . . . . . 169

# Resumen en español

El Teorema de Cramér–Wold [30, p.291] establece que una probabilidad de Borel en un espacio euclídeo está determinada unívocamente por sus proyecciones unidimensionales. En otras palabras, dos distribuciones son iguales si y solo si todas sus marginales unidimensionales son iguales.

Una versión mejorada del Teorema de Cramér–Wold aparece en el Teorema de Cuesta–Fraiman–Ransford [35, p.203]. En virtud de este resultado se tiene que una sola proyección unidimensional aleatoriamente elegida es suficiente para distinguir de forma casi segura a dos distribuciones definidas en un espacio de Hilbert separable siempre que los momentos de una de ellas satisfagan cierta condición. De forma más precisa, este resultado viene a decir que, bajo la condición mencionada, dadas dos distribuciones de probabilidad, si las proyectamos en el mismo subespacio unidimensional elegido con una distribución continua, entonces se tiene que casi seguro, las dos distribuciones son diferentes/iguales si y solo si las dos proyecciones son diferentes/iguales.

El Teorema de Cramér–Wold justifica el uso de las técnicas de *Projection Pursuit* (PP) en los tests de bondad de ajuste, ya que la base de estas técnicas es proyectar los datos en una serie de direcciones unidimensionales apropiadas. En particular, la selección de estas direcciones se realiza maximizando un criterio específico de optimalidad que mide el grado de "interés" de las direcciones. Por ejemplo, en tests de igualdad de dos distribuciones, las técnicas PP pretenden buscar las direcciones donde las distribuciones son lo más diferentes posible. Como alternativa a estas direcciones (pseudo-)deterministas, el Teorema de Cuesta–Fraiman–Ransford permite considerar proyecciones aleatorias. Éstas consisten simplemente en proyectar los datos iniciales de alta dimensión en un subespacio de baja dimensión seleccionado aleatoriamente. Esta técnica se ha usado en algoritmos de diversas áreas como optimización combinatoria (Vempala [132]), recuperación de la información (Bingham y Mannila [20]), aprendizaje automático (Arriaga y Vempala [9]), reconocimiento facial (Goel *et al.* [72]), etc.

Las proyecciones aleatorias son rápidas y estables, por ejemplo ver teorema 2.2.2 en el Capítulo 2. Por tanto, se usan en aplicaciones que reducen la dimensión y que requieren de cierta eficiencia computacional y preservación de la estructura local de los datos. Evidentemente, reemplazar datos $d$-dimensionales por otros de menor dimensión conlleva una pérdida de información, sin embargo esta desventaja no es tan relevante como pudiera parecer a priori, ver por ejemplo Cuesta-Albertos *et al.* [36] donde se compara esta técnica con procedimientos PP en test de bondad de ajuste a la normalidad. Todas estas propiedades llevan a considerar en este trabajo este tipo de proyecciones. Ahora bien, ¿cómo usar las proyecciones aleatorias? Podemos dividir las soluciones que aparecen en la literatura según dos puntos de vista: se elige un estadístico apropiado para el problema considerado en el caso unidimensional y entonces

*i*) Manejamos solamente una única proyección aleatoria y calculamos el valor del estadístico.

*ii*) Calculamos el valor esperado, dada la muestra, del estadístico.

La idea es, por ejemplo, en vez de llevar a cabo un test de igualdad de medias en datos funcionales, tomar una proyección unidimensional y realizar dicho test en los datos proyectados unidimensionales. Por lo tanto, *i*) sería el resultado de usar dicho test con una única proyección, mientras que *ii*) sería el resultado de realizar la integración de los valores de un estadístico (o de unos $p$-valores) en todas las direcciones posibles. Este hecho tiene una repercusión importante en la complejidad de cálculo ya que *i*) utiliza una única dirección por lo que el procedimiento acabaría en un paso, mientras que con *ii*) habría que integrar en las direcciones. La complejidad de esta integral dependerá de la rejilla elegida que, a su vez, tiene que crecer con la dimensión. En cierta manera *ii*) aplica el Teorema de Cramér–Wold, mientras que *i*) se basa en el Teorema de Cuesta–Fraiman–Ransford.

Notar que *i*) es una cantidad aleatoria, pero *ii*) no lo es. De hecho es el valor esperado de las cantidades aleatorias de i). Además mientras que es esperable que en la mayoría de las situaciones, *ii*) proporcione mayor potencia por manejar más información, podría suceder que, por ejemplo, en un test de bondad de ajuste tuviéramos dos probabilidades diferentes pero el valor esperado de un determinado test fuera igual al obtenido bajo la nula. Con *i*) esto no puede suceder porque con probabilidad uno todas las marginales son diferentes.

A continuación analizamos varias características de estos procedimientos.

**Sobre el uso de *i*):** Como hemos señalado, el uso de una única proyección aleatoria simplifica los cálculos en comparación con la integración sobre todas las proyecciones. Sin embargo en la práctica, *i*) tiene el inconveniente de que el resultado depende de la dirección en la que se proyecta y bajo la alternativa la potencia puede ser baja. La forma más usual de solventar este problema es proyectar en varias direcciones combinando los resultados usando, por ejemplo, Bonferroni o el *False Discovery Rate*. No obstante, no existe una guía clara para elegir el número de direcciones aleatorias a utilizar, sin contar además que las técnicas de *False Discovery Rate* son conservadoras (aunque menos que las de Bonferroni).

Como solución a los problemas relacionados con la potencia, en este trabajo proponemos usar el análisis secuencial, que resuelve al mismo tiempo, los problemas de precisar el número de proyecciones necesarias y la manera de combinar los $p$-valores. Además los procedimientos secuenciales utilizan en media menos observaciones que aquellos que las fijan de antemano para lograr la misma potencia, Tartakovski *et al.* [130].

El uso de la proyecciones aleatorias como en *i*), se ha aplicado en diversos problemas tales como los tests de bondad de ajuste (Cuesta-Albertos *et al.* [34], Cuesta-Albertos *et al.* [36] and Cuesta-Albertos *et al.* [38]), el análisis de la varianza (Cuesta-Albertos y Febrero-Bande [32]), tests de linealidad en regresión funcional (Cuesta-Albertos *et al.* [39]), para la construcción de profundidades (Cuesta-Albertos y Nieto-Reyes [33]), etc.

**Sobre el uso de *ii*):** Integrar un estadístico a través de todas las direcciones, como se propone en *ii*), fue considerado por primera vez en el contexto de la regresión lineal por Escanciano [50] en su test de bondad de ajuste *Cramér–von Mises proyectado*. Este test se ha exportado entre otros contextos al de datos funcionales por García-Portugués *et al.* [66]. Este último test fue comparado en Cuesta-Albertos *et al.* [39] con otro test propuesto en *ibíd*, que sigue la filosofía de *ii*), con resultados empíricos que evidenciaron que integrar a través de todas las direcciones en ese contexto suele dar una potencia superior en la práctica, aunque con un coste computacional muy superior $O(n^3)$ frente a $O(n)$.

Debido a las características comentadas, consideramos conveniente usar *i*) para un nuevo procedimiento en dimensión alta de detección de outliers y *ii*) para proponer una novedosa clase de tests de uniformidad en hiperesferas. A continuación precisamos estos problemas y las razones que nos han llevado a elegir *i*) o *ii*).

- El procedimiento de detección de outliers en datos multivariantes gausianos (ver Capítulos 3 y 4) consiste en proyectar los datos en un subespacio unidimensional elegido aleatoriamente donde se aplica un procedimiento de detección de outliers unidimensional, similar al de Tukey pero con un umbral que depende de la dimensión inicial de los datos y del tamaño muestral. Como ya hemos comentado anteriormente, hay varias razones para usar más de una proyección y para determinar el número de proyecciones necesarias usaremos aquí el análisis secuencial.

El problema de detección de outliers en situaciones donde es factible estimar la matriz de covarianzas puede considerarse resuelto. En cambio, el interés de nuestra propuesta radica en escenarios de alta dimensión donde esta estimación no es factible. A su vez, en aquellas situaciones en las que aparecen grandes diferencias entre los autovalores de la matriz de covarianzas, un outlier situado en la dirección del valor propio asociado al mayor autovalor, va a ser outlier en un conjunto “pequeño” de direcciones, por lo que un procedimiento basado en *ii*) no lo puede detectar. Por ello, en este contexto es preferible usar *i*).

- La clase de tests de uniformidad en la hiperesfera está basada en la integral a lo largo de todas las posibles direcciones de una discrepancia cuadrática ponderada entre la función de distribución empírica de los datos proyectados y la distribución uniforme proyectada (ver Capítulo 5). Por tanto, la familia propuesta de tests depende del peso que se use en la ponderación. Se obtienen expresiones sencillas para varios tests estadísticos en el círculo y en la esfera y otras relativamente manejables en dimensiones superiores.

  En este contexto hemos usado *ii*) ya que el procedimiento está concebido para su aplicación en dimensiones bajas o moderadas y prima la potencia sobre la complejidad de cálculo.

Veremos en las siguientes secciones de forma más específica las dos aplicaciones mencionadas, las conclusiones que hemos obtenido y finalizaremos con los problemas abiertos que han surgido a lo largo de esta tesis. Los principales resultados del Capítulo 4 aparecen en Navarro-Esteban and Cuesta-Albertos [109] y los del Capítulo 5 en García-Portugués *et al.* [70] y García-Portugués *et al.* [67].

## Detección de outliers

El estudio de las observaciones anómalas o outliers ha sido de gran interés desde mitad del siglo XX. De hecho, actualmente, hay una gran variedad de libros (ver por ejemplo Barnett y Lewis [14]) y diversos paquetes de software estadístico que versan sobre este problema. Su detección es una parte importante en el preprocesamiento de los datos ya que pueden llevar a un modelo mal especificado, a una estimación sesgada de los parámetros o a resultados incorrectos en general, ver Weisberg [138].

Comúnmente se acepta que un outlier es una observación, o conjunto de ellas, que parece ser inconsistente con el resto de los datos. Sin embargo, la formalización de la idea de tal inconsistencia no es directa. Aquí, nos centraremos en comprobar si vectores

$\mathbf{x} \in \mathbb{R}^d$ son outliers con respecto a una muestra de v.a.'s (vectores aleatorios) i.i.d. (independientes e idénticamente distribuidos) $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbb{R}^d$ con distribución normal de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\Sigma$, $N_d(\boldsymbol{\mu}, \Sigma)$. Para ello, tomaremos como referencia la Definición 0.0.1, la cual se basa en el resultado bien conocido de que el cuadrado de la norma de Mahalanobis de un vector $d$-dimensional con distribución $N_d(\boldsymbol{\mu}, \Sigma)$, sigue una distribución chi-cuadrado con $d$ grados de libertad, $\chi^2_d$.

Dado $0 < \delta < 1$, denotamos $C^d_n(\delta)$ a la raíz cuadrada del $\delta$-cuantil del máximo de una muestra aleatoria de tamaño $n$ con distribución $\chi^2_d$. En otras palabras, $C^d_n(\delta)$ es la solución de la ecuación:

$$\mathbf{P}\left(\max\left\{\|\mathbf{X}_1 - \boldsymbol{\mu}\|_\Sigma, \ldots, \|\mathbf{X}_n - \boldsymbol{\mu}\|_\Sigma\right\} \geq C^d_n(\delta)\right) = \delta,$$

donde $\|\mathbf{X} - \boldsymbol{\mu}\|_\Sigma = \|\Sigma^{-1/2}\left(\mathbf{X} - \boldsymbol{\mu}\right)\|$, siendo $\|\cdot\|$ la norma euclídea y $\mathbf{X}_1, \ldots, \mathbf{X}_n$ son v.a.'s i.i.d. con distribución $N_d(\boldsymbol{\mu}, \Sigma)$.

**Definición 0.0.1.** *Sea $\mathbf{x} \in \mathbb{R}^d$ y $\delta \in (0, 1)$. Diremos que $\mathbf{x}$ es un outlier al nivel $\delta$ con respecto a muestra aleatoria simple de tamaño $n$ y distribución $N_d(\boldsymbol{\mu}, \Sigma)$, si $\|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma \geq C^d_n(\delta)$.*

Así que, dado $\mathbf{x} \in \mathbb{R}^d$, las hipótesis a contrastar son

$$\mathbf{H}_0 : \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma < C^d_n(\delta) \quad \text{vs. } \mathbf{H}_1 : \quad \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma \geq C^d_n(\delta). \tag{0.1}$$

Un inconveniente de la Definición 0.0.1 es que la bola que interviene en ella depende de la matriz de covarianzas, la cual es desconocida en la práctica. Usaremos las proyecciones aleatorias unidimensionales para evitar su estimación, puesto que cuanto mayor es la dimensión, más compleja es esta estimación, siendo imposible en dimensiones altas a menos que se imponga una estructura a $\Sigma$.

El Algoritmo 1 muestra el esquema del procedimiento que proponemos para el contraste (0.1) .

La elección de los parámetros $a$ y $b$ se discute en la Sección 4.5. Resulta que dependen del tamaño muestral, de la dimensión del subespacio y de $\Sigma$, ver (4.5) y Proposición 4.2.5. Esta dependencia será analizada a través del número esperado de proyecciones aleatorias necesarias para tomar una decisión sobre el punto que estamos clasificando. Sobre la estimación de $\hat{\nu}_\mathbf{V}$ y $\hat{\lambda}_\mathbf{V}$ empezaremos asumiendo en el Capítulo 3 que $\boldsymbol{\mu}$ y $\Sigma$ son conocidas. Posteriormente, en el Capítulo 4, los estimaremos por medio de la media y desviación típica muestrales, las cuales serán reemplazadas por la mediana y la MAD (*median absolute deviation*) muestrales para robustificar el método.

**Algoritmo 1:** Procedimiento propuesto para contrastar (0.1)

```
1) Seleccionar $a, b \in \mathbb{R}^+$, $a \leq b$
2) a) Tomar una v.a. $\mathbf{V}$ con distribución $N_d(\mathbf{0}, I_d)$
   b) Considerar $\mathbf{V} = \mathbf{V}/\|\mathbf{V}\|$
3) Proyectar $\mathbf{x}$ y la muestra en el subespacio generado por $\mathbf{V}$. Es decir, calcular $\mathbf{x}'\mathbf{V}$
   y $\mathbf{X}_1'\mathbf{V}, \ldots, \mathbf{X}_n'\mathbf{V}$
4) a) Calcular $\hat{\nu}_{\mathbf{V}}$ y $\hat{\lambda}_{\mathbf{V}}$, estimadores del centro y de la dispersión de las
      proyecciones $\mathbf{X}_1'\mathbf{V}, \ldots, \mathbf{X}_n'\mathbf{V}$
   b) Considerar $y^{\mathbf{V}} := \frac{\mathbf{x}'\mathbf{V} - \hat{\nu}_{\mathbf{V}}}{\hat{\lambda}_{\mathbf{V}}}$
5) Si $|y^{\mathbf{V}}| \in [a, b]$ volver al Paso 2), en otro caso:
   - El punto $\mathbf{x}$ es declarado como outlier si $|y^{\mathbf{V}}| > b$
   - El punto $\mathbf{x}$ es declarado como no outlier si $|y^{\mathbf{V}}| < a$
```

Hemos corroborado las propiedades obtenidas con varias simulaciones considerando distintas dimensiones $d = 5, 50, 100, 500, 1000$, tamaños muestrales $n = 50, 100, 500$ y cinco familias de matrices de covarianzas (la identidad $I_d$, tres que tienen los valores propios muy dispersos y otra que los tiene más concentrados, ver subsección 3.5.2 para más detalles). Además comparamos el método propuesto (denotado RP a partir de ahora) con otros como Filzmoser *et al.* [61], que propusieron un método basado en componentes principales (lo denotamos como PCOut), y Ro *et al.* [116] con su método del producto de la diagonal mínima (denotado por MDP). Planteamos dos situaciones para realizar dicha comparación: *s1)* tenemos una muestra sin outliers y calculamos la proporción de puntos que declaramos como outliers; *s2)* introducimos una contaminación de un 10% de outliers y analizamos la proporción de ellos que son detectados. Para estas situaciones hemos considerado tres matrices adicionales que cubren escenarios con marginales independientes, otro con correlacciones relativamente altas y un tercero con correlacciones aleatorias (ver Subsection 4.7.2 para más detalles).

El objetivo de *s1)* es comprobar si variaciones en la matriz de covarianzas y/o la dimensión afectan a la estabilidad de los procedimientos. El resultado ha sido que MDP depende mucho de $\Sigma$ (si es diagonal o no), y cuando $\Sigma \neq I_d$, la dimensión también afecta. Además, a medida que disminuye $n$, el número de puntos regulares declarados incorrectamente como outliers aumenta. En cambio PCOut y RP son más estables cuando $d$ varía. No obstante, PCOut tiende a declarar más outliers cuando $n = 50$, notándose más esta diferencia cuando $n$ aumenta. Además PCOut parece declarar menos outliers cuando la dependencia no es demasiado fuerte, mientras que ocurre lo contrario con RP.

La conclusión es que RP parece dar unos resultados más estables que los que dan MDP o PCOut.

En *s2)* MDP funciona aceptablemente cuando $\Sigma = I_d$, teniendo mejores resultados que PCOut, sin embargo este comportamiento empeora en la mayoría de los casos cuando $\Sigma$ difiere de la identidad sobre todo cuando $d$ aumenta. En general, en esta situación, podríamos decir que PCOut es el ganador cuando $\Sigma$ tiene correlacciones altas, mientras que RP sería la mejor opción en el resto de los casos. Por tanto podríamos sugerir el uso de PCOut en situaciones altamente dependientes y RP en aquellas no tan dependientes. El problema con esta recomendación es que para saber en qué situación estamos deberíamos estimar $\Sigma$.

Asimismo hemos aplicado el método a dos conjuntos de datos reales, uno con espectros de resonancia magnética de muestras de vino y otro con espectros infrarrojos de muestras de gasolina. Ambos conjuntos fueron analizados en Hubert *et al.* [80] (Hub a partir de ahora). En ambos conjuntos vimos que PCOut y RP detectan mejor que Hub los outliers de forma. Además MDP detecta solamente como outliers aquellos puntos que a simple vista, en su representación, son anómalos. RP declara también como outliers curvas que tienen peculiaridades no apreciables a simple vista o aquellas que se encuentran en el borde de la mayoría de los datos, aunque lo hace con una probabilidad baja.

Los análisis empíricos se han llevado a cabo usando el lenguaje de programación estadístico R y los códigos se pueden solicitar a la autora (paula.navarro@unican.es).

## Tests de uniformidad en la hiperesfera

En ocasiones, contrastar la uniformidad de la distribución que genera una muestra $\mathbf{X}_1, \ldots, \mathbf{X}_n$ cuyo soporte está en la hiperesfera unidad $\Omega^{d-1}$ de $\mathbb{R}^d$, con $d \geq 2$, es uno de los primeros pasos que se realizan al analizar datos multivariantes para los cuales solo las direcciones (y no las magnitudes) son de interés, son los llamados *datos direccionales*. Los datos direccionales están presentes en diferentes disciplinas tales como bioinformática de proteínas, ciencias ambientales y biología, ver Ley y Verdebout [99] para un resumen de los casos de estudio más recientes. Debido a la peculiaridad del soporte, el análisis riguroso de los datos direccionales requiere una adaptación de los métodos estadísticos clásicos, siendo Mardia y Jupp [102], Ley y Verdebout [98] y Pewsey y García-Portugués [113] monografías y recopilaciones actuales de la *estadística direccional* y sus avances. Además los tests de uniformidad en $\Omega^{d-1}$ son una herramienta importante para: (*i*) contrastar distribuciones esféricamente simétricas en $\mathbb{R}^d$ (ver, e.g., Cai *et al.* [25]); (*ii*) realizar tests de bondad de ajuste en el círculo mediante la transformación integral de probabilidad (Mardia y Jupp [102, Section 6.4.2]); (*iii*) tests de bondad de ajuste en $\Omega^{d-1}$, $d \geq 2$ mediante una transformación casi canónica, Jupp and Kume [89, Proposition 1]; (*iv*); contrastar simetría rotacional en $\Omega^{d-1}$ (ver, e.g., García-Portugués *et al.* [68]).

Desde la segunda mitad del siglo XX, se ha propuesto un gran número de tests para evaluar la uniformidad en $\Omega^{d-1}$. Estas contribuciones varían en generalidad (dimensión arbitraria vs. tests de dimensión específica; consistencia contra cualquier tipo de desviación vs. consistencia solo contra ciertas alternativas) y de la metodología subyacente (paramétricos vs. no paramétricos), un resumen actualizado puede verse en García-Portugués y Verdebout [64]. Además de su propia importancia, los tests de uniformidad en $\Omega^{d-1}$ son herramientas auxiliares cruciales, entre otros, en los siguientes problemas estadísticos: (*i*) contraste de distribuciones esféricamente simétricas en $\mathbb{R}^d$ (ver, e.g., Cai *et al.* [25]); (*ii*) tests de bondad de ajuste en el círculo $\Omega^1$ a través de la transformación integral de probabilidad, Mardia and Jupp [102, Section 6.4.2]; (*iii*) tests de bondad de ajuste en $\Omega^{d-1}$, $d \geq 2$, a través de una transformación casi-canónica, Jupp and Kume [89, Proposition 1]; (*iv*) contrastar la simetría rotacional en $\Omega^{d-1}$ (ver, e.g., García-Portugués *et al.* [68]).

Dada una muestra $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de observaciones i.i.d. de $\mathbf{X}$, si llamamos $\mathbf{P}$ a la distribución de $\mathbf{X}$ contrastar la uniformidad en $\Omega^{d-1}$ es el contraste de

$$\mathbf{H}_0 : \mathbf{P} = \nu_{d-1} \quad \text{vs.} \quad \mathbf{H}_1 : \mathbf{P} \neq \nu_{d-1}, \tag{0.2}$$

donde $\nu_{d-1}$ es la distribución de probabilidad uniforme en $\Omega^{d-1}$.

Se propone el test que rechaza $\mathbf{H}_0$ de (0.2) para valores grandes del estadístico

$$P^W_{n,d-1} := n\mathrm{E}_\gamma \left( \int_{-1}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \,\mathrm{d}W(F_{d-1}(x)) \right), \tag{0.3}$$

donde $W$ una medida positiva $\sigma$-finita de Borel en $[0,1]$ y $F_{d-1}$ y $F_{n,\gamma}$ son, respectivamente, la función de distribución bajo la nula y la función de distribución empírica de la muestra proyectada en la dirección aleatoria $\gamma$, cuya distribución en $\Omega^{d-1}$ es independiente de la muestra.

Notar que este estadístico sigue la línea *i*) puesto que es la *esperanza* con respecto a las direcciones $\gamma$, que siguen una distribución $\nu_{d-1}$, de la conocida norma cuadrática ponderada de Anderson y Darling [7]. Para el peso $w : [0,1] \to \mathbb{R}^+$,

$$Q^w_{n,d-1,\gamma} := n \int_{-1}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, w(F_{d-1}(x)) \,\mathrm{d}F_{d-1}(x),$$

donde los casos particulares de $w \equiv 1$ y $w(u) = 1/(u(1-u))$ dan los estadísticos de Cramér–von Mises (CvM) y Anderson–Darling (AD), respectivamente.

En otras palabras, en vez de tomar varias direcciones aleatorias y agregralas después como en Cuesta-Albertos *et al.* [37] (ver Subsección 2.6.2 para más detalles), nuestro estadístico aúna la información de todas las direcciones de $\Omega^{d-1}$ a través de la esperanza de $Q^{w}_{n,d-1,\gamma}$ con respecto a $\gamma$. Además hemos reemplazado el peso $w$ por la integración con respecto de una medida $W$.

A pesar de tener diferente procedencia, demostramos que la clase propuesta está muy relacionada con la conocida clase de tests de uniformidad de Sobolev. De hecho, probamos que los estadísticos de ambas clases tienen forma de $U$-estadístico con kernel actuando en los ángulos entre pares de puntos de la muestra. Nuestra nueva parametrización va a ser ventajosa al permitir derivar nuevos tests para datos hiperesféricos que extienden claramente los tests circulares de Watson [135], Ajne [4] y Rothman [117], e introducir por primera vez un test de tipo AD para tales datos. Se obtienen las distribuciones asintóticas y la optimalidad local frente a determinadas alternativas de los nuevos tests.

Hemos comparado el funcionamiento empírico de los tests propuestos a través de simulaciones considerando distintas alternativas, dimensiones $d = 2, 3, 4, 11$ y tamaños muestrales $n = 50, 100, 200$. Entre las conclusiones que hemos obtenido resulta que el test AD presenta un comportamiento destacable frente a las alternativas unimodales y una notable robustez frente a las alternativas no unimodales.

Hemos ilustrado la relevancia práctica de los tests propuestos con el análisis de tres conjuntos de datos reales vinculados a la astronomía. Esta disciplina es una fuente natural de datos esféricos y del estudio de la uniformidad de éstos, se puede obtener, por ejemplo, información sobre el origen de los cuerpos celestes. Los dos primeros conjuntos de datos considerados ya han sido analizados por otros autores y versan sobre las manchas solares y los cometas de periodos largos. En general, nuestros resultados confirman los análisis anteriores. Acabamos con el análisis de la distribución de los cráteres de Rea, una de las lunas de Saturno. Este conjunto no había sido analizado con anterioridad desde esta perspectiva

Los análisis empíricos se han llevado a cabo usando el lenguaje de programación estadístico R y son reproducibles en la librería `sphunif`, García-Portugués and Verdebout [65].

## Trabajo futuro

Esta tesis deja varios problemas abiertos, algunos de ellos planteados desde los inicios y otros que han ido surgiendo durante su desarrollo. A continuación exponemos algunos de los problemas que permanecen abiertos para su futuro estudio. Dividimos dichos problemas según el tema que tratan:

- Detección de outliers:

  a) En esta investigación nos hemos centrado en la detección de outliers en alta dimensión (ver Capítulos 3 and 4), sin embargo queda pendiente el salto a dimensión infinita, lo cual parece plenamente factible.

  b) Sería de interés también, la extensión a otras familias de distribuciones (como las familias elípticas generales) e incluso al caso no paramétrico. En estos casos el procedimiento unidimensional podría estar, por ejemplo, basado en *bootstrap* o en *kernel density functions*.

- Otras aplicaciones del análisis secuencial:

  a) Como ya comentamos anteriormente, el análisis secuencial puede ser aplicado a otros tipos de problemas, pudiéndose elaborar una teoría general que pueda incluir tests variados (incluso en el caso funcional). Los problemas deberían presentar algún tipo de invarianza por linealidad como son los analizados en [32] a [39] (excluido Cuesta-Albertos *et al.* [35]).

- Tests de uniformidad en la hiperesfera:

  a) Una investigación alternativa e inmediata surgiría al proceder *à la* Escanciano [50] y reemplazar $\nu_{d-1}$ por $F_{n,\gamma}$ en (0.3). Esta aproximación reemplazaría la compleja integración analítica en $\Omega^{d-1}$ por una suma de $n$ términos, no obstante tendría una conexión menos explícita con los tests de dimensión específica. Además si suponemos que se hacen los cálculos con las expresiones exactas de los estadísticos y fijamos $d$, se pasaría de un test con complejidad $O(n^2)$ a uno con $O(n^3)$. Sin embargo, si admitimos que se pueden aproximar las expresiones de los estadísticos, las complejidades dependerán de cómo se realicen estas aproximaciones.

  b) Otra alternativa sería reemplazar la norma CvM en (0.3) por el "3-point CvM statistic" de Feltz y Goldin [58].

  c) Hemos visto que ciertos resultados como la Proposición 5.1.5 nos ayudan a definir nuevos tests de uniformidad basados en diferentes elecciones de las medidas $W$. Por ejemplo, la medida $W_{a,b}(x) := \mathrm{I}_x(a,b)$, donde $\mathrm{I}_x(a,b)$ es la función beta incompleta regularizada, genera una familia flexible de dos parámetros de tests de uniformidad en $\Omega^{d-1}$, aunque con expresiones engorrosas para las funciones involucradas.

Este es solamente un ejemplo concreto de $W$ entre muchos otros que podríamos considerar para construir $P^W_{n,d-1}$. Además, se podría estudiar también la extensión a dimensiones más altas de otros tets conocidos.

d) Sería posible, aunque complicado, introducir tests de bondad de ajuste para distribuciones no uniformes en $\Omega^{d-1}$ reemplazando $F_{d-1}$ en (0.3) por la distribución apropiada.

e) Bakshaev [11] propuso un test de uniformidad para el que solo obtuvo la distribución asintótica en dimensiones $2$ y $3$ (y de forma no explícita). El hecho de que este test pertenezca a la familia introducida aquí (ver nota 5.1.8) hace que nos planteemos la posibilidad de obtener una expresión explícita para esta distribución en cualquier dimensión.

# 1 Introduction

The Cramér–Wold Theorem [30, p.291] asserts that a Borel probability measure on an Euclidean space is uniquely determined by all its one-dimensional projections. In other words, two distributions are equal if and only if all their one-dimensional marginals are equal.

This result justifies the use of Projection Pursuit (PP) techniques in goodness of fit tests since the core of those techniques is to project the data onto one-dimensional appropriate directions. In particular, the selection of the directions is done by maximizing a certain optimality criteria that measures the degree of "interestingness" of the directions. In this way, in the case of checking the equality of two distributions, PP techniques help to search the directions where the two distributions are as different as possible. As alternative to these (pseudo-)deterministic directions, random projections will be considered in this thesis. Random projections simply consist in projecting the original high-dimensional data into a low-dimensional randomly chosen subspace. This technique has been utilized in numerous algorithms of different areas such as combinatorial optimization (Vempala [132]), information retrieval (Bingham and Mannila [20]), machine learning (Arriaga and Vempala [9]), and face recognition (Goel *et al.* [72]).

Random projections run quite fast and are stable, see for instance, Theorem 2.2.2 in Chapter 2. Hence they are used in dimension reduction applications that need computational efficiency and to preserve the local structure of the data. Obviously, replacing data by lower dimensional ones implies a loss of information, however, this disadvantage is not as relevant as it could seem at a first view, see for instance Cuesta-Albertos *et al.* [36], where this technique is compared with PP procedures in normality goodness of fit tests. This makes random projections a useful dimension reduction technique.

A sharp form of the Cramér–Wold Theorem was given in the Cuesta–Fraiman–Ransford Theorem [35, p.203]. From that, it is known that a.s. just a one-dimensional random projection is enough to distinguish between two distributions defined on a separable Hilbert space if one of them satisfies a certain condition on their moments: if two distributions are given, the above mentioned condition is satisfied, we choose a one dimensional subspace using a continuous distribution and we compute the marginals of those distributions on this subspace, then we have that a.s., the two distributions are different/equal if and only if the two marginals are different/equal. From these results a question arises, how to manage random projections? In the literature some applications have been proposed following two points of view: choose a statistic suitable for your problem in the one-dimensional case and then

*i*) Take just a random direction and compute the value of the statistic.

*ii*) Compute the expected value, given the sample, of this statistic.

The idea is, for instance, instead of carrying out a test of equality of means of functional data, to take just a one-dimensional projection and to conduct the test of equality of means on the projected one-dimensional data. Then, *i*) is the result of an only handled test, whereas *ii*) is the result of integrating the values of the statistic along all the directions. In addition, *i*) is random, whereas *ii*) does not it. In fact, *ii*) is the expected value of *i*). This fact has a significant impact on the computational complexity since *i*) uses an only direction, thus, the procedure ends in one step, while with *ii*), it would be necessary to integrate in the directions. The complexity of this integral will depend on the chosen grid which, in turn, has to grow with the dimension. Somehow *i*) is based on the Cuesta–Fraiman–Ransford Theorem while *ii*) applies the Cramér–Wold Theorem.

The practical use of *i*) has a main drawback: the procedures based on this technique are conditional given the chosen random direction and the power under the alternative is low. The most common way to alleviate this problem is to choose several projections and, then, combine the obtained results making use, for instance, of the Bonferroni or the False Discovery Rate techniques. However, no clear guidance has been given on the right number of random directions to be selected and, additionally, the use of the False Discovery Rate is slightly conservative (but less than Bonferroni's). On the other hand, the advantage of *i*) is its computational speed which, contrary to *ii*), is not too affected by the dimension.

To solve the power problems of *i*), we propose the use of sequential analysis. This choice is due to the fact that it needs on average smaller sample sizes than fixed sample size procedures to achieve the same power (Tartakovski *et al.* [130]). Therefore, the use of sequential analysis resolves, at the same time, the problems to fix the number of required projections in a data-driven way and the manner to combine the obtained $p$-values.

This use of random projections, as in *i*), has been applied in goodness of fit (Cuesta-Albertos *et al.* [34], Cuesta-Albertos *et al.* [36] and Cuesta-Albertos *et al.* [38]), analysis of variance (Cuesta-Albertos and Febrero-Bande [32]), testing linearity in functional regression (Cuesta-Albertos *et al.* [39]), constructing depths (Cuesta-Albertos and Nieto-Reyes [33]), etc.

Integrating the values of a statistic along all the unit-norm directions, as *ii*) proposes, was firstly considered, within the regression context, by Escanciano [50] in his *Projected Cramér–von Mises* goodness-of-fit test. His test has been exported to, among other settings, the functional data context in García-Portugués *et al.* [66]. Relevant to the consideration of this point of view, the test in *ibid* was compared in Cuesta-Albertos *et al.* [39] against a proposal based on the *ii*) paradigm of Cuesta-Albertos *et al.* [37], with empirical results evidencing that integrating along all the directions within the test tends to provide superior power in practice, in spite of a much slower computation.

In this thesis we analyse two problems using random projections. We propose:

- A new procedure to detect outliers in Gaussian high-dimensional data (see Chapters 3 and 4). It consists in projecting the data in a one-dimensional randomly chosen subspace where an appropriate univariate outlier detection method is applied. The used unidimensional method is similar to Tukey's method but with a threshold depending on the initial dimension and the sample size. As previously stated, there are several reasons to use more than one projection and here the required number of projections is determined using sequential analysis.

  The task of detecting outliers in circumstances where the estimation of the covariance matrix is feasible can be considered solved. However, the interest of our proposal resides in high-dimensional settings where such estimation is not viable. In turn, in those situations in which big differences appear among the eigenvalues of the covariance matrix, an outlier located in the direction of the eigenvalue associated with the highest eigenvalue will be an outlier in a “ small ” set of directions. Thus, a procedure based on *ii*) cannot detect it and therefore, in this context it is preferable to use *i*).

- A novel projection-based class of uniformity tests on the hypersphere integrating along all possible directions a weighted quadratic discrepancy between the empirical cumulative distribution function (cdf) of the projected data and the cdf of the projected uniform distribution (see Chapter 5). Thus, we propose a full family of tests depending on the used weight. In addition, simple expressions for several test statistics are obtained for the circle and sphere, and relatively tractable forms for higher dimensions.

  We use here the philosophy of *ii*) because those tests are intended to be used in low or moderate dimensional spaces.

We next explain the two problems above referred in more detail. The main results of Chapter 4 appear in Navarro-Esteban and Cuesta-Albertos [109] and those of Chapter 5 in García-Portugués *et al.* [70] and García-Portugués *et al.* [67].

## 1.1 Outlier detection

Outlying, or unusual, observations have been of enormous interest since the second half of the twentieth century. Indeed, outlier detection methods are nowadays widely discussed in a variety of textbooks on Statistics (see Barnett and Lewis [14], for instance) and implemented in statistical software packages. Their detection is an important part of the preprocessing of the data because they may lead in the subsequent analysis to model misspecification, biased parameter estimation and incorrect results in general, see Weisberg [138] for instance. Typically, once suspicious observations have been flagged, the action to be taken remains the personal decision of the analyst. Mostly those observations are deleted, but this procedure is not always correct. A curious example of wrong use of the systematic elimination of outliers is the discovery of the ozone hole. In 1985, three scientists were shocked when they realized that the levels of the Antarctic ozone had decreased more than the usual, Farman *et al.* [53]. However, these low measures had been registered for several years by then; as they were so low, the measuring satellites had eliminated them because they had been programmed automatically to flag ozone losses of this magnitude as measurement failures and to delete them.

It is universally accepted that outliers (abnormalities, anomalies or irregularities) are observations which appear to be inconsistent with the rest of the data. However, the idea of such inconsistency is not straightforward to formalize. We provide some formalizations of this concept in Chapter 2 which justify our choice of Definition 1.1.1. This is based on the well known fact that if a $d$-dimensional rv (random vector) has normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma$, $N_d(\boldsymbol{\mu}, \Sigma)$, then the square of its $\Sigma$-based Mahalanobis distance to $\boldsymbol{\mu}$ follows a chi-squared distribution with $d$ degrees of freedom, $\chi^2_d$.

Given $0 < \delta < 1$, denote by $C^d_n(\delta)$ the square root of the $\delta$-quantile of the maximum of a random sample with size $n$ and distribution $\chi^2_d$, i.e. $C^d_n(\delta)$ is the solution of the equation:

$$\mathbf{P}\left(\max\left\{\|\mathbf{X}_1 - \boldsymbol{\mu}\|_\Sigma, \dots, \|\mathbf{X}_n - \boldsymbol{\mu}\|_\Sigma\right\} \geq C^d_n(\delta)\right) = \delta, \tag{1.1}$$

where $\|\mathbf{X} - \boldsymbol{\mu}\|_\Sigma = \|\Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})\|$, with $\|\cdot\|$ being the Euclidean norm and $\mathbf{X}_1, \dots, \mathbf{X}_n$ are iid (independent and identically distributed) rv's with distribution $N_d(\boldsymbol{\mu}, \Sigma)$. Thus, $C^d_n(\delta)$ is the square root of $F^{-1}_{\chi^2_d}\left((1-\delta)^{1/n}\right)$, where $F^{-1}_{\chi^2_d}$ is the quantile function of the distribution $\chi^2_d$. To ease the notation, we omit $\delta$ in $C^d_n(\delta)$ when its value is clear from the context or its exact value is irrelevant.

**Definition 1.1.1.** *Let* $\mathbf{x} \in \mathbb{R}^d$ *and* $\delta \in (0,1)$. *We say that* $\mathbf{x}$ *is an outlier at the level* $\delta$ *with respect to a simple random sample with size* $n$ *and a distribution* $N_d(\boldsymbol{\mu}, \Sigma)$*, if* $\|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma \geq C_n^d(\delta)$.

Definition 1.1.1 is easily modified to cover dependent data or other elliptical non-gaussian distributions. The only difference in the dependent case will be the expression for $C_n^d(\delta)$ which will be more complex.

Therefore, in this work we focus on testing outlyingness of some vectors $\mathbf{x}$ in $\mathbb{R}^d$ with respect to a sample of iid rv's $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in $\mathbb{R}^d$ with normal distribution $N_d(\boldsymbol{\mu}, \Sigma)$. Thus, the hypotheses to be tested are

$$\mathbf{H}_0 : \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma < C_n^d(\delta) \quad \text{vs.} \quad \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma \geq C_n^d(\delta). \tag{1.2}$$

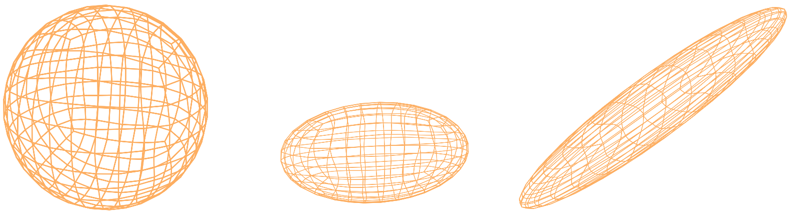


Figure 1.1.: Representation of the set $\{\mathbf{x} : \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma = C_n^d(\delta)\}$ for $d = 3$, $n = 100$, $\delta = 0.05$ and different covariance matrices (from left to right): the identity, $\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 10 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{pmatrix}$.

A drawback of this definition is that the ball depends on the covariance matrix, see Figure 1.1, which is unknown in practice. We use one-dimensional random projections (directions) to avoid its estimation because the higher the dimension, the more complex the estimation of the matrix, becoming impossible in high dimensions unless a structure in $\Sigma$ is imposed. Algorithm 2 gives the sketch of the proposed procedure to test (1.2).

**Remark 1.1.2.** *The standardization in Step 4.b) makes irrelevant the norm of the selected* $\mathbf{V}$. *This fact allows us to assume without loss of generality, when appropriated, that* $\|\mathbf{V}\| = 1$ *in Step 2.a).*

The choice of parameters $a$ and $b$ is discussed in Section 4.5. It turns out that they depend on the sample size, on the dimension of the space and on $\Sigma$, see (4.5) and Proposition 4.2.5. This dependency will be analysed through the expected number of required projections to reach the decision about the point we are classifying. Concerning the estimation of $\hat{\nu}_\mathbf{V}$ and $\hat{\lambda}_\mathbf{V}$ we begin assuming that $\boldsymbol{\mu}$ and $\Sigma$ are known in Chapter 3. Then, we employ the sample mean and the sample standard deviation which will

**Algorithm 2:** Proposed procedure to test if a point $\mathbf{x}$ is an outlier or not

```
1) Select $a, b \in \mathbb{R}^+$, $a \leq b$
2)   a) Take a rv $\mathbf{V}$ with $N_d(\mathbf{0}, I_d)$ distribution
     b) Make $\mathbf{V} = \frac{\mathbf{V}}{\|\mathbf{V}\|}$
3) Project $\mathbf{x}$ and the sample on the subspace generated by $\mathbf{V}$, i.e. compute $\mathbf{x}'\mathbf{V}$ and
   $\mathbf{X}_1'\mathbf{V}, \ldots, \mathbf{X}_n'\mathbf{V}$
4)   a) Calculate $\hat{\nu}_{\mathbf{V}}$ and $\hat{\lambda}_{\mathbf{V}}$ estimators of the centre and of the dispersion of the
        projections $\mathbf{X}_1'\mathbf{V}, \ldots, \mathbf{X}_n'\mathbf{V}$
     b) Consider $y^{\mathbf{V}} := \frac{\mathbf{x}'\mathbf{V} - \hat{\nu}_{\mathbf{V}}}{\hat{\lambda}_{\mathbf{V}}}$
5) If $|y^{\mathbf{V}}| \in [a, b]$ go back to Step 2), else:
     - The point $\mathbf{x}$ is declared as an outlier if $|y^{\mathbf{V}}| > b$
     - The point $\mathbf{x}$ is declared as non-outlier if $|y^{\mathbf{V}}| < a$
```

be replaced by the sample median and the sample median absolute deviation, MAD, respectively, to make them more robust in Chapter 4.

## 1.2 Uniformity tests on the hypersphere

Testing the uniformity of the distribution generating a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ supported on the unit hypersphere $\Omega^{d-1}$ of $\mathbb{R}^d$, with $d \geq 2$, is one of the first steps when analysing multivariate data for which only the directions (and not the magnitudes) are of interest – the so-called *directional data*. Directional data arise in many applied disciplines, such as protein bioinformatics, environmental science, and biology; we refer to Ley and Verdebout [99] and references therein for an overview of recent applications and case studies. Due to the peculiarity of the support, a rigorous analysis of directional data requires from the consideration of adapted statistical methods, with Mardia and Jupp [102], Ley and Verdebout [98] and Pewsey and García-Portugués [113] being the current reference monographs and reviews on *directional statistics*.

A sizeable number of tests for assessing uniformity on $\Omega^{d-1}$ have been proposed. These contributions range notably in generality (arbitrary dimension vs. specific-dimension tests; consistency against all kind of deviations vs. consistency only against certain alternatives) and underlying methodology (parametric vs. nonparametric tests); see

García-Portugués and Verdebout [64] for an updated review. In addition to its self-importance, uniformity tests on $\Omega^{d-1}$ are important auxiliary tools for, among others, the following statistical problems: (*i*) testing for spherically-symmetric distributions on $\mathbb{R}^d$ (see, e.g., Cai *et al.* [25]); (*ii*) goodness-of-fit tests on the circle $\Omega^1$ via the probability integral transform, Mardia and Jupp [102, Section 6.4.2]; (*iii*) goodness-of-fit tests on $\Omega^{d-1}$, $d \geq 2$, via an almost-canonical transformation, Jupp and Kume [89, Proposition 1]; (*iv*) testing for rotational symmetry on $\Omega^{d-1}$ (see, e.g., García-Portugués *et al.* [68]).

Given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of iid observations of $\mathbf{X}$, where $\mathbf{P}$ is the distribution of $\mathbf{X}$, testing uniformity on $\Omega^{d-1}$ is formalized as the testing of

$$\mathbf{H}_0 : \mathbf{P} = \nu_{d-1} \quad \text{vs.} \quad \mathbf{H}_1 : \mathbf{P} \neq \nu_{d-1}, \tag{1.3}$$

where $\nu_{d-1}$ stands for the uniform probability distribution on $\Omega^{d-1}$.

As we mentioned, we develop a projection-based class of uniformity tests on the hypersphere to test (1.3) in Chapter 5. Despite its different origin, the proposed class is shown to be related to the well-studied Sobolev class of uniformity tests. We will see that, in virtue of (2.13) and (5.17), both types of statistics have a $U$-statistic form with kernels acting on the angles between pairs of points in the sample. Our new parametrization proves itself advantageous by allowing to derive new tests for hyperspherical data that neatly extend the circular tests by Watson [135], Ajne [4], and Rothman [117], and by introducing the first instance of an Anderson–Darling-like test for such data. The asymptotic distributions and the local optimality against certain alternatives of the new tests are obtained.

## 1.3 Analysis of simulated and real datasets

Several analysis of simulated and real datasets have been carried out for all the methods proposed throughout this work. Specifically, the chosen real datasets are:

- For the novel outlier detection method, two well-known datasets: the wine (magnetic resonance spectra of some wine samples) and octane (infrared spectra of some gasoline samples) datasets.

- For the proposed uniformity tests, three datasets coming from astronomy: the first two ones build on previous applications in the circle and the sphere, while the third is a novel case study on the uniform distribution of the craters of Rhea, a moon of Saturn.

We use the statistical programming language R and all the codes can be provided on demand (paula.navarro@unican.es). The end-to-end reproduction of the analysis of the three real datasets in Chapter 5 is possible trough the `sphunif`, package García-Portugués and Verdebout [65]. The results concerning to the outlier detection method appear in Navarro-Esteban and Cuesta-Albertos [109] and those related with the proposed uniformity tests in García-Portugués *et al.* [67].

# 2 Background

> *For most of history, Anonymous was a woman.*
>
> — **Virginia Woolf**

In order to make this work self-contained, this chapter presents some notation, non-original definitions and results that will be relevant to the following chapters. It is organised as follows. Section 2.1 is devoted to the notation of special functions which will be used later, Section 2.2 shows the main properties of random projections, Section 2.3 presents different outlier definitions, Section 2.4 exposes some existing outlier detection methods and Section 2.5, several properties of the spherical distributions. Section 2.6 briefly introduces the uniformity tests on the hypersphere showing some particular cases, in particular those belonging to the class of Sobolev tests, which will be relevant for Chapter 5. Section 2.7 summarizes relevant results from the theory of integral equations that are required in the proof of Theorem 5.2.2. We conclude this chapter with an introduction to sequential analysis which is the key of the proposed outlier detection method. The reader can skip those sections if she/he is familiar with.

## 2.1 General notation

Apart from the notation employed in the Introduction, we will adopt the common convention of using boldface letters for vectors, while regular font is used for both matrices and scalars. Capital letters are used for rv's and matrices, with the context ensuring no ambiguity. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the Euclidean norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|$ and $\mathbf{x}'$ is its transpose. For a square matrix $T = (t_{ij})$, $T'$ denotes its transpose and $|T|$ its determinant. The identity matrix of dimension $d$ is denoted by $I_d$.

We refer by $\Omega_\Sigma^{d-1}(r)$ to the Mahalanobis hypersphere of radius $r$ associated to the positive definite matrix $\Sigma$. Thus $\Omega_\Sigma^{d-1}(r) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_\Sigma = r\}$. In an abuse of notation, $\Omega_r^{d-1}$ denotes such a sphere when $\Sigma = I_d$, $\omega_t^{d-1}$ is its surface area, $\omega_r^{d-1} = 2\pi^{d/2} r^d / \Gamma(d/2)$, and we omit the sub-index $r$ when $r = 1$.

We assume that all the rv's are defined on the same, rich enough, probability space $(\Upsilon, \mathcal{A}, \mathbf{P})$. Given two rv's $\mathbf{X}$ and $\mathbf{Y}$, $\mathrm{E}(\mathbf{X}|\mathbf{Y})$ will denote the conditional expectation of $\mathbf{X}$ given $\mathbf{Y}$ and $\mathbf{P}(\mathbf{X}|\mathbf{Y})$ will be a regular conditional distribution for $\mathbf{X}$ given $\mathbf{Y}$.

By $\mathbf{X} \overset{d}{=} \mathbf{P}$ we mean that $\mathbf{P}$ is the distribution of the rv $\mathbf{X}$. By pdf we mean probability density function and $1_A$ denotes the indicator function of the set $A$. In an abuse of notation, the pdf of $\nu_{d-1}$, the uniform probability on $\Omega^{d-1}$, with respect to $\omega^{d-1}$, the Lebesgue measure on $\Omega^{d-1}$, is denoted by $1/\omega^{d-1}(\Omega^{d-1})$. The symbol $\overset{d}{\rightsquigarrow}$ means convergence in distribution.

We use $m(\mathbf{P})$ and $M^*(\mathbf{P})$ to denote the median and the median absolute deviation (MAD) of a distribution $\mathbf{P}$ and, $\hat{m}(\mathbf{P})$ and $\hat{M}^*(\mathbf{P})$ to denote their sample versions; we omit the distribution when they are clear from the context.

In some integrals, given a vector $\mathbf{z} = (z_1, \ldots, z_d)$ we denote $\mathbf{z}_{-i} := z_1 \cdots z_{i-1} z_{i+1} \ldots z_d$ and $\mathrm{d}\mathbf{z}_{-i} := \mathrm{d}z_1 \cdots \mathrm{d}z_{i-1} dz_{i+1} \ldots \mathrm{d}z_d$. The positive part of a function $f$ will be denoted by $(f)_+$. Given two real functions of real variables $f(t)$ and $g(t)$, the symbol $f \sim g$ means that $\lim_{t\to\infty} \frac{f(t)}{g(t)} = 1$.

### 2.1.1 Special functions and orthogonal polynomials

Let $a, b$ be strictly positive real numbers and $z \in [0,1]$. Let us define

$$\mathrm{I}_z(a,b) := \frac{\mathrm{B}(z;a,b)}{\mathrm{B}(a,b)},$$

where $\mathrm{B}(z;a,b) := \displaystyle\int_0^z y^{a-1}(1-y)^{b-1}\,\mathrm{d}y$ is the lower incomplete function with parameters $a$ and $b$ and $\mathrm{B}(a,b)$ is the beta function.

We also denote $\mathrm{B}^{-1}(y;a,b)$ the inverse of the lower incomplete beta function. Elementary properties of those functions are $\mathrm{B}(a,b) = \mathrm{B}(b,a)$ and $\mathrm{I}_z(a,b) = 1 - \mathrm{I}_{1-z}(b,a)$. Given $a, x > 0$, we denote

- The incomplete lower gamma function with parameter $a$ as $\gamma(a,x) := \int_0^x y^{a-1} e^{-y}\, dy$.

- The regularized lower gamma function with parameter $a$ as $\mathcal{P}(a,x) := \gamma(a,x)/\Gamma(a)$.

Let $x$ be real, the error function is defined as $\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2}\,\mathrm{d}y$.

Let $k \in \mathbf{N}$, $\alpha > -1/2$ and $\alpha \neq 0$, for $x \in [-1,1]$, we denote the $k$th Gegenbauer polynomial of order $\alpha$ as $C_k^\alpha(x) := \sum_{j=0}^{\lfloor k/2 \rfloor} (-1)^j \frac{\Gamma(\alpha+k-j)}{j!(k-2j)!\Gamma(\alpha)} (2x)^{k-2j}$ and the $k$th Chebyshev polynomial of the first kind as $T_k(x) := \cos(k \cos^{-1} x)$.

Given $d \geq 3$, the Gegenbauer polynomials $C_k^{d/2-1}$ form an orthogonal basis on $(L_{d-1}^2[-1,1], \upsilon_{d-1})$, the space of square-integrable real functions in $[-1,1]$, with respect to the weight $\upsilon_{d-1} := (1-x^2)^{(d-3)/2}$. Therefore, they satisfy, for $k \geq 0$,

$$\int_{-1}^{1} C_k^{d/2-1}(x) C_\ell^{d/2-1}(x)(1-x^2)^{(d-3)/2}\,\mathrm{d}x = \delta_{k\ell} c_{k,d-1},$$

where $c_{k,d-1} := 2^{2-d}\pi\Gamma(d+k-2)/\left((d+2k-2)k!\Gamma(d/2-1)^2\right)$, and therefore any function $g \in (L_{d-1}^2[-1,1], \upsilon_{d-1})$ can be uniquely expressed into the basis of Gegenbauer polynomials of order $d/2-1$ as

$$g(x) = \sum_{k=0}^{\infty} b_{k,d-1} C_k^{d/2-1}(x), \text{ where } b_{k,d-1} = \frac{1}{c_{k,d-1}} \int_{-1}^{1} g(x) C_k^{d/2-1}(z)(1-x^2)^{(d-3)/2}\,\mathrm{d}x.$$

For $d = 2$, the Chebyshev polynomials of the first kind form an orthonormal basis on $(L_1^2[-1,1], \upsilon_1)$ with respect to the weight $\upsilon_1 := (1-x^2)^{-1/2}$, with normalizing constants given by $c_{k,1} = (1+\delta_{k0})\pi/2$ for $k \geq 0$.

More specialized notation is introduced in context.

## 2.2 Random projections

In this section we show the most relevant properties of random projections which are the foundations of the proposed outlier detection method, Chapter 4, and the uniformity test on the hypersphere, Chapter 5.

Johnson and Lindenstrauss' Lemma, [85], is the basis of the feasibility of random projections. Their most useful property for us is due to Theorem 3.1 in Cuesta-Albertos *et al.* [35]. Next we detail those results.

**Theorem 2.2.1** (Johnson and Lindenstrauss). *Let $\varepsilon \in (0, 1/2)$ and let $D \subset \mathbb{R}^d$ be a set of $n$ points and $k = O(\log n/\varepsilon^2)$. There exits a mapping $f : \mathbb{R}^d \longrightarrow \mathbb{R}^k$ such that for all $u, v \in D$,*

$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2.$$

*Furthermore, this map can be found in randomized polynomial time.*

In Theorem 2.2.1, randomness comes from the fact that in the proof, $f$ has the demanded properties with probability at least $1/n$, thus if the process is repeated $\mathcal{O}(n)$ times, the success probability can be increased to the desired constant, see Dasgupta and Gupta [42]. On the other hand, one of such mapping is the projection to a random subspace of dimension $k$. Algorithmically the projection is obtained by the multiplication of a random $d \times k$ matrix $T = (t_{ij})$ to the right of the matrix of the data (whose rows correspond to the observations). For instance, this matrix can be defined choosing iid $t_{ij}$ with:

$$t_{ij} := \begin{cases} 1 & \text{with probability } 1/2s, \\ 0 & \text{with probability } 1 - 1/s, \\ -1 & \text{with probability } 1/2s, \end{cases}$$

where $s \in \{1, 3\}$, see Achlioptas [1]. Another type of random projection matrix, our choice, selects the $t_{ij}$'s with standard normal distribution or uniform distribution (see, for instance Lemma 2.2 in Dasgupta and Gupta [42] or Vempala [132, Section 1.2] for two precise formulations of this). Formally, the result is given in the following random projection theorem.

**Theorem 2.2.2** (Dasgupta and Gupta [42], Vempala [132]). *Let $\varepsilon, \delta \in (0, 1/2)$ and $T$ be a random $k \times d$ matrix whose rows are chosen independently from $N_d(\mathbf{0}, I_d)$. If $k = O(\log(\delta^{-1})/\varepsilon^2)$ and $\mathbf{x}$ is a unit-length $d$-dimensional vector,*

$$\mathbf{P}\left( \left| \left\| \frac{1}{\sqrt{k}} T\mathbf{x} \right\|^2 - 1 \right| > \varepsilon \right) < \delta.$$

On the other hand, random projections have also the amazing property of, in some sense, allowing to identify distributions: by Cuesta-Albertos *et al.* [35], it is known that only a one dimensional random projection is enough to distinguish between two distributions defined on a separable Hilbert space if one of them satisfies a certain condition on their moments. Let us state this result formally in the $d$-dimensional case. To this, we use the following notation: given two Borel probability measures $\mathbf{P}, \mathbf{Q}$ on $\mathbb{R}^d$, we define $\mathcal{C}(\mathbf{P}, \mathbf{Q}) := \{x \in \mathbb{R}^d : \mathbf{P}_{\langle x \rangle} = \mathbf{Q}_{\langle x \rangle}\}$, where $\langle x \rangle$ denotes the one-dimensional subspace spanned by $x$ and $\mathbf{P}_{\langle x \rangle}$ is the probability $\mathbf{P}$ projected onto the subspace $\langle x \rangle$, i.e. the probability measure on $\langle x \rangle$ given by

$$\mathbf{P}_{\langle x \rangle}(B) = \mathbf{P}(\pi^{-1}_{\langle x \rangle}(B)),$$

where $B$ is a Borel subset of $\langle x \rangle$ and $\pi_{\langle x \rangle}$ denotes the orthogonal projection onto the subspace $\langle x \rangle$.

**Theorem 2.2.3** (Cuesta-Albertos et al. (2007)). *Let $\mathbf{P}, \mathbf{Q}$ be Borel probability measures on $\mathbb{R}^d$, where $d \geq 2$. Assume that:*

- *The absolute moments $m_n := \int \|x\|^n \, \mathrm{d}\mathbf{P}(x)$ are finite and satisfy $\sum_{n \geq 1} m_n^{-1/n} = \infty$ (which implies that $\mathbf{P}$ is characterized by its moments).*

- *The set $\mathcal{C}(\mathbf{P},\mathbf{Q})$ is of positive Lebesgue measure in $\mathbb{R}^d$.*

*Then $\mathbf{P} = \mathbf{Q}$ and $\mathcal{C}(\mathbf{P},\mathbf{Q}) = \mathbb{R}^d$.*

Some applications of this result to statistical problems were mentioned in the Introduction.

## 2.3 What is an outlier?

As we mention in the Introduction, the idea of outlier is universally accepted, but its formalization is not. We classify the notions of outlier in two families: distribution-based and distance-based.

For instance, Hawkins [76] and Febrero *et al.* [56] and [57] give the following distribution-based notion: A multidimensional outlier is an observation generated by a rv with a different distribution than the one of regular observations. When this definition is put into practice, strictly speaking, every element of the sample might be an outlier. Furthermore, as in practice we have no way of exactly knowing the distribution with which the sample was drawn, we could always conclude that no observation is outlier. For example, take a sample generated by the standard normal distribution and a point generated with a Cauchy distribution; represented in Figure 2.1 with black and red dots respectively. In practice, we are not able of distinguishing the red point from the black ones. Therefore, we would declare the red point as regular. Consequently, for this work we discard this family of notions and focus on the second one, which is more convenient to practical use.

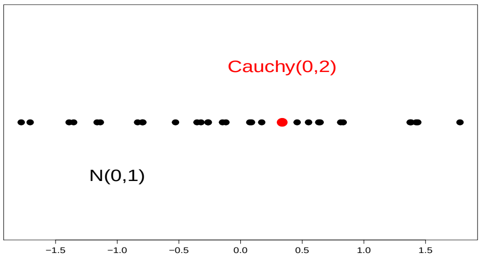


**Figure 2.1.:** Sample drawn from a $N_1(0,1)$ distribution (black) and a point drawn from a Cauchy distribution (red). It is shown the difficulty of noticing the red point as an outlier.

The second family of notions considers as outliers those points lying at a distance greater than a given threshold from the centre of the sample, independently of the distribution which produced them. It is important to emphasize that, in this case, the outlier identification problem does not consist in stating which observations are irregular but rather specifying the observations that lie in a particular region.

The idea that an outlier is a point too separated from the centre of a data set can dated back to 1968 in Healy [77]. It was made more precise in Davies and Gather [44] for dimension $d = 1$ and in Becker and Gather [15] for multidimensional data. Those papers propose computing (robust) estimators of the centre and of the covariance matrix of the data set at hand, and, then declaring outliers those points whose Mahalanobis distances to the estimated centre are greater than a previously fixed threshold. An important characteristic is that the threshold should depend on both $d$ and $n$ (see Theorem 3.4.3 in Chapter 3). Some computational problems were reported, for instance, in Cerioli *et al.* [28] and Cerioli [27], albeit they will not appear in our implementation in Chapter 4.

### 2.3.1 When is a point far away from the rest?

In one dimension the most popular definition of outlier is due to Tukey [131]. There, an outlier is defined as a point which lies outside of the interval

$$(q_1 - 1.5(q_3 - q_1), q_3 + 1.5(q_3 - q_1)), \tag{2.1}$$

where $q_1$ and $q_3$ are respectively the first and third quantile of the sample under consideration. The choice of the constant 1.5 is due to the fact that under normality, we expect to declare as outliers the $0.7\%$ of the data. There is a conservative version of this definition where $1.5$ is replaced by $3$, and therefore only the $0.00023\%$ of the data are expected to be flagged as outliers, which in this case are known as extreme outliers. The weakness of this proposal is that the probability of declaring a point as outlier is not sample size dependent. In fact, as commented in Hoaglin *et al.* [79], this method is highly liberal with the constant $1.5$. For instance, there is a probability of $0.5$ of declaring at least a point as outlier in a outlier-free sample with normal distribution and size $100$. This probability increases to $0.9$ if the sample size is $328$.

The following example shows the importance of taking into account the sample size. Take a sample $X_1, \ldots, X_n$ drawn from the $N_1(0,1)$ with $n = 10$ and another one with $n = 1000$, and introduce the point $3$ in both samples (see Figure 2.2). We see that the point $3$ is inconsistent with the rest of the sample when $n = 10$ (left panel), meanwhile it seems part of the cloud of the data when $n = 1000$ (right panel). In fact, we have

$$\begin{aligned}\mathbf{P}(\max\{X_1, \ldots, X_{10}\} \geq 3) &= 0.01,\\ \mathbf{P}(\max\{X_1, \ldots, X_{1000}\} \geq 3) &= 0.74.\end{aligned}$$

This implies that the point $3$ could be declared as an outlier when $n = 10$ and should not when $n = 1000$. However, the interval defined in equation (2.1) is $(-2.698, 2.698)$, i.e. according to Tukey [131] 3 is always declared as outlier, independently on the sample size.

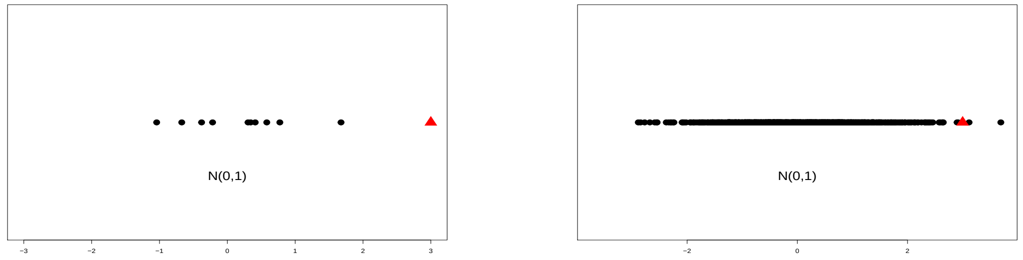


Figure 2.2.: Plots of two simulated random samples (in black) with $N_1(0,1)$ and sample size $n = 10$ (left) and $n = 1000$ (right). The point 3 (in red and triangular) is inconsistent in the left plot, while it seems part of the cloud of the data in the right plot.

Iglewicz and Banerjee [81] provided a modification of Tukey's definition which takes into account the sample size.

**Definition 2.3.1** (Iglewicz and Banerjee). *Let $\delta \in (0,1)$ be the desired probability of declaring at least one outlier for a sample with no outliers of size $n$ and distribution $N(0,1)$. An outlier with respect to that sample is a point which lies out of the interval*

$$(q_1 - f(n)g(n)(q_3 - q_1), q_3 + f(n)g(n)(q_3 - q_1)),$$

*where*

$$f(n) := \frac{\Phi^{-1}[(1-\delta/2)^{1/n}] - .6745}{1.349},$$
$$g(n) := 1 + \frac{8.9764}{n} - \frac{126.6262}{n^2} + \frac{1531.7064}{n^3} - \frac{10729.3439}{n^4},$$

*with $\Phi^{-1}(\cdot)$ the quantile function of a rv with distribution $N(0,1)$.*

The function $g(n)$ of Definition 2.3.1 is an approximate formula which has been calculated through simulations in order to control $\delta$. Note that $g(n)$ is strictly decreasing for $n \geq 30$ and that $g(2000) \approx 1.005$, i.e. back to the above example, the intervals defined in Definition 2.3.1 are $(-2.995, 2.995)$ and $(-4.083, 4.083)$ for $n = 10$ and $n = 1000$, respectively, and $\delta = .05$. Therefore, the point 3 would be declared as outlier when $n = 10$, while it would declared as no outlier when $n = 1000$.

Definition 2.3.1 is still a one dimensional procedure and our scenario is $d$-dimensional, but we can transform the above interval into a ball whose shape should be determined by the covariance matrix $\Sigma$, since we focus on multivariate normally distributed data.

That is, a point must be an outlier if it is outside a ball centred at the centre of the sample with a radius big enough as to contain most points in the sample. Those ideas lead to Definition 1.1.1 presented in the Introduction.

## 2.4 Methods to detect outliers

If $\boldsymbol{\mu}$ and $\Sigma$ are known it is simple to check whether a given point satisfies Definition 1.1.1 or not. However, in practice, those parameters must be estimated and, consequently, somehow obtaining accurate robust estimations of $\boldsymbol{\mu}$ and $\Sigma$ and detecting outliers are essentially equivalent. For that reason we pay some attention in Subsection 2.4.1 to methods devoted to obtain robust estimations of those parameters. However, we already mentioned in the Introduction that those estimators habitually do not work properly in high-dimensional settings. Hence, high-dimensional methods which skip the estimation of $\Sigma$ are explored in Subsection 2.4.2 where we also analyse some dimension reduction procedures. This is the family the random projections method belongs to.

### 2.4.1 Methods that estimate the covariance matrix

There exist many detection methods when $d$ is low or moderate in comparison with $n$, see, for instance Barnett and Lewis [14] and Aggarwal [3] and references therein. First we focus on methods that estimate $\Sigma$. Once $\Sigma$ is estimated, they use Definition 1.1.1. If the sample includes outliers, it is known that the classical least-squares estimates can be strongly affected or completely fail. Hence classical estimators of $\boldsymbol{\mu}$ and $\Sigma$ are replaced by some highly robust ones (Rousseeuw and Van Zomeren [120], Becker and Gather [15] and Peña and Prieto [112]).

A key concept in robust estimation is the breakdown point, a measure used to describe the resistance to the presence of outliers of the estimators. Roughly speaking, the breakdown point of an estimator is the largest fraction of arbitrarily contaminated observations that the sample may contain before the value of the estimator becomes arbitrarily large, see Maronna *et al.* [105] for further references.

Obviously the highest possible breakdown point is $50\%$. Among the methods with highest breakdown point, given $h$ with $n/2 < h < n$, we have the minimum volume ellipsoid estimator (MVE), (Rousseeuw [118]), which searches for the ellipsoid with minimum volume containing $h$ data points, and the minimum covariance determinant (MCD) estimator [118], which is the covariance matrix of the $h$ observations whose covariance matrix has minimal determinant. Davies [43] showed that the MVE has a lower convergence rate, $n^{-1/3}$, than the MCD, $n^{-1/2}$. However the MCD involves a high computational complexity. Additionally, both methods have limitations, because although they estimate feasibly the shape of the covariance matrix, they do not give a good approximation of their scale. The problem is that, both methods require the value $h$ to be previously fixed by the user and, as a consequence, they tend to underestimate or overestimate the covariance matrix depending on the relation between $n - h$ and the real number of outliers in the sample.

There are also some improvements or extensions of those methods to make them more computationally efficient, for instance Rousseeuw and Van Driessen [119], Cerioli [27] and Jobe and Pokojovy [84]. However, those kind of methods do not work well for high-dimensional data or in presence of $n/(d+1)$ outliers or more. For instance, in Adrover and Yohai [2] the maximum bias of the MCD covariance estimator was computed numerically and it was showed that, when the dimension increases, the maximum bias of the MCD grows almost exponentially.

There are some alternatives which require the estimation of $\Sigma$ but they do not use Definition 1.1.1. As an example we finish this section showing a depth based method which although its main goal is not to obtain an estimation of $\Sigma$, it will be required eventually. Such method was proposed by Sun and Genton [128] and it is a $d$-dimensional version of the classical boxplot. We present it here from the multivariate point of view albeit it was proposed in the functional setting. Given a sample in $\mathbb{R}^d$, a depth is a map $D : \mathbb{R}^d \mapsto \mathbb{R}$ such that if $x, y \in \mathbb{R}^d$, then $D(x) \geq D(y)$ is equivalent to that $x$ is deeper than $y$ inside the sample (see Liu *et al.* [100] for more details). The procedure goes as follows: order the points according to this depth, i.e. the first point is the deepest point and so on. We estimate the centre of the data with the deepest point. Compute the $50\%$ convex central region (analogous to the interquartile range in the univariate Tukey's method), in other words: compute the convex set generated by the $50\%$ deeper points in the sample in each direction. Define the fences as a positive constant times the width of this central region intersected with the direction. Any vector such that the segment joining it with the center crosses the fences is declared as an outlier. In Sun and Genton [128], the constant factor is taken as $1.5$ (not size-dependent) and in Sun and Genton [129] is chosen through a procedure that requires the estimation of the covariance matrix. Moreover, both versions of this procedure are useless when $d > n$.

## 2.4.2 Methods to handle high-dimensional cases

When the dimension is higher than the sample size the literature is not so abundant as in the low-dimensional case, but we can mention Fritsch *et al.* [63] and Ro *et al.* [116].

The first paper added a regularization term to the MCD approach to guarantee that the estimation of the parameters is well-posed in the high-dimensional case. However in practice, it is not clear how to determine suitable cut-off values to obtain a desired significance level because there is no result in the paper on the distribution of the involved statistic. On the other hand, the method introduced in Ro *et al.* is based on a modification of the Mahalanobis distance which involves only the diagonal of $\Sigma$. Thus, it is equivalent to consider uncorrelated marginals and this does not usually occur in practice.

Next we mention other methods which do not suffer from the mentioned problems. They are based in dimension reduction techniques.

### Outlier detection methods that use dimensionality reduction

Until now, most procedures that we have mentioned are useless when $d > n$. Moreover, the estimation of the covariance matrix has unexpected features if both $d$ and $n$ are large. For instance, when $d/n \to c \in (0,1)$ and the covariance matrix is the identity, then the distribution of the eigenvalues of the sample covariance matrix follows the Marchenko-Pastur law, which is supported on $((1-\sqrt{c})^2, (1+\sqrt{c}^2)$. Thus, the larger $d/n$, the more spread out the eigenvalues are, see Bickel and Levina [18], and therefore their estimators become inconsistent. Consequently, it is hard or inappropriate to estimate covariance matrices without imposing any structure (e.g. sparse matrices such as in [24]) or assumption (like $n \gg d$).

For that reason, in order to analyse high-dimensional data, dimension reduction techniques such as PP, due to Kruskal [91] and firstly successful implemented in Friedman and Tukey [62], are generally used. These techniques consist in finding a lower dimensional representation of the high-dimensional data, in which the original structure of the data is highly preserved so that the low-dimensional data can be effectively used. This strategy is an important tool and is widely used in many fields of data analysis, data mining, data visualization, and machine learning.

Obviously, a key issue is the choice of the optimality criteria. Principal component analysis (PCA), (Jolliffe [87]), is one of the most widely used technique which employs a optimality criteria. It consists in projecting orthogonally the data on a low-dimensional subspace that captures as much of the variation of the data as possible. More precisely, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $d$-dimensional centred rv's. The first principal direction, $\mathbf{v}_1^n$, is defined as the vector

$$\mathbf{v}_1^n := \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1} \{\mathrm{Var}(\mathbf{X}_1'\mathbf{v}, \ldots, \mathbf{X}_n'\mathbf{v})\}.$$

In general, the estimation of the $i$-th principal direction, $\mathbf{v}_i^n$, is defined as

$$\mathbf{v}_i^n := \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\{\mathrm{Var}(\mathbf{X}_1'\mathbf{v},\ldots,\mathbf{X}_n'\mathbf{v})\},$$

with the additional restriction $(\mathbf{v}_j^n)'\mathbf{v}_i^n = 0$, where $1 \le j < i$. Let $\hat{\sigma}_1^2 \ge \ldots \ge \hat{\sigma}_n^2$ be the ordered eigenvalues of the sample covariance matrix. Then $\mathbf{v}_1^n,\ldots,\mathbf{v}_d^n$ are associate eigenvectors and $\mathrm{Var}(\mathbf{X}_1'\mathbf{v}_j^n,\ldots,\mathbf{X}_n'\mathbf{v}_j^n) = \hat{\sigma}_j^2$. Obviously, there are at most $\min\{n-1,d\}$ nonzero eigenvalues. Therefore PCA fails to yield consistent estimators of all the eigenvectors in very high-dimensional settings, (Johnstone and Lu [86]). For instance, it occurs that it is only possible to obtain the asymptotic distribution for $O(n^{1/5})$ coefficients in the linear functional regression model when a PCA-based estimator is used, (Cardot *et al.* [26]).

Some methods are based on principal components such as the proposed by Maronna and Zamar [104], where it is defined the orthogonalized Gnanadesikan–Kettenring estimate, and the one by Filzmoser *et al.* [61], who designed a computationally efficient high-dimensional method which uses two steps to detect the so-called location and scatter outliers which, roughly speaking, can be identified with clusters of outliers and isolated outliers respectively.

Other procedures that use PP methods are the Stahel-Donoho estimators. Stahel [125] and Donoho [48] independently defined the first equivariant robust multivariate estimator with a high breakdown point of one-half for large data sets, regardless of the dimensions of the data. They look for a univariate projection that makes an observation be an outlier, based on the idea: “The outlyingness measure is based on the idea that if a point is a multivariate outlier, then there must be some one-dimensional projection of the data for which the point is a (univariate) outlier” (Maronna and Yohai [103]). Define for any $\mathbf{x} \in \mathbb{R}^d$ its outlyingness with respect to the sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ by

$$\mathrm{out}(\mathbf{x};\mathbf{X}_1,\ldots,\mathbf{X}_n) := \max_{\mathbf{a}\in\mathbb{R}^d:\|\mathbf{a}\|=1} \frac{|\mathbf{a}'\mathbf{x} - \hat{m}(\mathbf{a}'(\mathbf{X}_1|\cdots|\mathbf{X}_n))|}{\hat{M}^*(\mathbf{a}'(\mathbf{X}_1|\cdots|\mathbf{X}_n))}, \tag{2.2}$$

where $(\mathbf{X}_1|\cdots|\mathbf{X}_n)$ denotes the matrix whose columns are the vectors of the sample.

Obviously, examining all the possible directions, as in equation (2.2), is the optimal method but unachievable in practice. To overcome this computational burden, there exist methods which entail only finitely many projections. For instance, the number of directions can be approximated by random subsamples as in Stahel [125], but there is a high computational burden to obtain satisfactory results, see Maronna and Yohai [103]. Peña and Prieto [112] proposed another way of restricting the search to a finite number of directions. In their work, the data are projected onto a certain set of $2d$ directions, where $d$ is the dimension of the data. Such directions are those that maximize and minimize the kurtosis coefficient of the projected data. This choice is due to the fact that a small number of outliers would cause heavy tails and lead to a larger kurtosis coefficient, meanwhile a number of clustered outliers would start introducing bimodality and decrease the kurtosis coefficient. However, they necessitate the estimation of the covariance matrix at the beginning of the process.

Other paper which uses finitely many deterministic directions is Pan *et al.* [111]. There, for $\mathbf{X}$ a $d$-dimensional random vector with distribution $F$ and $\mathbf{a} \in \Omega^{d-1}$, it is defined the *projected Hampel identifier* as:

$$V(\mathbf{x}, \mathbf{a}, F) := \frac{\mathbf{a}'\mathbf{x} - m(F^{\mathbf{a}})}{M^*(F^{\mathbf{a}})}, \tag{2.3}$$

where $F^{\mathbf{a}}$ denotes the distribution of $\mathbf{a}'\mathbf{X}$. Then, this outlier identifier is based on the differences between $V(\mathbf{x}, \mathbf{a}, F_n) - V(\mathbf{x}, \mathbf{a}, F)$, where $V(\mathbf{x}, \mathbf{a}, F_n)$ is defined as in equation (2.3) bit replacing $F^{\mathbf{a}}$ by the empirical distribution. Since the theoretical distribution of the data is usually unknown, a bootstrap step is proposed. Furthermore, they assume that $F$ has elliptical contours and they do not provide the exact number of the required directions, which are chosen following an algorithm described in Fang and Wang [51].

Serfling and Mazumder [122] describe another alternative method which uses a finite number of projections. Although they do not assume elliptical contours, they need of the estimation of the covariance matrix.

## 2.5 Spherical Distributions

The uniform distribution on $\Omega_r^{d-1}$, $d \geq 2$, is a particular case of the spherical distribution, as stated in Proposition 2.5.3 . To begin this section, we define spherical distributions. Then we include some well-known results for further reference.

**Definition 2.5.1.** *Let $\mathbf{U}$ be a $\Omega_r^{d-1}$-valued rv. It is said that $\mathbf{U}$ has a spherical distribution (or just that it is spherical) if $T\mathbf{U}$ has the same distribution as $\mathbf{U}$, for every orthogonal $d \times d$ matrix $T$.*

The proofs of the following results are immediate, but see Section 3.3 of McNeil *et al.* [107] and Theorem 2.5.1 in Fang and Zhang [52]. Proposition 2.5.2 gives a characterization for the spherical distributions.

**Proposition 2.5.2.** *The following statements are equivalent.*

1. $\mathbf{U} = (U_1, \ldots, U_d)$ *has spherical distribution.*

2. *For all* $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{a}'\mathbf{U}$ *has the same distribution as* $\|\mathbf{a}\|U_1$.

**Proposition 2.5.3.** *Let* $\mathbf{U}$ *be a rv with uniform distribution on* $\Omega^{d-1}$*, then* $\mathbf{U}$ *is spherical.*

### 2.5.1 Uniform distribution on the hypersphere

The following lemma gives the cdf of the marginal of a uniform distribution on $\Omega^{d-1}$. We include its proof because it was done by the author, although alternative proofs can be found in the literature, for instance see Mardia and Jupp [102, page 161]) where the tangent-normal decomposition is used.

**Lemma 2.5.4.** *Let* $\mathbf{U} = (U_1, \ldots, U_d)$ *be a rv with distribution uniform on* $\Omega^{d-1}$*. The distribution function of* $U_1$ *is given by the following expression:*

$$F_{d-1}(u) := \frac{1}{2}\left(1 + \operatorname{sign}(u)\mathrm{I}_{u^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right), u \in [-1, 1].$$

*Proof.* By (2.5.11) in Fang and Zhang [52],

$$F_{d-1}(u) = \mathrm{B}\left(\tfrac{1}{2}, \tfrac{d-1}{2}\right)^{-1} \int_{-1}^{u} (1-y^2)^{(d-3)/2}\, \mathrm{d}y.$$

By the change of variable $y^2 = s$,

$$F_{d-1}(u) = \frac{1}{2\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} \left(\operatorname{sign}(u)\mathrm{B}\left(u^2; \frac{1}{2}, \frac{d-1}{2}\right) + \mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right).$$

Then, the result is deduced from the definition of the incomplete beta function. □

Some particular cases for $d = 2, 3$ in Lemma 2.5.4 are, for $u \in [-1, 1]$,

$$F_1(u) = 1 - \frac{1}{\pi}\cos^{-1}(u), \qquad F_2(u) = \frac{1}{2}(u+1). \tag{2.4}$$

Due to the recurrence properties of the incomplete beta function, the next relation holds for $d \geq 4$

$$F_{d-1}(u) = F_{d-3}(u) + \frac{u(1-u^2)^{(d-3)/2}}{(d-3)\mathrm{B}\left(\frac{1}{2}, \frac{d-3}{2}\right)}. \tag{2.5}$$

## 2.6 Uniformity tests on the hypersphere

We introduce in this section some well-known uniformity tests which will be mentioned in Chapter 5. We begin reviewing the construction and specifics of the Sobolev tests which are closely related with the projection-based uniformity tests on the hypersphere that we propose here, as we will see in Section 5.2.1 of Chapter 5, and contains most of the nonparametric uniformity tests in $\Omega^{d-1}$ for $d \geq 2$. Then, we briefly see a test based on a finite number of random projections, thus being a test which follows the line ii) described in the Introduction as opposed to the test we introduce in Chapter 5.

### 2.6.1 Sobolev tests of uniformity

Sobolev tests, as introduced in Beran [16] and Giné [71], are based on the eigenfunctions of the Laplacian on $\Omega^{d-1}$, which form an orthonormal basis (the so-called *spherical harmonics*) on $L^2(\Omega^{d-1}, \nu_{d-1})$, the space of square-integrable functions on $\Omega^{d-1}$ with respect to the uniform measure $\nu_{d-1}$ of $\Omega^{d-1}$. Denoting by $\mathcal{E}_k$ to the space of eigenfunctions corresponding to the $k$-th non-zero eigenvalue of the Laplacian, it is well-known that $L^2(\Omega^{d-1}, \nu_{d-1}) = \bigoplus_{k=0}^{\infty} \mathcal{E}_k$, where $\mathcal{E}_0$ is the space of constant functions. Moreover,

$$\ell_{k,d-1} := \dim(\mathcal{E}_k) = \binom{d+k-3}{d-2} + \binom{d+k-2}{d-2}. \tag{2.6}$$

Then, $\cup_{k=1}^{\infty}\{g_{i,k} : i = 1, \ldots, \ell_{k,d-1}\}$ is an orthonormal basis for $L^2(\Omega^{d-1}, \nu_{d-1})$ and so is $\{g_{i,k} : i = 1, \ldots, \ell_{k,d-1}\}$ for $\mathcal{E}_k$. Theorem 1.5.1 in Dai and Xu [41]) gives expressions of $g_{i,k}$ for $d \geq 2$. This allows to write $\psi \in L^2(\Omega^{d-1}, \nu_{d-1})$, a pdf with respect to $\nu_{d-1}$, as

$$\psi(\mathbf{u}) = \sum_{k=0}^{\infty} \sum_{i=1}^{\ell_{k,d-1}} e_{i,k} g_{i,k}(\mathbf{u}),$$

where $e_{i,k} = \langle f, g_{i,k} \rangle := \int_{\Omega^{d-1}} \psi(\mathbf{u}) g_{i,k}(\mathbf{u})\, \nu_{d-1}(\mathrm{d}\mathbf{u}) = \mathrm{E}_f[g_{i,k}(\mathbf{X})]$.

Importantly, due to the orthogonality of the sets $\mathcal{E}_k$, $\psi$ equals the uniform pdf if and only if $e_{i,k} = 0$ for all $i = 1, \ldots, \ell_{k,d-1}$ and $k \geq 1$. Thus, $\mathbf{H}_0$ is conveniently characterized within $L^2(\Omega^{d-1}, \nu_{d-1})$.

Sobolev tests exploit the previous characterization with statistics that inspect the empirical version of $e_{i,k}$'s and that weight the observed deviations from zero on each $\mathcal{E}_k$ by a sequence of weights. Formally, they consider the mapping

$$\mathbf{u} \in \Omega^{d-1} \mapsto \mathrm{t}(\mathbf{u}) := \sum_{k=1}^{\infty} v_{k,d-1} \mathrm{t}_k(\mathbf{u}) \in L^2(\Omega^{d-1}, \nu_{d-1}),$$

where $\mathrm{t}_k(\mathbf{u}) := \sum_{i=1}^{\ell_{k,d-1}} g_{i,k}(\mathbf{u})g_{i,k} \in \mathcal{E}_k$ (which allows to construct the estimates $(\hat{e}_{1,k},\ldots,\hat{e}_{\ell_{k,d-1},k})$), and $\{v_{k,d-1}\}$ is an arbitrary real sequence such that $\sum_{k=1}^{\infty} v_{k,d-1}^2 \ell_{k,d-1} < \infty$. Then, the Sobolev test for $\{v_{k,d-1}\}$ rejects $\mathbf{H}_0$ in (1.3) for large values of the statistic

$$S_{n,d-1}(\{v_{k,d-1}\}) := \frac{1}{n}\left\|\sum_{i=1}^{n} \mathrm{t}(\mathbf{X}_i)\right\|^2 = \frac{1}{n}\sum_{i,j=1}^{n}\sum_{k=1}^{\infty} v_{k,d-1}^2 \langle \mathrm{t}_k(\mathbf{X}_i), \mathrm{t}_k(\mathbf{X}_j)\rangle. \tag{2.7}$$

Given $\mathbf{u},\mathbf{v} \in \Omega^{d-1}$, an explicit form (Prentice [114, Proposition 2.1]) for the addends in (2.7) is

$$\langle \mathrm{t}_k(\mathbf{u}), \mathrm{t}_k(\mathbf{v})\rangle = \begin{cases} 2T_k(\mathbf{u}'\mathbf{v}), & d = 2, \\ \left(1 + \frac{2k}{q-1}\right)C_k^{(d-2)/2}(\mathbf{u}'\mathbf{v}), & d \geq 3, \end{cases} \tag{2.8}$$

where $T_k$ and $C_k^{(d-2)/2}$ are the $k$th Chebyshev polynomial of the first kind and the $k$th Gegenbauer polynomial of order $(d-2)/2$, respectively (defined in Subsection 2.1.1). Since

$$\lim_{\alpha\to 0^+} \frac{1}{\alpha} C_k^{\alpha}(z) = \frac{2}{k} T_k(z) \ \text{ for } \ k \geq 1, \tag{2.9}$$

see equation 18.7.25 in NIST [46], the case $d = 2$ in (2.8) can be seen as an extension of the $d \geq 3$ case, which enables writing $\langle \mathrm{t}_k(\mathbf{u}), \mathrm{t}_k(\mathbf{v})\rangle = \left(1 + \frac{2k}{d-2}\right)C_k^{(d-2)/2}(\mathbf{u}'\mathbf{v})$ for $d \geq 2$ by assuming implicitly such extension for $d = 2$. We do so henceforth for the sake of brevity, unless an explicit separation of the two cases is beneficial for clarity.

Alternatively, Sobolev tests can be constructed as the locally most powerful rotation-invariant tests for testing $\mathbf{H}_0$ in (1.3) against specified alternatives $f(\cdot'\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \Omega^{d-1}$ (Beran [16] and Giné [71]). The construction is based on an application of the Neyman–Pearson Lemma and a locality argument (assuming $f \approx 1$), providing an equivalent expression for the test statistic (2.7):

$$S_{n,d-1}(\{v_{k,d-1}\}) = \frac{1}{n}\int_{\Omega^{d-1}} \left(\sum_{i=1}^{n} f(\mathbf{X}_i'\boldsymbol{\sigma}) - n\right)^2 \nu_{d-1}(\mathrm{d}\boldsymbol{\sigma}), \tag{2.10}$$

where

$$f(z) := 1 + \sum_{k=1}^{\infty}\left(1 + \frac{2k}{d-2}\right) v_{k,d-1} C_k^{(d-2)/2}(z), \quad z \in [-1,1]. \tag{2.11}$$

Proposition 2.1 in Prentice [114] states the existence of $g_f \in L_{d-1}^2[-1,1]$ such that

$$g_f(z) = \sum_{k=1}^{\infty} b_{k,d-1} C_k^{(d-2)/2}(z), \quad b_{k,d-1} = (1 + 2k/(d-2))v_{k,d-1}^2, \quad k \geq 1, \tag{2.12}$$

and such that

$$S_{n,d-1}(\{v_{k,d-1}\}) = \frac{1}{n}\sum_{i,j=1}^{n} g_f\left(\mathbf{X}_i'\mathbf{X}_j\right). \tag{2.13}$$

When $d = 2$, the continuity extension (2.9) is assumed in (2.11) and (2.12), resulting $f(z) = 1 + \sum_{k=1}^{\infty}(2b_{k,1})^{1/2}T_k(z)$ and $g_f(z) = \sum_{k=1}^{\infty} b_{k,1}T_k(z)$ with $b_{k,1} = 2v_{k,1}^2$, $k \geq 1$.

**Remark 2.6.1.** *Equation* (2.13) *and the comment which follows it, show that it is rather* $\{v_{k,d-1}^2\}$ *who determines* $S_{n,d-1}(\{v_{k,d-1}\})$. *Moreover,* $f$ *immediately provides* $\{v_{k,d-1}^2\}$, *hence* $S_{n,d-1}(\{v_{k,d-1}\})$. *Because of this, we will use the notation* $S_{n,d-1}(f)$ *to refer to* $S_{n,d-1}(\{v_{k,d-1}\})$ *when the indexing by* $f \in \mathcal{F}_{d-1}$ *provides better clarity. For that, we define*

$$\mathcal{F}_{d-1} := \left\{f \in L_{d-1}^2[-1,1] : f(z) = 1 + \sum_{k=1}^{\infty}\left(1 + \tfrac{2k}{d-2}\right) v_{k,d-1}C_k^{\frac{d-2)}{2}}(z),\ \sum_{k=1}^{\infty} v_{k,d-1}^2\ell_{k,d-1} < \infty\right\}.$$

*Note that two different functions* $f, f^* \in \mathcal{F}_{d-1}$ *such that* $v_{k,d-1} = \pm v_{k,d-1}^*$, $k \geq 1$, *determine the same squared coefficients, hence yielding the same statistic.*

The following precise definition of the Sobolev class of test statistics will be useful in Chapter 5.

**Definition 2.6.2** (Sobolev class). *The* Sobolev class *of test statistics is defined as* $\mathcal{S} := \left\{S_{n,d-1}(\{v_{k,d-1}\}) : \{v_{k,d-1}\} \subset \mathbb{R}, \sum_{k=1}^{\infty} v_{k,d-1}^2\ell_{k,d-1} < \infty\right\} = \left\{S_{n,d-1}(f) : f \in \mathcal{F}_{d-1}\right\}$. *The* $h$-finite Sobolev class *is defined as*

$$\mathcal{S}_h := \{S_{n,d-1}(\{v_{k,d-1}\}) \in \mathcal{S} : \{v_{k,d-1}\} \text{ has at most } h \text{ non-null terms}\}.$$

Clearly, $\mathcal{S}_h \subset \mathcal{S}$, $\forall h \geq 1$. The $h$-finite Sobolev class has been studied in Jupp [88] and Jammalamadaka *et al.* [83].

For the sake of reference, we collect in the following theorem the main results on the tests based on (2.7) and (2.10), as stated in Giné [71] and Prentice [114].

**Theorem 2.6.3** (Giné [71], Prentice [114]). *Let* $\{v_{k,d-1}\}$ *be a real sequence satisfying* $\sum_{k=1}^{\infty} v_{k,d-1}^2\ell_{k,d-1} < \infty$. *Let* $Y_k$ *be independent rv's with* $\chi_{\ell_{k,d-1}}^2$ *distribution,* $k \geq 1$. *Then, under* $\mathbf{H}_0$ *in* (1.3),

$$S_{n,d-1}(\{v_{k,d-1}\}) \overset{d}{\rightsquigarrow} \sum_{k=1}^{\infty} v_{k,d-1}^2 Y_k. \tag{2.14}$$

*In addition, the test that rejects for large values of $S_{n,d-1}(\{v_{k,d-1}\})$ is asymptotically and locally (in $\kappa \to 0$) most powerful rotation-invariant (except $O(\kappa^3)$ terms) against the alternative with pdf*

$$f_{\boldsymbol{\mu},\kappa}(\mathbf{x}) := (1-\kappa)\frac{1}{\omega^{d-1}} + \kappa\frac{f(\mathbf{x}'\boldsymbol{\mu})}{\omega^{d-1}}, \quad 0 < \kappa \leq 1, \tag{2.15}$$

*where $\boldsymbol{\mu} \in \Omega^{d-1}$ is unspecified and $f$ is given by (2.11). Furthermore, if $v_{k,d-1} \neq 0$, for all $k \geq 1$, the test is consistent against all non-uniform alternatives with pdf in $L^2(\Omega^{d-1}, \nu_{d-1})$.*

Further details and insights on Sobolev tests can be seen in Section 3 of García-Portugués and Verdebout [64] and references therein.

We finish this subsection showing some well-known Sobolev tests which will be mentioned in Chapter 5. Some remarks about them: i) Watson, Rothman, Ajne and Rayleigh tests were initially proposed to test (1.3) for the circular case (although Ajne and Rayleigh tests are immediately generalizable to $\Omega^{d-1}$ for $d > 2$, both through the Sobolev class of tests); meanwhile, Bingham was proposed for the sphere. ii) Although Rayleigh, Watson, Rothman, Ajne, and Bingham tests were introduced before the definition of Sobolev tests, Giné showed that they belong to this class, see Giné [71].

**Rayleigh test:** The Rayleigh test [115] is based on the fact that if $\mathbf{X}$ has $\nu_1$ distribution, then $\mathrm{E}(\mathbf{X}) = 0$ or, equivalently, $\|\mathrm{E}(\mathbf{X})\|^2 = 0$. This leads to define the Rayleigh's statistic as

$$R_n := 2n\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\right\|^2 = \frac{2}{n}\left(\left(\sum_{i=1}^{n}\cos\Theta_i\right)^2 + \left(\sum_{i=1}^{n}\sin\Theta_i\right)^2\right).$$

**Watson test:** The Watson [135] statistic is defined as

$$U_n^2 := n\int_0^{2\pi}\left\{G_n(\theta) - G(\theta) - \int_0^{2\pi}(G_n(\varphi) - G(\varphi))\,\mathrm{d}G(\varphi)\right\}^2\mathrm{d}G(\theta), \tag{2.16}$$

where $G_n(\theta) := (1/n)\sum_{i=1}^{n} 1_{\{\Theta_i \leq \theta\}}$ is the empirical cdf of the circular sample $\Theta_1, \ldots, \Theta_n$ in $[0, 2\pi)$, $G(\theta) := \theta/(2\pi)$ is the uniform cdf on $[0, 2\pi)$, where the origin is implicitly assumed to be $0$. The statistic $U_n^2$ has several neat connections with other circular and linear uniformity tests; particularly it can be regarded as the rotation-invariant version of the CvM statistic that selects the origin in such a way that the discrepancy of the sample with respect to $\mathbf{H}_0$ is minimized (see, e.g., García-Portugués and Verdebout [64]).

Alternatively, $U_n^2$ has the form (see, e.g., Mardia and Jupp [102, p. 111]):

$$U_n^2 = \frac{1}{n}\sum_{i,j=1}^{n} h\left(\theta_{ij}\right) = \frac{2}{n}\sum_{i<j} h\left(\theta_{ij}\right) + \frac{1}{12}, \quad h(\theta) = \frac{1}{2}\left(\frac{\theta^2}{4\pi^2} - \frac{\theta}{2\pi} + \frac{1}{6}\right), \tag{2.17}$$

where $\theta_{ij} = \cos^{-1}(\cos(\Theta_i - \Theta_j)) \in [0, \pi]$ is the shortest angle distance between $\Theta_i$ and $\Theta_j$.

**Rothman test:** The test statistic by Rothman [117] compares the number of expected and observed data points in arcs of $\Omega^1$ of length $2\pi t$ in a rotationally-invariant way:

$$R_{n,t} := \frac{1}{2\pi n}\int_0^{2\pi} \left(N(\alpha, t) - nt\right)^2 \, \mathrm{d}\alpha, \tag{2.18}$$

where $N(\alpha, t) := \#\left\{\Theta_1, \ldots, \Theta_n : \cos^{-1}(\cos(\Theta_i - (\alpha + t\pi))) < t\pi, i = 1, \ldots, n\right\}$ represents the number of observations in the arc $[\alpha, \alpha + 2\pi t)$, for $\alpha \in [0, 2\pi)$ and $t \in (0, 1)$. The Ajne [4] statistic arises as a particular case of $R_{n,t}$ with $t = 1/2$:

$$A_n := \frac{1}{2\pi n}\int_0^{2\pi} \left(N\left(\alpha, \tfrac{1}{2}\right) - \tfrac{n}{2}\right)^2 \mathrm{d}\alpha = \frac{n}{4} - \frac{1}{n\pi}\sum_{i<j}\theta_{ij}. \tag{2.19}$$

**Bingham test:** Bingham test [19] is based on the fact that if $\mathbf{X}$ has $\nu_{d-1}$ distribution, then $\mathrm{E}(\mathbf{X}'\mathbf{X}) = I_d/d$ or equivalently, the trace $\mathrm{tr}\left(\mathrm{E}(\mathbf{X}'\mathbf{X})^2\right) - d^{-1} = 0$. It is defined as

$$B_n := \frac{nd(d+2)}{2}\left(\mathrm{tr}(S_n^2) - \frac{1}{d}\right),$$

where $S_n := n^{-1}\sum_{i=1}^n \mathbf{X}_i'\mathbf{X}_i$ is the empirical covariance matrix of the $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

**Giné test:** The Giné statistic [71] designed for $d = 2$, extended in Prentice [114] for $d \geq 2$, is defined as

$$G_n := \frac{n}{2} - \frac{(d-1)\Gamma((d-1)/2)^2}{2n\Gamma(d/2)^2}\sum_{1\leq i<j\leq n} \sin(\cos^{-1}(\mathbf{X}_i'\mathbf{X}_j)).$$

**Bakshaev test:** Bakshaev [11] proposed the chordal-based test statistic

$$N_{n,d-1} := n\mathrm{E}\|\mathbf{X}_1 - \mathbf{X}_2\| - \frac{1}{n}\sum_{i,j=1}^{n}\|\mathbf{X}_i - \mathbf{X}_j\|$$

for testing uniformity on $\Omega^d$. Observing that $\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{2 - 2\cos(\theta_{ij})} = 2\sin(\theta_{ij}/2)$ and that $\mathrm{E}\|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{2}\int_{-1}^{1}\sqrt{1-t}\,\mathrm{d}F_{d-1}(t)$ under $\mathbf{H}_0$, when $d = 3$ the test statistic can be written as

$$N_{n,2} = \frac{4n}{3} - \frac{4}{n}\sum_{i<j}\sin\left(\frac{\theta_{ij}}{2}\right). \tag{2.20}$$

### 2.6.2 A test based on random projections

Cuesta-Albertos *et al.* [37] proposed a uniformity test based on random projections. It is based on the characterization in Theorem 2.2.3. As a consequence of this characterization, testing $\mathbf{H}_0$ in (1.3) is (almost surely) equivalent to testing $\mathbf{H}_0^\gamma : \gamma'\mathbf{X}$ has distribution $F_{d-1}$, where $F_{d-1}$ (see Lemma 2.5.4) is the common cdf of the sample of random projections $\gamma'\mathbf{X}_1, \ldots, \gamma'\mathbf{X}_n$ under $\mathbf{H}_0$ (see Proposition 3.2.2), and $\gamma$ is a rv with distribution on $\Omega^{d-1}$ independent of the sample. Denoting by $F_{n,\gamma}$ to the empirical cdf of the projections $\gamma'\mathbf{X}_1, \ldots, \gamma'\mathbf{X}_n$, the test rejects $\mathbf{H}_0^\gamma$, and consequently $\mathbf{H}_0$, for large values of the Kolmogorov–Smirnov test statistic

$$D_{n,\gamma} := \sup_{x\in[-1,1]} |F_{n,\gamma}(x) - F_{d-1}(x)|. \tag{2.21}$$

As already mentioned, this test depends on the random direction $\gamma$. In order to alleviate this, Cuesta-Albertos *et al.* [37] considered $k$ random projections on the iid rv's $\gamma_1, \ldots, \gamma_k$ and used the *aggregated* statistic $C_n := \min\{p_{\gamma_1}, \ldots, p_{\gamma_k}\}$, where $p_{\gamma_j}$ represents the $p$-value associated to the test performed with the random direction $\gamma_j$. The test rejects for low values of $C_{n,\gamma_1,\ldots,\gamma_k}$. The distribution of $C_n$ under $\mathbf{H}_0$ is unknown, but the authors approximate it by Monte Carlo simulations (conditionally on the sample $\gamma_1, \ldots, \gamma_k$).

## 2.7 Required results about integral equations

We show next some results from the theory of integral equations that are needed to prove Theorem 5.2.2 and are particularizations for real $L^2[-1,1]$ kernels of results in Smithies [123, Chapters 7 and 8].

**Definition 2.7.1** ($L^2$-kernel and its adjoint)**.** *A real Borel measurable function $K$ defined on $[-1,1]\times[-1,1]$ is called an* $L^2$-kernel *if $\int_{-1}^{1}\int_{-1}^{1} K(s,t)^2\,\mathrm{d}s\,\mathrm{d}t$, $\int_{-1}^{1} K(s,t)^2\,\mathrm{d}s$, and $\int_{-1}^{1} K(s,t)^2\,\mathrm{d}t$ are finite for every $s,t\in[-1,1]$. The* adjoint kernel *of $K$, denoted by $K^*$, is defined as $K^*(s,t) := K(t,s)$, for $s,t\in[-1,1]$.*

**Definition 2.7.2** (Singular values and singular functions)**.** *Let $K$ be an $L^2$-kernel and $u, v \in L^2[-1,1]$ such that $u, v \neq 0$. The real number $\mu$ is called a* singular value *of $K$ and $[u, v]$ are referred to as a pair of* singular functions *of $K$ associated to the singular value $\mu$ if*

$$u(s) = \mu \int_{-1}^{1} K^*(s,t)v(t)\,\mathrm{d}t \quad \textit{and} \quad v(s) = \mu \int_{-1}^{1} K(s,t)u(t)\,\mathrm{d}t.$$

It happens that the set of the singular values of $K$ is either finite or denumerable with no finite limit points. In addition, given a singular value $\mu$, the set of singular functions associated to $\mu$ constitutes a finite dimensional linear subspace. Then, each singular value $\mu$ admits a finite number of pairs of singular functions, say $\{[u_i, v_i]\}_{i=1}^{d_\mu}$, which are an orthonormal basis of the corresponding singular functions. The union of those bases forms a *full orthonormal system* of singular functions.

In what follows we consider ordered systems of singular values $\{\mu_n\}_{n=1}^{\infty}, 0 < \mu_1 \leq \mu_2 \leq \ldots$, where each singular value is repeated as many times as the dimension of the associated subspace of singular functions $d_{\mu_n}$. With an abuse of notation, the collection of singular values and singular functions $\{([u_n, v_n]; \mu_n)\}_{n=1}^{\infty}$, referred to as *singular system*, is treated as infinite.

**Proposition 2.7.3** (Theorem 8.7.1 in Smithies [123])**.** *Let $\{([u_n, v_n]; \mu_n)\}_{n=1}^{\infty}$ be a singular system of the $L^2$-kernel $K$ and let $y \in L^2[-1,1]$. Then, the equation*

$$y(s) = \int_{-1}^{1} K(s,t)x(t)\,\mathrm{d}t$$

*has a solution $x \in L^2[-1,1]$ for almost every $s \in [-1,1]$ if and only if*

*a)* $\sum_{n=1}^{\infty} \mu_n^2 \int_{-1}^{1} (y(s)u_n(s))^2\,\mathrm{d}s < \infty$*;*

*b)* $\int_{-1}^{1} y(s)u(s)\,\mathrm{d}s = 0$ *for every function $u \in L^2[-1,1]$ such that $\int_{-1}^{1} K^*(s,t)u(s)\,\mathrm{d}s = 0$ for almost every $t \in [-1,1]$.*

## 2.8 Sequential analysis

A sequential method is characterized by a stopping rule that decides whether to stop the observation process with $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ or to get an additional observation $\mathbf{X}_{n+1}$ for $n \geq 1$. Therefore, the sample size is not predetermined, in fact it is a random variable.

Those procedures started with Wald's sequential probability ratio tests (SPRTs) [133] around the 1930s and they have been used very often (see for instance Box *et al.* [22], Montgomery [108] and Wetherill and Brown [139]). They were initially developed as methods of quality control, and since then, they form the basis of many more sophisticated developments in quality control methodologies. They are also widely used in general test of hypotheses problems. In fact, they are applied in a wide range of situations: biomedical signal and image processing, intrusion detection in computer networks and security systems, detection of road traffic incidents, human motion analysis, etc, (Tartakovski *et al.* [130]).

**Wald's SPRT.** Wald's SPRT was originally formulated as a test between two simple hypotheses. Let $x_1, \dots, x_n$ be a sequence of iid observations of a rv $X$ and $f(x, \theta)$ be the pdf of $X$, where the parameter $\theta$ is unknown. The hypotheses to be sequentially tested are of the form $\mathbf{H}_0 : \theta = \theta_0$ and $\mathbf{H}_1 : \theta = \theta_1$. Let $m$ be a positive integer, the likelihood of a sample $x_1, \dots, x_m$ is $\prod_{i=1}^m f(x_i, \theta_0)$, when $\mathbf{H}_0$ is true and $\prod_{i=1}^m f(x_i, \theta_1)$, when $H_1$ is true. Let $0 < a < b$ be two constants. Consider the ratios:

$$\Lambda_m := \prod_{i=1}^{m} \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)}.$$

At any stage of the process ($m$-th step with $m \geq 1$), $\Lambda_m$ is computed and

- Stop and accept $\mathbf{H}_1$ if $\Lambda_m \geq b$.

- Stop and accept $\mathbf{H}_0$ if $\Lambda_m \leq a$.

- Continue and take an additional observation if $a < \Lambda_m < b$.

The constants $a$ and $b$ are chosen in order to control the prescribed errors of type I and II, $\alpha$ and $\beta$. Wald suggested that in practice those constants can be approximated by $a \approx (1 - \beta)/\alpha$ and $b \approx (1 - \alpha)/\beta$.

Two functions characterize the SPRT:

- The operating characteristic (OC), denoted by $L(\theta)$, is the probability of accepting $\mathbf{H}_0$ when the parameter is $\theta$. Thus, it is one minus the power function.

- The average sample number (ASN), or expected duration of the experiment, denoted by $\mathrm{E}_\theta(K)$, is the mean number of sample points which are necessary for testing the hypotheses when the true value of the parameter is $\theta$.

The performance of SPRTs has been studied extensively. It is well known that Wald's SPRT is the optimal method (in the sense that it requires the smallest number of required observations to achieve the given errors $\alpha$ and $\beta$) when the observations are iid and the hypotheses to be tested are simple, (Wald and Wolfowitz [134] and Matthes [106]). In particular, let $L(\theta)$ and $\mathrm{E}_\theta(K)$ be the OC and the ASN respectively for a SPRT, and

$\tilde{L}(\theta)$ and $\tilde{\mathrm{E}}_\theta(K)$ be the same functions but for another test. If $\tilde{L}(\theta_0) \geq L(\theta_0)$ and $\tilde{L}(\theta_1) \leq L(\theta_1)$, then $\tilde{\mathrm{E}}_{\theta_0}(K) \geq \mathrm{E}_{\theta_0}(K)$ and $\tilde{\mathrm{E}}_{\theta_1}(K) \geq \mathrm{E}_{\theta_1}(K)$.

However, these assumptions may be quite restrictive in real applications (the observations are not always iid and the hypothesis to be tested are often composite). Indeed, theoretically, SPRTs require neither the hypotheses to be simple nor the observations to be iid, neither the thresholds for the test to be constant over time. Nevertheless, studying SPRTs properties and behaviour becomes much more complicated without these assumptions. For instance, there exist some non-parametric SPRTs with dependent data such as Lai [93], although the optimal parametric SPRTs with dependent data is still an open problem (see Niu and Varshney [110]). There also exist generalized SPRTs where the thresholds are not constant, unfortunately the determination of how those thresholds should be chosen is still not solved (see Gölz *et al.* [73]).

# 3 Outliers detection: known parameters

> *Let us choose for ourselves our path in life, and let us try to strew that path with flowers.*
>
> — **Emilie du Chatelet**

This chapter aims to propose a method to decide whether a point is an outlier or not with respect to a multivariate normal distribution with known parameters. This is a previous step to study the method when the parameters are unknown, which is the most common situation in practice. We also provide the expected number of projections required to declare a point as outlier or not and also the values of the constants $a$, $b$ which determine the test when the covariance matrix is the identity (see Propositions 3.3.1 and 3.5.4 respectively).

## 3.1 Scenario of the method and notation

Let $\boldsymbol{\mu}$ and $\Sigma$ be known parameters and $C_n^d(\delta)$ defined in (1.1). Basic ideas presented in the Introduction lead to test the following hypotheses for a given point $\mathbf{x} \in \mathbb{R}^d$ :

$$\begin{aligned} \mathbf{H}_0: &\quad \|\mathbf{x}-\boldsymbol{\mu}\|_\Sigma \leq C_n^d(\delta), \text{ i.e. } \mathbf{x} \text{ is no outlier wrt } N_d(\boldsymbol{\mu},\Sigma), \\ \mathbf{H}_1: &\quad \|\mathbf{x}-\boldsymbol{\mu}\|_\Sigma > C_n^d(\delta), \text{ i.e. } \mathbf{x} \text{ is an outlier wrt } N_d(\boldsymbol{\mu},\Sigma). \end{aligned} \tag{3.1}$$

The procedure to test (3.1) is analogous to Algorithm 2 replacing $\hat{\nu}_\mathbf{V}$ and $\hat{\lambda}_\mathbf{V}$ by $\nu_\mathbf{V}$ and $\lambda_\mathbf{V}$, which are measures of central tendency and of dispersion, respectively, of the projected distribution.

Our selection in this work for $\nu_\mathbf{V}$ and $\lambda_\mathbf{V}$ will be, respectively, the mean, $\boldsymbol{\mu}_\mathbf{V}$, or the median, $m_\mathbf{V}$, and the standard deviation, $\sigma_\mathbf{V}$, or the MAD, denoted by $M_\mathbf{V}^*$. It is well known that under normality the MAD overestimates the standard deviation (see Maronna *et al.* [105]). To make it consistent (see ibid), we use the normalized MAD, abridged to MADN: $M_\mathbf{V} = M_\mathbf{V}^*/q_3$, where $q_3$ is the third quantile of a $N_1(0,1)$

distribution. In this chapter no estimation is needed, and $\boldsymbol{\mu}_{\mathbf{V}} = m_{\mathbf{V}}$ and $\sigma_{\mathbf{V}} = M_{\mathbf{V}}$. Hence, to simplify, we use $\boldsymbol{\mu}_{\mathbf{V}}$ and $\sigma_{\mathbf{V}}$ through this chapter.

A remaining task is to select $a$ and $b$ in order to have a test with prescribed errors of type I and II equal to $\alpha$ and $\beta$, respectively. It turns out that $a$ and $b$ depend on the sample size and on the dimension of the space, but to avoid excessive notational burden, we do not make explicit this dependency, which will be analysed giving the expected number of required projections to reach the decision about the point we want to classify (related with the efficiency of the method).

Under $\mathbf{H}_0$, the only relevant quantity here is the value of $t = \|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma$, thus instead of assuming that we have a fixed point, we will replace the point $\mathbf{x}$ by a generic point on the Mahalanobis sphere with centre at $\boldsymbol{\mu}$ and radius $t$, for some $t > 0$. Being more precise, we will replace the point $\mathbf{x} \in \mathbb{R}^d$ by a random observation $\mathbf{X}$ coming from a $N_d(\boldsymbol{\mu}, \Sigma)$ distribution given that $\|\mathbf{X} - \boldsymbol{\mu}\|_\Sigma = t$. We begin with two assumptions and some notation:

(A1) $\mathbf{X}$ is a rv with distribution $N_d(\boldsymbol{\mu}, \Sigma)$.

(A2) $\mathbf{V}$ and $\mathbf{V}_1, \dots, \mathbf{V}_n$ are iid rv's with distribution $N_d(\mathbf{0}, I_d)$ which also are independent from the rv in (A1).

**Notation.** Under assumptions (A1) and (A2), denote

$$Y^{\mathbf{V}} := \frac{\mathbf{X}'\mathbf{V} - \mu_{\mathbf{V}}}{\sigma_{\mathbf{V}}}. \tag{3.2}$$

When $\mathbf{V} = \mathbf{V}_k$, we ease the notation writing $Y^k$ instead of $Y^{\mathbf{V}_k}$. The rv number of random projections which we need to decide if $\mathbf{X}$ is an outlier or not with respect to the distribution $N_d(\boldsymbol{\mu}, \Sigma)$ is denoted by $K^{a,b}(\Sigma)$. Thus, given $0 < a \le b$

$$K^{a,b}(\Sigma) := \inf\left\{k : |Y^k| < a \text{ or } |Y^k| > b\right\}. \tag{3.3}$$

However, when there is no possibility of confusion, or the specific values of $a, b$ or $\Sigma$ are not important, we omit those parameters and simplify, for instance, to $K$. Note that if $K$ is finite, then $Y^K$ is well defined.

For $s$, $t > 0$, let $\mathbf{X}$ be a rv with distribution $N_d(\boldsymbol{\mu}, \Sigma)$ and $\mathbf{Z}$ with distribution $N_d(\boldsymbol{\mu}, I_d)$, denote $f_t$ the pdf of $\mathbf{X}$ given that $\|\mathbf{X}\|_\Sigma = t$ and

$$\begin{array}{rcl} y^{\mathbf{V}} & := & (\mathbf{x}'\mathbf{V} - \mu_{\mathbf{V}})/\sigma_{\mathbf{V}}, \\ F(s,t) & := & \mathbf{P}\left(\frac{|\mathbf{Z}'\mathbf{V} - \mu_{\mathbf{V}}|}{\|\mathbf{V}\|} < s \,\middle|\, \|\mathbf{Z}\| = t\right), \\ F_\Sigma(a,b,t) & := & \mathbf{P}(|Y^K| > b \mid \|\mathbf{X}\|_\Sigma = t). \end{array} \tag{3.4}$$

Notice that the distribution of $Y^{\mathbf{V}}$ does not depend on $\boldsymbol{\mu}$ neither that on scale. Since our method relies on $Y^{\mathbf{V}}$, we can assume w.l.o.g. that $\boldsymbol{\mu} = \mathbf{0}$.

## 3.2 The distribution of the projections $Y^{\mathbf{V}}$

Let $s, t > 0$, we firstly show in Proposition 3.2.1 that $\mathbf{P}\left(|Y^{\mathbf{V}}| < s \mid \|\mathbf{X}\|_{\Sigma} = t\right)$ does not depend on the covariance matrix. This allows us to obtain en explicit expression for it, as it can be seen in Proposition 3.2.2.

**Proposition 3.2.1.** *Let $y \in [-t, t]$, with $t > 0$. Under assumptions* (A1) *and* (A2)*, we have that $\mathbf{P}\left(Y^{\mathbf{V}} < s \mid \|\mathbf{X}\|_{\Sigma} = t\right)$ does not depend on $\Sigma$.*

*Proof.* Write $\mathbf{X} = \Sigma^{1/2}\mathbf{X}_0$, with $\mathbf{X}_0$ a rv with distribution $N_d(\mathbf{0}, I_d)$, then

$$\begin{aligned}\mathbf{P}\left(Y^{\mathbf{V}} < s \mid \|\mathbf{X}\|_{\Sigma} = t\right) &= \mathbf{P}\left(\frac{\mathbf{X}_0'\Sigma^{1/2}\mathbf{V}}{\sigma_{\mathbf{V}}} < s \;\middle|\; \|\mathbf{X}_0\| = t\right)\\ &= \mathrm{E}\left[\mathbf{P}(\mathbf{X}_0'U_{\Sigma} < s \mid U_{\Sigma}, \|\mathbf{X}_0\| = t) \mid \|\mathbf{X}_0\| = t\right],\end{aligned}$$

where $U_{\Sigma} := \Sigma^{1/2}\mathbf{V}/\sigma_{\mathbf{V}}$. Then, it is enough to see that $\mathbf{P}(\mathbf{X}_0'U_{\Sigma} < s \, |U_{\Sigma}, \|\mathbf{X}_0\| = t)$ does not depend on $\Sigma$. Note that,

$$\mathbf{P}(\mathbf{X}_0'U_{\Sigma} < s \, |U_{\Sigma}, \|\mathbf{X}_0\| = t) = \mathbf{P}\left(\frac{\mathbf{X}_0'}{\|\mathbf{X}_0\|}U_{\Sigma} < \frac{s}{\|\mathbf{X}_0\|} \;\middle|\; U_{\Sigma}, \|\mathbf{X}_0\| = t\right).$$

Obviously $\mathbf{X}_0'/\left\|\mathbf{X}_0\right\|$ has a uniform distribution on $\Omega^{d-1}$. Since $\|U_{\Sigma}\| = 1$, we have that 2. in Proposition 2.5.2 gives

$$\begin{aligned}\mathbf{P}\left(\frac{\mathbf{X}_0'}{\|\mathbf{X}_0\|}U_{\Sigma} < \frac{s}{\|\mathbf{X}_0\|} \;\middle|\; U_{\Sigma}, \|\mathbf{X}_0\| = t\right) &= \mathbf{P}\left(\|U_{\Sigma}\| \, U_0^1 < \frac{s}{\|\mathbf{X}_0\|} \;\middle|\; U_{\Sigma}, \|\mathbf{X}_0\| = t\right)\\ &= \mathbf{P}\left(U_0^1 < \frac{s}{\|\mathbf{X}_0\|} \;\middle|\; U_{\Sigma}, \|\mathbf{X}_0\| = t\right), \qquad (3.5)\end{aligned}$$

where $U_0^1$ is the one dimensional marginal of a r.v. with uniform distribution on $\Omega^{d-1}$. Consequently, (3.5) does not depend on $U_{\Sigma}$ neither on $\Sigma$. □

**Proposition 3.2.2.** *Let $t > 0$ and $s \in \mathbb{R}$. Under assumptions* (A1) *and* (A2)*, we have*

$$\mathbf{P}\left(Y^{\mathbf{V}} < s \mid \|\mathbf{X}\|_{\Sigma} = t\right) = \begin{cases} 0 & \text{if } s \leq -t, \\ \frac{1}{2} + \operatorname{sign}(s)\frac{1}{2}\mathrm{I}_{s^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) & \text{if } s \in (-t, t), \\ 1 & \text{if } s \geq t, \end{cases}$$

*where* $\mathrm{I}_{s^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ *denotes the incomplete beta function.*

*Proof.* The result when $s \notin [-t, t]$ is trivial. If $s \in (-t, t)$, by Proposition 3.2.1 we have that, if $\mathbf{Z}$ is a rv with $N_d(\mathbf{0}, I_d)$ distribution, then

$$\begin{aligned} \mathbf{P}\left(Y^{\mathbf{V}} < s \mid \|\mathbf{X}\|_{\Sigma} = t\right) &= \mathbf{P}\left(\frac{\mathbf{Z}'\mathbf{V}}{\|\mathbf{V}\|} < s \,\middle|\, \|\mathbf{Z}\| = t\right) \\ &= \mathrm{E}\left[\mathbf{P}\left(\frac{\mathbf{Z}'\mathbf{V}}{\|\mathbf{V}\|} < s \,\middle|\, \mathbf{Z} = \mathbf{z}\right) \,\middle|\, \|\mathbf{Z}\| = t\right] \\ &= \int_{\Omega_t^{d-1}} \mathbf{P}\left(\frac{\mathbf{Z}'\mathbf{V}}{\|\mathbf{V}\|} < s \,\middle|\, \mathbf{Z} = \mathbf{z}\right) f_{\mathbf{Z} \mid \|\mathbf{Z}\|=t}(\mathbf{z})\, \mathrm{d}\mathbf{z}_{-d}. \end{aligned} \tag{3.6}$$

Denoting by $\mathbf{U}_0^1$ the one-dimensional marginal of a rv uniform on $\Omega^{d-1}$, by Proposition 2.5.2, given $\mathbf{z} \in \Omega_t^{d-1}$:

$$\mathbf{P}\left(\frac{\mathbf{Z}'\mathbf{V}}{\|\mathbf{V}\|} < s \,\middle|\, \mathbf{Z} = \mathbf{z}\right) = \mathbf{P}\left(\mathbf{U}_0^1 < st^{-1} \mid \mathbf{Z} = \mathbf{z}\right).$$

This expression does not depend on $\mathbf{z}$, therefore we can take this expression out of the integral in (3.6). By $f_{\mathbf{Z}| \ \|\mathbf{Z}\|=t}$ being a pdf and Lemma 2.5.4, the result is deduced when $s \in (-t, t)$. □

**Remark 3.2.3.** *From previous propositions, it is deduced that*

*i) The expression of* $\mathbf{P}\left(Y^{\mathbf{V}} < s \mid \|\mathbf{X}\|_{\Sigma} = t\right)$ *given in Proposition 3.2.2 depends on the dimension. Moreover, the computation of* $\mathrm{I}_{s^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ *is not too daunting and explicit expressions can be obtained. For instance, we have that:*

  - *If* $d = 2 : \mathrm{I}_{s^2/t^2}\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{2}{\pi}\sin^{-1}\left(st^{-1}\right)$.

  - *If* $d = 11 : \mathrm{I}_{s^2/t^2}\left(\frac{1}{2}, 5\right) = \frac{1}{128t^9}\left(35s^9 - 180s^7t^2 + 378s^5t^4 - 420s^3t^6 + 315st^8\right)$.

*ii) Proposition 3.2.1 shows that the distribution of* $Y^{\mathbf{V}}$ *given* $\|\mathbf{X}\|_{\Sigma}$ *does not depend on* $\Sigma$*. Moreover, from the proof of Proposition 3.2.2, we infer that if* $\Sigma = I_d$*, then* $\mathbf{X}'\mathbf{V}/\|\mathbf{V}\|$ *only depends on* $\|\mathbf{X}\|$ *but not on the precise value of* $\mathbf{X}$.

Figure 3.1 shows the graph of the probability $\mathbf{P}\left(\mathbf{X}'\mathbf{V}/\|\mathbf{V}\| < s \mid \|\mathbf{X}\| = t\right)$, computed as shown in Proposition 3.2.2 for different values of $n$ and $d$ with $\Sigma = I_d$ and $t = C_n^d$. We see that the larger the dimension the more concentrated the distribution on the central part of the interval $(-C_n^d, C_n^d)$. We appreciate no big differences when the sample size increases.

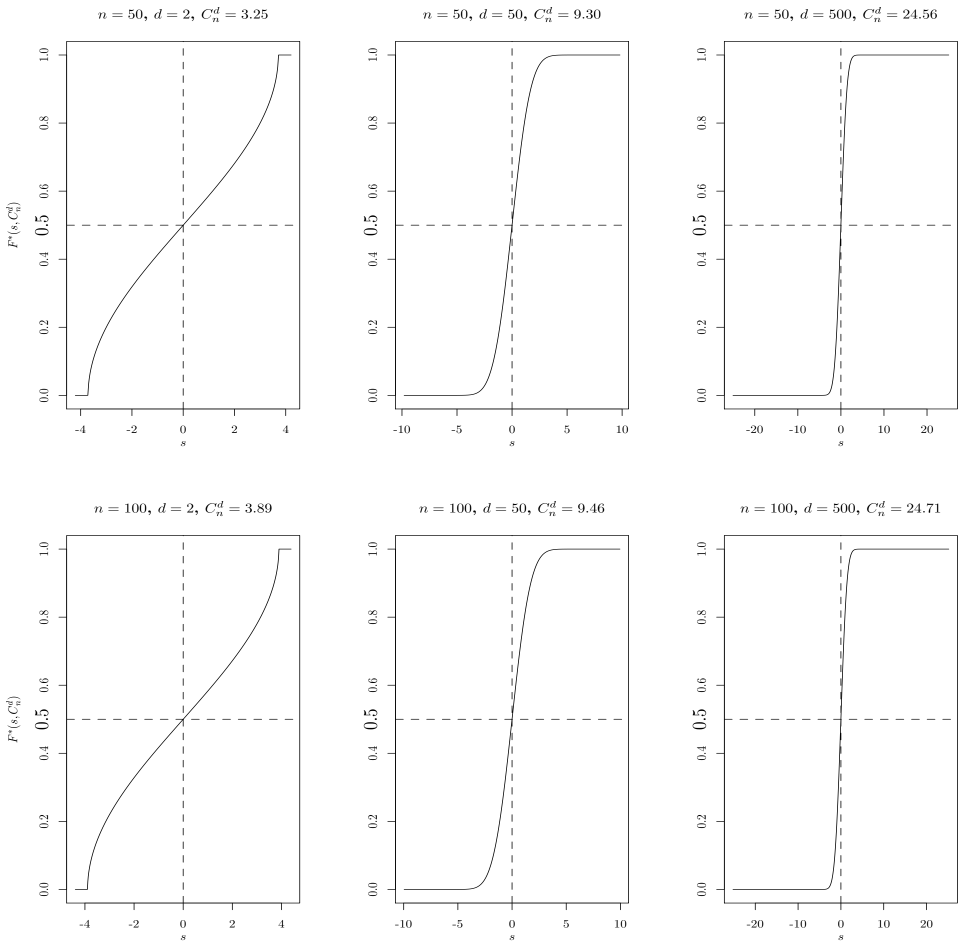


**Figure 3.1.:** Representation of $F^*(s, C_n^d) := \mathbf{P}\left(\frac{\mathbf{X}'\mathbf{V}}{\|\mathbf{V}\|} < s \,\middle|\, \|\mathbf{X}\| = C_n^d\right)$ for different values of $n$ and $d$.

Next example shows that the property ii) of Remark 3.2.3 may fail if $\Sigma \neq I_d$.

**Example 3.2.4.** *Let $n = 100$, $d = 2$, and consider the distribution $N_2(\mathbf{0}, \Sigma)$, where $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 10 \end{smallmatrix}\right)$. In this case, the set $\{\mathbf{x} : \|\mathbf{x}\|_\Sigma = t\}$, for any $t > 0$ is an ellipse on the plane with its major axis $10^{1/2}$ times larger than its minor axis. We also have that $C_n^d = 3.8925$. Take $\mathbf{x}_M = (0, 10^{1/2} C_n^d)'$ and $\mathbf{x}_m = (C_n^d, 0)'$, thus $\|\mathbf{x}_M\|_\Sigma = \|\mathbf{x}_m\|_\Sigma = C_n^d$. Let $\mathbf{V}$ be* *a rv with uniform distribution on $\Omega^1$, by Lemma 2.5.4 and taking into account that $\mathrm{I}_y\left(1/2, 1/2\right) = 2\sin^{-1}(\sqrt{y})/\pi$ for any $y > 0$, we have that, if $s = 3.8920$, then*

$$\mathbf{P}\left(\frac{|\mathbf{x}_M'\mathbf{V}|}{\sigma_\mathbf{V}} < s\right) = \frac{2}{\pi}\sin^{-1}\left(\frac{s}{(10(C_n^d)^2 - 9s^2)^{1/2}}\right) = 0.9678,$$
$$\mathbf{P}\left(\frac{|\mathbf{x}_m'\mathbf{V}|}{\sigma_\mathbf{V}} < s\right) = \frac{2}{\pi}\sin^{-1}\left(\frac{10^{1/2}s}{((C_n^d)^2 + 9s^2)^{1/2}}\right) = 0.9968.$$

*This difference between the probabilities increases with the quotient between the maximum and the minimum of the eigenvalues of the covariance matrix. For instance, if we repeat the computations using $\Sigma = \left(\begin{smallmatrix}\tau & 0\\ 0 & 10^3\tau\end{smallmatrix}\right)$, $\mathbf{x}_M = (0, 10^{3/2}\tau C_n^d)'$ and $\mathbf{x}_m = (\tau C_n^d, 0)'$ for some $\tau > 0$, then*

$$\mathbf{P}\left(\frac{|\mathbf{x}_M'\mathbf{V}|}{\sigma_\mathbf{V}} < s\right) = \frac{2}{\pi}\sin^{-1}\left(\frac{s}{(10^3(C_n^d)^2 - 999s^2)^{1/2}}\right) = 0.7013,$$
$$\mathbf{P}\left(\frac{|\mathbf{x}_m'\mathbf{V}|}{\sigma_\mathbf{V}} < s\right) = \frac{2}{\pi}\sin^{-1}\left(\frac{10^{3/2}s}{((C_n^d)^2 + 999s^2)^{1/2}}\right) = 0.9997.$$

Given $s > 0$, Example 3.2.4 leads us to study the variation on $\mathbf{x}$ of the function $\mathbf{P}\left(y^\mathbf{V} < s\right)$ on the set $\{\mathbf{x} : \|\mathbf{x}\|_\Sigma = t\}$ for a given $t > 0$, which is given in Proposition 3.2.7. To prove it, we need some previous results.

Notice that, in Lemma 3.2.5, in the sets $\mathcal{R}_i$, defined there, we assume that $v_2 \neq 0$ just to simplify the writing. It is enough that $v_i \neq 0$ for some $i \in \{2, \ldots, d\}$.

**Lemma 3.2.5.** *Given $0 < \sigma_1 \leq \ldots \leq \sigma_d$, $\mathbf{x} = (x_1, \ldots, x_d)'$ and $\mathbf{v} = (v_1, \ldots, v_d)'$, let $0 \neq t = \|\mathbf{x}\|$, then the map $\mathcal{H} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ given by*

$$\mathcal{H}(v_1, \ldots, v_d) := \left(\frac{x_1 v_1 + \psi}{(\sigma_1^2 v_1^2 + \varphi)^{1/2}}, v_2, \ldots, v_d\right),$$

*where $\psi(= \psi_\mathbf{v}) := x_2 v_2 + \cdots + x_d v_d$ and $\varphi(= \varphi_\mathbf{v}) := \sigma_2^2 v_2^2 + \cdots + \sigma_d^2 v_d^2$, is injective when restricted to each of the following regions:*

$$\mathcal{R}_1 := \left\{\mathbf{v} : v_1 < \tfrac{x_1\varphi_\mathbf{v}}{\sigma_1^2\psi_\mathbf{v}}, \psi_\mathbf{v} > 0, v_2 \neq 0\right\}; \quad \mathcal{R}_2 := \left\{\mathbf{v} : v_1 < \tfrac{x_1\varphi_\mathbf{v}}{\sigma_1^2\psi_\mathbf{v}}, \psi_\mathbf{v} < 0, v_2 \neq 0\right\};$$
$$\mathcal{R}_3 := \left\{\mathbf{v} : v_1 > \tfrac{x_1\varphi_\mathbf{v}}{\sigma_1^2\psi_\mathbf{v}}, \psi_\mathbf{v} > 0, v_2 \neq 0\right\}; \quad \mathcal{R}_4 := \left\{\mathbf{v} : v_1 > \tfrac{x_1\varphi_\mathbf{v}}{\sigma_1^2\psi_\mathbf{v}}, \psi_\mathbf{v} < 0, v_2 \neq 0\right\}.$$

*Proof.* The projection of the last $d - 1$ components of $\mathcal{H}$ coincides with the identity function, which is obviously injective. Therefore we assume that $v_2, \ldots, v_d$ are fixed and study the monotonicity of the function $\mathcal{H}^1(v_1, \ldots, v_d) = (x_1 v_1 + \psi)/(\sigma_1^2 v_1^2 + \varphi)^{1/2}$. We have that

$$\begin{aligned}\frac{\partial \mathcal{H}^1(v_1, \ldots, v_d)}{\partial v_1} &= \frac{x_1\sqrt{\sigma_1^2 v_1^2 + \varphi} - (x_1 v_1 + \psi)\sigma_1^2 v_1(\sigma_1 v_1^2 + \varphi)^{-1/2}}{\sigma_1^2 v_1^2 + \varphi} \\ &= \frac{\sigma_1(x_1\varphi/\sigma_1^1 - \psi v_1)}{(\sigma_1^2 v_1^2 + \varphi)^{3/2}}.\end{aligned}$$

Then $\frac{\partial \mathcal{H}^1(v_1,\ldots,v_d)}{\partial V_1} = 0$ if and only if $v_1 = \frac{x_1\varphi}{\sigma_1^2\psi}$, or $\psi = 0$ and $x_1 = 0$. It is easy to check that $\mathcal{H}^1$ is strictly increasing on $\mathcal{R}_1$ and $\mathcal{R}_4$, while it is strictly decreasing on $\mathcal{R}_2$ and $\mathcal{R}_3$. Consequently, the result is proven. □

For the sake of simplicity we denote the function $\mathcal{H}$ restricted to the regions $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$ as $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ and $\mathcal{H}_4$, respectively.

**Corollary 3.2.6.** *With the notation above introduced, we have that:*

- *On* $\mathcal{H}(\mathcal{R}_1) = \left(-\frac{|x_1|}{\sigma_1}, \sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}\right) \times \{(v_2, \ldots, v_d)' : \psi_{\mathbf{v}} > 0\}$*:*

$$\mathcal{H}_1^{-1}(y, v_2, \ldots, v_d) = \begin{cases} (h_+(y), v_2, \ldots, v_d)\,, & \textit{if } -\frac{|x_1|}{\sigma_1} < y < \frac{|x_1|}{\sigma_1}, \\ (h_-(y), v_2, \ldots, v_d)\,, & \textit{if } \frac{|x_1|}{\sigma_1} < y < \sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}. \end{cases}$$

- *On* $\mathcal{H}(\mathcal{R}_2) = \left(-\sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}, -\frac{|x_1|}{\sigma_1}\right) \times \{(v_2, \ldots, v_d)' : \psi_{\mathbf{v}} < 0\}$*:*

$$\mathcal{H}_2^{-1}(y, v_2, \ldots, v_d) = (h_+(y), v_2, \ldots, v_d)\,, \textit{ for } -\sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}} < y < -\frac{|x_1|}{\sigma_1}.$$

- *On* $\mathcal{H}(\mathcal{R}_3) = \left(\frac{|x_1|}{\sigma_1}, \sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}\right) \times \{(v_2, \ldots, v_d)' : \psi_{\mathbf{v}} > 0\}$*:*

$$\mathcal{H}_3^{-1}(y, v_2, \ldots, v_d) = (h_+(y), v_2, \ldots, v_d)\,, \textit{ for } \frac{|x_1|}{\sigma_1} < y < \sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}.$$

- *On* $\mathcal{H}(\mathcal{R}_4) = \left(-\sqrt{\frac{x_1^2}{\sigma_1^2} + \frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}}, \frac{|x_1|}{\sigma_1}\right) \times \{(v_2, \ldots, v_d)' : \psi_{\mathbf{v}} < 0\}$*:*

$$\mathcal{H}_4^{-1}(y,v_2,\dots,v_d) = \begin{cases} (h_-(y),v_2,\dots,v_d)\,, & \textit{if } \; -\sqrt{\frac{x_1^2}{\sigma_1^2}+\frac{\psi_{\mathbf{v}}^2}{\varphi_{\mathbf{v}}}} < y < -\frac{|x_1|}{\sigma_1}, \\ (h_+(y),v_2,\dots,v_d)\,, & \textit{if } \; -\frac{|x_1|}{\sigma_1} < y < \frac{|x_1|}{\sigma_1}, \end{cases}$$

*where* $h_\pm(y) := \left(x_1\psi_{\mathbf{v}} \pm |y|\sqrt{x_1^2\varphi_{\mathbf{v}} + \sigma_1^2\left(\psi_{\mathbf{v}}^2 - y^2\varphi_{\mathbf{v}}\right)}\right) / \left(\sigma_1^2 y^2 - x_1^2\right)$.

*Proof.* From Lemma 3.2.5, the explicit expression of the inverse of $\mathcal{H}$ is:

$$\mathcal{H}^{-1}(y,v_2,\dots,v_d) = (h_\pm(y),v_2,\dots,v_d)\,.$$

It remains to determine when the first coordinate of $\mathcal{H}^{-1}(y,v_2,\dots,v_d)$ is $h_+$ or $h_-$. Suppose $\psi_{\mathbf{v}} > 0$ and $y > |x_1|/\sigma_1$, (the rest of the cases are analogous) then

$$\frac{x_1\psi_{\mathbf{v}} + |y|\sqrt{x_1^2\varphi_{\mathbf{v}} + \sigma_1^2\left(\psi_{\mathbf{v}}^2 - y^2\varphi_{\mathbf{v}}\right)}}{\sigma_1^2 y^2 - x_1^2} > \frac{x_1\psi_{\mathbf{v}} - |y|\sqrt{x_1^2\varphi_{\mathbf{v}} + \sigma_1^2\left(\psi_{\mathbf{v}}^2 - y^2\varphi_{\mathbf{v}}\right)}}{\sigma_1^2 y^2 - x_1^2}.$$

Hence by the definition of the regions $\mathcal{R}_i$: $(h_-(y),v_2,\dots,v_d) = \mathcal{H}_1^{-1}(y,v_2,\dots,v_d)$, if $(y,v_2,\dots,v_d) \in \mathcal{R}_1$, and $(h_+(y),v_2,\dots,v_d) = \mathcal{H}_3^{-1}(y,v_2,\dots,v_d)$, if $(y,v_2,\dots,v_d) \in \mathcal{R}_3$. □

Proposition 3.2.7 gives an expression of the cdf of the standardized random projection of a given $d$-dimensional vector. Remember that, by Proposition 3.2.1, this distribution does not depend on such vector if $\Sigma = I_d$ and it may depend if $\Sigma \neq I_d$ (see Example 3.2.4).

**Proposition 3.2.7.** *Let* $\mathbf{x} = (x_1,\dots,x_d)' \in \mathbb{R}^d$. *Assume that* $\Sigma$ *is diagonal with eigenvalues* $0 < \sigma_1^2 \leq \ldots \leq \sigma_d^2$ *and that* $0 \neq t := \|\mathbf{x}\|_\Sigma$. *If* $\mathbf{V}$ *has* $N_d(\mathbf{0}, I_d)$ *distribution, then the distribution of* $y^{\mathbf{V}}$ *is supported by* $[-t,t]$ *and*

$$\mathbf{P}(y^{\mathbf{V}} < s) = \begin{cases} \tau \displaystyle\int_{A_-} \Delta(s) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2}\, \mathrm{d}\mathbf{v}_{-1} & \textit{if } \; -t < s < -|x_1|/\sigma_1, \\ \frac{1}{2} - \operatorname{sign}(s)\tau \displaystyle\int_{A_+} \Delta(s) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2}\, \mathrm{d}\mathbf{v}_{-1} & \textit{if } \; -|x_1|/\sigma_1 \leq s \leq |x_1|/\sigma_1, \\ 1 - \tau \displaystyle\int_{A_+} \Delta(s) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2}\, \mathrm{d}\mathbf{v}_{-1} & \textit{if } \; |x_1|/\sigma_1 < s < t, \end{cases}$$

*where* $\Delta(s) := \operatorname{erf}\left(h_+(s)/\sqrt{2}\right) - \operatorname{erf}\left(h_-(s)/\sqrt{2}\right)$, $\mathbf{v}_{-1} = (v_2,\dots,v_d)'$, $\tau := \left(2^{\frac{d+3}{2}}\pi^{\frac{d-1}{2}}\right)^{-1}$, $A_+ := \{\mathbf{v}_{-1} : \psi_{\mathbf{v}} > 0\}$ *and* $A_- := \{\mathbf{v}_{-1} : \psi_{\mathbf{v}} < 0\}$ *with* $\psi_{\mathbf{v}}$ *and* $\varphi_{\mathbf{v}}$ *defined as in Lemma 3.2.5 and* $h_\pm(\cdot)$ *as in Corollary 3.2.6.*

*Proof.* To ease the notation, we omit the sub-index $\mathbf{v}$ in $\psi_{\mathbf{v}}$ and $\varphi_{\mathbf{v}}$. Due to the symmetry of the distribution of $\mathbf{V}$, we assume that $x_i \geq 0$ for $i = 1, \ldots, d$. It is clear that $y^{\mathbf{V}} \in [-t, t]$. Take the transformation $\mathcal{H}$ defined on Lemma 3.2.5.

We have that if $B, B_0$ are two Borel sets and $V^1, \ldots, V^4$ are rv's such that the distribution of $V^i$ is that of $V$ given that $V \in \mathcal{R}_i$, for $i = 1, \ldots, 4$, denoting the conditional probability of $B_0$ given $B$ by $\mathbf{P}(B_0|B)$ with $\mathbf{P}(B) > 0$, then:

$$\begin{aligned}
\mathbf{P}(Y \in B) &= \mathbf{P}(\mathcal{H}(V) \in B) \\
&= \sum_{i=1}^{4} \mathbf{P}(V \in \mathcal{R}_i)\mathbf{P}(\mathcal{H}(V) \in B | V \in \mathcal{R}_i) \\
&= \sum_{i=1}^{4} \mathbf{P}(V \in \mathcal{R}_i)\mathbf{P}(\mathcal{H}_{i,1}(V^i) \in B), \qquad (3.7)
\end{aligned}$$

where, as stated, $\mathcal{H}_i = (\mathcal{H}_{i,1}, \ldots, \mathcal{H}_{i,d})$ is the restriction of $\mathcal{H}$ to the set $\mathcal{R}_i$, $i = 1, \ldots, 4$. Since all $\mathcal{H}_i$ are injective and derivable we have that

$$\mathbf{P}(\mathcal{H}_{i,1}(V^i) \in B) = \int_B \int_{\mathbb{R}^{d-1}} f_{V^i}(\mathcal{H}_i^{-1}(s, \mathbf{v}_{-1})) |J_{\mathcal{H}_i}(s, \mathbf{v}_{-1})| \, \mathrm{d}\mathbf{v}_{-1} \, \mathrm{d}s, \qquad (3.8)$$

where $f_{\mathbf{V}^i}$ is the pdf of the rv $\mathbf{V}^i$. We trivially have that

$$|J_{\mathcal{H}_i}(s, \mathbf{v}_{-1})| = \left| \frac{(\partial \mathcal{H}_i)_1^{-1}(s)}{\partial s} \right|,$$

and

$$f_{V^i}(\mathcal{H}_i^{-1}(s, \mathbf{v}_{-1})) = \frac{1}{\mathbf{P}(V \in \mathcal{R}_i)} f_V(\mathcal{H}_i^{-1}(s, \mathbf{v}_{-1})) 1_{\mathcal{R}_i}(\mathcal{H}_i^{-1}(y, \mathbf{v}_{-1})).$$

This expression jointly with (3.7) and (3.8) give

$$\mathbf{P}(Y \in B) = \int_B \sum_{i=1}^{4} \int_{\mathbb{R}^{d-1}} f_V(\mathcal{H}_i^{-1}(s, \mathbf{v}_{-1})) |J_{\mathcal{H}_i}(s, \mathbf{v}_{-1})| \, \mathrm{d}\mathbf{v}_{-1} \, \mathrm{d}s,$$

where we have used the fact that, by definition, $1_{\mathcal{R}_i}((\mathcal{H}_i)^{-1}(s, \mathbf{v}_{-1})) = 1$.

We study the sign of the determinant of the Jacobian. By Corollary 3.2.6,

$$\begin{aligned}
\left| \frac{(\mathcal{H}_i)_1^{-1}(s)}{\partial s} \right| &= \left| \frac{\partial h_{\pm}(s)}{\partial s} \right| \\
&= \frac{\mp \frac{x_1^4}{\sigma_1^4}\varphi + \frac{x_1^2}{\sigma_1^2}(\pm s^2 \varphi \mp \psi^2) \mp s^2 \psi^2 - 2\frac{x_1}{\sigma_1}\psi\varphi^{1/2} s \left( \frac{x_1^2}{\sigma_1^2} + \frac{\psi^2}{\varphi} - s^2 \right)^{1/2}}{\sigma_1 (s^2 - x_1^2/\sigma_1^2)^2 (x_1^2/\sigma_1^2 \varphi + \psi^2 - s^2 \varphi)^{1/2}},
\end{aligned}$$

where the signs depend on the particular index and $s$.

We have that $\left|\frac{\partial h_{\pm}(s)}{\partial s}\right| = 0$ only when $s \in \{0, \pm x_1/\sigma_1\}$. As those values are not in the mentioned regions, we can state that:

$$\begin{aligned}
&\text{- If } s > 0, \quad \partial h_+(y)/\partial s < 0 \quad \text{and} \quad \partial h_-(s)/\partial s > 0.\\
&\text{- If } s < 0, \quad \partial h_+(y)/\partial s > 0 \quad \text{and} \quad \partial h_-(s)/\partial s < 0.
\end{aligned}$$

Take $B = (-\infty, r]$, with $r \in (-t, -|x_1|/\sigma_1)$, then

$$\begin{aligned}
\mathbf{P}(y^{\mathbf{V}} < r) = \int_{A^+} \Bigg( \int_{-t}^{r} & f_V(\mathcal{H}_1^{-1}(s, \mathbf{v}_{-1}))|J_{\mathcal{H}_1}(s, \mathbf{v}_{-1})| \, \mathrm{d}s \\
& + \int_{-t}^{r} f_V(\mathcal{H}_3^{-1}(s, \mathbf{v}_{-1}))|J_{\mathcal{H}_3}(s, \mathbf{v}_{-1})| \, \mathrm{d}s \Bigg) \, \mathrm{d}\mathbf{v}_{-1} \\
+ \int_{A^-} \Bigg( \int_{-t}^{r} & f_V(\mathcal{H}_2^{-1}(s, \mathbf{v}_{-1}))|J_{\mathcal{H}_2}(s, \mathbf{v}_{-1})| \, \mathrm{d}s \\
& + \int_{-t}^{r} f_V(\mathcal{H}_4^{-1}(s, \mathbf{v}_{-1}))|J_{\mathcal{H}_4}(s, \mathbf{v}_{-1})| \, \mathrm{d}s \Bigg) \, \mathrm{d}\mathbf{v}_{-1}. \qquad (3.9)
\end{aligned}$$

We have $f_V(\mathcal{H}_i^{-1}(s, \mathbf{v}_{-1}))|J_{\mathcal{H}_i}(s, \mathbf{v}_{-1})| = 0$ when $r \in \left(-t, -\sqrt{s_1^2 + \psi^2/\phi}\right)$ for $i = 1, \ldots, 4$ and, using Corollary 3.2.6,

$$\begin{aligned}
\mathbf{P}(y^{\mathbf{V}} < r) = \frac{1}{(2\pi)^d} \int_{A^-} \Bigg( \Bigg( & \int_{-\sqrt{x_1^2/\sigma_1^2 + \psi^2/\phi}}^{r} e^{-\frac{1}{2}(h_+^2(s) + v_2^2 + \cdots + v_d^2)} \frac{\partial h_+(s)}{\partial y} \, \mathrm{d}s \\
& - \int_{-\sqrt{x_1^2/\sigma_1^2 + \psi^2/\phi}}^{r} e^{-\frac{1}{2}(h_-^2(s) + v_2^2 + \cdots + v_d^2)} \frac{\partial h_-(s)}{\partial s} \, \mathrm{d}s \Bigg) \, \mathrm{d}\mathbf{v}_{-1}.
\end{aligned}$$

From $h_-\left(-\sqrt{x_1^2/\sigma_1^2 + \psi^2/\phi}\right) = h_+\left(-\sqrt{x_1^2/\sigma_1^2 + \psi^2/\phi}\right)$, the result is obtained. The case $-|x_1|/\sigma_1 < r < 0$ is analogous and the cases when $r > 0$ are deduced by symmetry. □

The expression obtained in Proposition 3.2.7 can be written in an explicit way when $d = 2$.

**Corollary 3.2.8.** *Under the same assumptions as in Proposition 3.2.7, if we additionally assume that $d = 2$ and $\mathbf{V}$ has $N_2(\mathbf{0}, I_2)$ distribution, then the distribution of $y^{\mathbf{V}}$ is supported by $[-t, t]$ and*

$$\mathbf{P}(y^{\mathbf{V}} < s) = \begin{cases} \frac{1}{2\pi}\left(\tan^{-1}(\tilde{h}_+(s)) - \tan^{-1}(\tilde{h}_-(s))\right) & \textit{if} \quad -t < s < -\frac{|x_1|}{\sigma_1}, \\ \frac{1}{2} + \operatorname{sign}(s)\frac{1}{2\pi}\left(\tan^{-1}(\tilde{h}_-(s)) - \tan^{-1}(\tilde{h}_+(s))\right) & \textit{if} \quad -\frac{|x_1|}{\sigma_1} < s < \frac{|x_1|}{\sigma_1}, \\ 1 - \frac{1}{2\pi}\left(\tan^{-1}(\tilde{h}_+(s)) - \tan^{-1}(\tilde{h}_-(s))\right) & \textit{if} \quad \frac{|x_1|}{\sigma_1} < s < t, \end{cases}$$

*where* $\tilde{h}_{\pm}(s) = \left(x_1 x_2 \pm |s|\left(\sigma_2^2 x_1^2 + \sigma_1^2(x_2^2 - \sigma_2^2 s^2)\right)^{1/2}\right) / \left(\sigma_1^2 s^2 - x_1^2\right).$

*Proof.* Without loss of generality assume that $x_2 > 0$. Since $d = 2$, $A_+ := \{v_2 : v_2 > 0\}$ and $A_- := \{v_2 : v_2 < 0\}$. Let us consider the case $r \in (-t, -|x_1|/\sigma_1)$, the other cases are analogous. Expression (3.9) now is

$$\begin{aligned}\mathbf{P}(Y < r) = &\int_{A^+}\int_{-t}^{r} \left(f_V(\mathcal{H}_1^{-1}(s), v_2)|J_{\mathcal{H}_1}(s, v_2)| + f_V(\mathcal{H}_3^{-1}(s), v_2)|J_{\mathcal{H}_3}(s, v_2)|\right) \mathrm{d}s\, \mathrm{d}v_2 \\ &+ \int_{A^-}\int_{-t}^{r} \left(f_V(\mathcal{H}_2^{-1}(s), v_2)|J_{\mathcal{H}_2}(s, v_2)| + f_V(\mathcal{H}_4^{-1}(s), v_2)|J_{\mathcal{H}_4}(s, v_2)|\right) \mathrm{d}s\, \mathrm{d}v_2.\end{aligned}$$

For $i = 1, \ldots, 4$, the determinant of the Jacobian is $|J_{\mathcal{H}_i}(s, v_2)| = \left|v_2 \frac{\partial \tilde{h}_\pm(s)}{\partial s}\right|$. By Corollary 3.2.6 and since $r \in (-t, -|x_1|/\sigma_1)$, we have

$$\begin{aligned}\mathbf{P}(y^{\mathbf{V}} < r) &= \frac{1}{2\pi}\int_{-t}^{r}\left(\frac{\partial \tilde{h}_+(s)}{\partial s}\int_{-\infty}^{0} v_2 e^{-\frac{v_2^2}{2}(1+\tilde{h}_+^2(s))}\, \mathrm{d}v_2 - \frac{\partial \tilde{h}_-(s)}{\partial y}\int_{-\infty}^{0} v_2 e^{-\frac{v_2^2}{2}(1+\tilde{h}_-^2(s))}\, \mathrm{d}v_2\right) \mathrm{d}s \\ &= \frac{1}{2\pi}\left(\int_{-t}^{r}\frac{\partial \tilde{h}_+(s)}{\partial s}\frac{1}{1+\tilde{h}_+^2(s)}\, \mathrm{d}s - \int_{-t}^{r}\frac{\partial \tilde{h}_-(s)}{\partial s}\frac{1}{1+\tilde{h}_-^2(s)}\, \mathrm{d}s\right).\end{aligned}$$

The proof is concluded solving these integrals and applying $\tilde{h}_-(-t) = \tilde{h}_+(-t)$. □

### 3.2.1 Several projections

Our decision will be based on the value of $Y^K$. Most of the times we will have $K > 1$ and, therefore, the distribution of $Y^K$ will depend on the joint distribution of the one-dimensional rv's $Y^1, \ldots, Y^k, \ldots$ In this subsection we pay some attention to this problem.

We begin with Proposition 3.2.10, where we show that the projections are not conditionally independent given the Mahalanobis norm of the point which is investigated unless $\Sigma = I_d$, meanwhile they are trivially independent when the point is given. This fact is stated for further reference in Lemma 3.2.9. After this we will obtain in Proposition 3.2.11 an expression of the probability to declare a generic point in $\Omega_\Sigma^{d-1}(t)$ as an outlier.

**Lemma 3.2.9.** *Under assumption (A2), $Y^{\mathbf{V}_1}, \ldots, Y^{\mathbf{V}_k}$ are conditionally independent given* $\mathbf{X}$.

**Proposition 3.2.10.** *Under assumptions* (A1) *and* (A2) *the rv's $Y^{\mathbf{V}_1}, \ldots, Y^{\mathbf{V}_k}$ are conditionally independent given $\|\mathbf{X}\|_\Sigma$ if and only if $\Sigma = I_d$.*

*Proof.* Let $t > 0$ and $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = t$. Define $\delta(\mathbf{x}) := \mathbf{P}(Y^{\mathbf{V}} < a|\mathbf{X} = \mathbf{x}) - \mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t)$. From Proposition 3.2.7 it follows that the map $\mathbf{x} \mapsto \mathbf{P}(Y^{\mathbf{V}} < a|\mathbf{X} = \mathbf{x})$

is continuous and not constant on $\mathbf{x}$ for any real $a$ if $\Sigma \neq I_d$. Thus, if $t > 0$, then $\int \delta^2(\mathbf{x})\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(d\mathbf{x}) > 0$. However, by definition of $\delta(\mathbf{x})$,

$$\begin{aligned}\mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t) &= \int \mathbf{P}\left(Y^{\mathbf{V}} < a|\mathbf{X} = \mathbf{x}\right)\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}) \\ &= \mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t) + \int \delta(\mathbf{x})\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}),\end{aligned}$$

and, consequently, $\int \delta(\mathbf{x})\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(d\mathbf{x}) = 0$. Denote $g_t(a) := \mathbf{P}(Y^{\mathbf{V}_1} < a, Y^{\mathbf{V}_2} < a \,|\|\mathbf{X}\|_\Sigma = t)$, then, by Lemma 3.2.9.

$$\begin{aligned}g_t(a) &= \int \mathbf{P}\left(Y^{\mathbf{V}_1} < a, Y^{\mathbf{V}_2} < a|\mathbf{X} = \mathbf{x}\right)\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}) \\ &= \int \mathbf{P}(Y^{\mathbf{V}} < a|\mathbf{X} = \mathbf{x})^2\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}) \\ &= \int \left(\mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t)^2 + \delta^2(\mathbf{x}) + 2\delta(\mathbf{x})\mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t)\right)\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}) \\ &= \mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t)^2 + \int \delta^2(\mathbf{x})\mathbf{P}_{\mathbf{X}|\|\mathbf{X}\|_\Sigma=t}(\mathrm{d}\mathbf{x}) \\ &> \mathbf{P}(Y^{\mathbf{V}} < a|\|\mathbf{X}\|_\Sigma = t)^2.\end{aligned}$$

However, if $\Sigma = I_d$, by *ii)* in Remark 3.2.3, we have that $\mathbf{P}(Y^{\mathbf{V}} < a|\mathbf{X} = \mathbf{x})$ is constant on $\mathbf{x}$ and the same reasoning shows the independence in this case. □

### 3.2.2 The testing problem revisited

We provide now some characteristics of the distribution of $Y^K$ which will be crucial in the determination of the constants $a$ and $b$. Given $\alpha \in (0, 1)$, the intended error of type I, our goal is to calculate values of $a$ and $b$ with $0 < a \leq b < C_n^d(\delta)$ such that $\mathbf{P}(K < \infty) = 1$, and the probability of declaring a point $\mathbf{X}$ as outlier when it is not it is less or equal than $\alpha$, i.e. $a$ and $b$ should verify

$$\sup_{t \leq C_n^d(\delta)} F_\Sigma(a, b, t) = \alpha. \tag{3.10}$$

Proposition 3.2.11 gives an expression for the term $F_\Sigma(a, b, t)$ which appears in (3.10). The reason to exclude the case $a = 0$ in the rest of this chapter, is that for any $\mathbf{x}$ we have that $\mathbf{P}(|y^{\mathbf{v}}| \leq 0) = 0$ a.s. what, issues aside, would lead to $\mathbf{P}(K < \infty) = 0$.

**Proposition 3.2.11.** *Under assumptions* (A1) *and* (A2)*, suppose that $a$, $b$ and $t$ are strictly positive constants such that $a \leq b$, then*

$$F_\Sigma(a,b,t) = \begin{cases} 0, & \textit{if } b \geq t, \\ \displaystyle\int_{\Omega_\Sigma^{d-1}(t)} \frac{\mathbf{P}(|y^{\mathbf{V}}| > b)}{1 - \mathbf{P}(|y^{\mathbf{V}}| \in (a,b))} f_t(\mathbf{x})\,\mathrm{d}\mathbf{x}, & \textit{if } 0 < a \leq b < t. \end{cases}$$

*Proof.* By definition, see (3.4),

$$\begin{aligned} F_\Sigma(a,b,t) &= \sum_{k=1}^{\infty} \mathbf{P}(\{|Y^K| > b\} \cap \{K = k\} \mid \|\mathbf{X}\|_\Sigma = t) \\ &= \sum_{k=1}^{\infty} \mathbf{P}\left(\text{for } i = 1,\dots,k-1;\ |Y^{\mathbf{V}_i}| \in [a,b] \text{ and } |Y^{\mathbf{V}_k}| > b \middle| \|\mathbf{X}\|_\Sigma = t\right). \end{aligned}$$

Since $a > 0$, $\mathbf{P}(|Y^K| > b| \|\mathbf{X}\|_\Sigma = t) = \mathrm{E}\left(\mathbf{P}(Y^K > b\,|\mathbf{X})| \|\mathbf{X}\|_\Sigma = t\right)$. By Lemma 3.2.9

$$\begin{aligned} F_\Sigma(a,b,t) &= \mathrm{E}\left( \sum_{k=1}^{\infty} \left(\mathbf{P}(|Y^k| \in (a,b)|\mathbf{X})\right)^{k-1} \mathbf{P}(|Y| > b\,|\mathbf{X}) \middle| \|\mathbf{X}\|_\Sigma = t\right) \\ &= \mathrm{E}\left( \frac{\mathbf{P}(|Y^{\mathbf{V}}| > b)}{1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b))} \middle| \|\mathbf{X}\|_\Sigma = t\right). \end{aligned}$$

□

By Propositions 3.2.2 and 3.2.10, if $\Sigma = I_d$, Proposition 3.2.11 simplifies as shown in Corollary 3.2.12.

**Corollary 3.2.12.** *Under the assumptions of Proposition 3.2.11, if, additionally,* $\mathbf{X}$ *is a rv with distribution* $N_d(\mathbf{0}, I_d)$*, then,*

$$\mathbf{P}\left(|Y^K| > b\,|\,\|\mathbf{X}\| = t\right) = \begin{cases} 0, & \textit{if } a > 0, b \geq t, \\ \frac{1-\mathrm{I}_{b^2/t^2}\left(\frac{1}{2},\frac{d-1}{2}\right)}{1+\mathrm{I}_{a^2/t^2}\left(\frac{1}{2},\frac{d-1}{2}\right)-\mathrm{I}_{b^2/t^2}\left(\frac{1}{2},\frac{d-1}{2}\right)}, & \textit{if } 0 < a \leq b < t. \end{cases}$$

Proposition 3.2.11 gives the mean value of the probability to declare a point as non-outlier given its Mahalanobis norm. However, this probability depends on the precise location of the point, and it varies between the extremes shown in Theorem 3.2.14.

For further reference we state the following lemma, whose proof is trivial.

**Lemma 3.2.13.** *Let* $v$*,* $w$*,* $v^1$ *and* $w^1$ *be such that* $0 < v < v^1$ *and* $0 < w^1 < w$*, then*

$$\frac{v}{v+w} < \frac{v^1}{v^1 + w^1}.$$

**Theorem 3.2.14.** *Let* $\Sigma$ *be a diagonal matrix with ordered eigenvalues* $0 < \sigma_1^2 \leq \ldots \leq \sigma_d^2$*; let* $\mathbf{x} \in \mathbb{R}^d$ *with* $0 \neq t = \|\mathbf{x}\|_\Sigma$*. Assume that* $a$*,* $b$ *are strictly positive constants such that*

*$a \le b < t$, and consider $Y^K$ as in (3.2). Under assumption (A2), we have that the map $\mathbf{x} \mapsto \mathbf{P}(|Y^K| > b|\mathbf{X} = \mathbf{x})$ attains a minimum and a maximum when $\mathbf{x} = (\pm t\sigma_1, 0, \dots, 0)$ and $\mathbf{x} = (0, \dots, 0, \pm t\sigma_d)$ respectively.*

*Proof.* The same reasoning as in Proposition 3.2.11 gives

$$\mathbf{P}(|Y^K| > b|\mathbf{X} = \mathbf{x}) = \frac{1 - \mathbf{P}(|y^{\mathbf{V}}| < b)}{1 - \mathbf{P}(|y^{\mathbf{V}}| < b) + \mathbf{P}(|y^{\mathbf{V}}| < a)}.$$

Choose a basis whose first element is $\mathbf{x}/\|\mathbf{x}\|$. Write $\mathbf{V} = (V_1, \dots, V_d)$, then

$$\mathbf{P}(|y^{\mathbf{V}}| < s) = \mathbf{P}\left(\frac{\|\mathbf{x}\|\,|V_1|}{\sigma_{\mathbf{V}}} < s\right) = \mathbf{P}\left(\frac{|V_1|}{\sigma_{\mathbf{V}}} < \frac{s}{\|\mathbf{x}\|}\right).$$

Hence $\mathbf{P}(|y^{\mathbf{V}}| < b)$ is strictly decreasing in $\|\mathbf{x}\|$ and we only have to search the extreme values of $\|\mathbf{x}\|$, or equivalently, of $\|\mathbf{x}\|^2$ under the restriction $t^2 = \|\mathbf{x}\|_\Sigma^2 = \sum_{i=1}^d \sigma_i^{-2} x_i^2$, if we write $\mathbf{x}$ on the basis of the eigenvectors of $\Sigma$. But it is clear that the extreme values of $\|\mathbf{x}\|$ are attained in $(t\sigma_1, 0, \dots, 0)$ and $(0, \dots, 0, t\sigma_d)$. □

From the proof of Theorem 3.2.14, Proposition 3.2.15 follows and states that the probability of declaring a point as outlier is strictly increasing with respect to its Mahalanobis norm.

**Proposition 3.2.15.** *Under the same hypotheses of Theorem 3.2.14, the map $t \mapsto \mathbf{P}(|Y^K| > b\,|\,\|\mathbf{X}\|_\Sigma = t)$ is strictly increasing.*

By Proposition 3.2.15, equation (3.10) is equivalent to

$$F_\Sigma(a, b, C_n^d) = \alpha. \tag{3.11}$$

## 3.3 Moments of $K$

In this section we compute the expected number of projections that we need to take a decision on whether a point is an outlier or not. This is a way to determine the efficiency of the method. It is also helpful to know the variance of the required number of projections to have a more precise idea of the projections that we will need. Proposition 3.3.1 gives such expressions.

**Proposition 3.3.1.** *Under assumption* (A1)*, assume that $a$, $b$ and $t$ are strictly positive numbers such that $a \leq b$ and consider $K$ and $y^{\mathbf{V}}$ defined as in* (3.3) *and* (3.4) *respectively. Then*

$$E(K|\, \|\mathbf{X}\|_\Sigma = t) = \int_{\Omega_\Sigma^{d-1}(t)} \frac{1}{1 - \mathbf{P}(|y^{\mathbf{V}}| \in (a,b))} f_t(\mathbf{x})\, \mathrm{d}\mathbf{x},$$
$$\mathrm{Var}(K|\, \|\mathbf{X}\|_\Sigma = t) = \int_{\Omega_\Sigma^{d-1}(t)} \frac{\mathbf{P}(|y^{\mathbf{V}}| < b) - \mathbf{P}(|y^{\mathbf{V}}| < a)}{(1 - \mathbf{P}(|y^{\mathbf{V}}| \in (a,b)))^2} f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

*Proof.* Beginning with the expectation; by Lemma 3.2.9, we have that

$$\begin{aligned}
\mathrm{E}(K|\, \|\mathbf{X}\|_\Sigma = t) &= \mathrm{E}\left(\mathrm{E}(K|\mathbf{X})|\, \|\mathbf{X}\|_\Sigma = t\right)\\
&= \mathrm{E}\left[\left((1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)|\mathbf{X}))\sum_{i=1}^{\infty} i\mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)|\mathbf{X})^{i-1}\right)\middle|\, \|\mathbf{X}\|_\Sigma = t\right]\\
&= \mathrm{E}\left[\left(\frac{1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b))}{(1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)))^2}\right)\middle|\, \|\mathbf{X}\|_\Sigma = t\right]\\
&= \mathrm{E}\left[\left(\frac{1}{1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b))}\right)\middle|\, \|\mathbf{X}\|_\Sigma = t\right]. \qquad (3.12)
\end{aligned}$$

For the variance, the result follows from (3.12) and

$$\begin{aligned}
\mathrm{Var}(K|\, \|\mathbf{X}\|_\Sigma = t) &= \mathrm{E}\left[K^2 \,|\, \|\mathbf{X}\|_\Sigma = t)\right] - \left(\mathrm{E}\left[K \,|\, \|\mathbf{X}\|_\Sigma = t)\right]\right)^2\\
&= \mathrm{E}\left[\mathrm{E}(K^2 \,|\mathbf{X}, \|\mathbf{X}\|_\Sigma = t)\right] - \left(\mathrm{E}(K \,|\, \|\mathbf{X}\|_\Sigma = t)\right)^2\\
&= \mathrm{E}\left[g_{\mathbf{X}}^t\right] - \left(\mathrm{E}(K \,|\, \|\mathbf{X}\|_\Sigma = t)\right)^2,
\end{aligned}$$

where

$$\begin{aligned}
g_{\mathbf{X}}^t &= \mathrm{E}\left[(1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)|\mathbf{X}))\sum_{i=1}^{\infty} i^2\mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)|\mathbf{X})^{i-1}\,\middle|\, \|\mathbf{X}\|_\Sigma = t\right]\\
&= \mathrm{E}\left[\frac{(1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)))(\mathbf{P}(|Y^{\mathbf{V}}| < b) + \mathbf{P}(|Y^{\mathbf{V}}| > a))}{(1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)))^3}\,\middle|\, \|\mathbf{X}\|_\Sigma = t\right]\\
&= \mathrm{E}\left[\frac{\mathbf{P}(|Y^{\mathbf{V}}| < b) + \mathbf{P}(|Y^{\mathbf{V}}| > a)}{(1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a,b)))^2}\,\middle|\, \|\mathbf{X}\|_\Sigma = t\right].
\end{aligned}$$

□

Observe that the expressions of Proposition 3.3.1 are well-defined, because only the values $a = 0$ and $b \geq t$ would make the integrand infinite.

Notice that Proposition 3.2.7 gives explicit expressions for the probabilities $\mathbf{P}(|y^{\mathbf{V}}| < b)$ which appear in Proposition 3.3.1. The conclusions in Proposition 3.3.1 get simplified in the case $\Sigma = I_d$, as shown in next corollary which follows from Propositions 3.3.1, 3.2.2 and 3.2.10.

**Corollary 3.3.2.** *Under the assumptions in Proposition 3.3.1, if $\Sigma = I_d$, then*

$$E(K|\,\|\mathbf{X}\| = t) = \frac{1}{1 + \mathrm{I}_{a^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) - \mathrm{I}_{b^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)},$$

$$\mathrm{Var}(K|\,\|\mathbf{X}\| = t) = \frac{\mathrm{I}_{b^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) - \mathrm{I}_{a^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)}{\left(1 + \mathrm{I}_{a^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) - \mathrm{I}_{b^2/t^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right)^2}.$$

A graphical representation of the functions in Corollary 3.3.2 appears in the left panel of Figure 3.2 that shows the curves $t \mapsto \mathrm{E}(K|\|\mathbf{X}\| = t) \pm (\mathrm{Var}\,(K|\|\mathbf{X}\| = t))^{1/2}$ and $t \mapsto \mathrm{E}(K|\|\mathbf{X}\| = t)$. Since those curves only depend on $p^t_{a,b} := \mathbf{P}\left(|Y^{\mathbf{V}}| \in (a,b)|\|\mathbf{X}\| = t\right)$, those are the values that we represent in the axis of abscissas. We see that, as expected, the higher the probability that the projection on one single vector belongs to the interval $(a,b)$, $p^t_{a,b}$, the higher the value of the expectation of the number of projections.

Notice that if $p^t_{a,b} = 0$, then $a = b$ and $\mathrm{E}(K|\|\mathbf{X}\| = t) = 1$ and $\mathrm{Var}(K|\|\mathbf{X}\| = t) = 0$. That means that we make our decision based on just one random projection. The other extreme case is when $p^t_{a,b} = 1$, then $a = 0$ and $b = t$ what implies that we should take infinite random projections in order to decide if a point is an outlier or not and the variance is also infinite. Taking into account the continuity of the scalar product and that the procedure we are handling to select the random directions is equivalent to choose them with a uniform distribution on the unit sphere, it happens that, in this case, we are considering all possible directions.

The right panel in Figure 3.2 shows the coefficient of variation of the number of the required random projections, i.e. the values of $(\mathrm{Var}(K|\,\|\mathbf{X}\| = t))^{1/2}/\mathrm{E}(K|\,\|\mathbf{X}\| = t)$. We see that the closer the value of $p^t_{a,b}$ to one, the higher this ratio, which is bounded by one.

## 3.4 Asymptotic properties of the threshold $C^d_n$

In this section we pay attention to the asymptotic behaviour of $C^d_n(\delta)$, defined in (1.1) of Chapter 2, when $n$ or $d$ diverge. This information will be helpful in Section 3.5.

Theorem 3.4.3 demonstrates that $C^d_n(\delta)$ increases to infinity when $n$ or $d$ increases and provides its rate. To prove it, we first need some lemmas. Lemma 3.4.1 gives an upper bound of $F^{-1}_{\chi^2_d}$. It is the first part in Lemma 1 of Laurent and Massart [95]. Lemma 3.4.2 is a simple and curious consequence of L'Hôpital's rule.

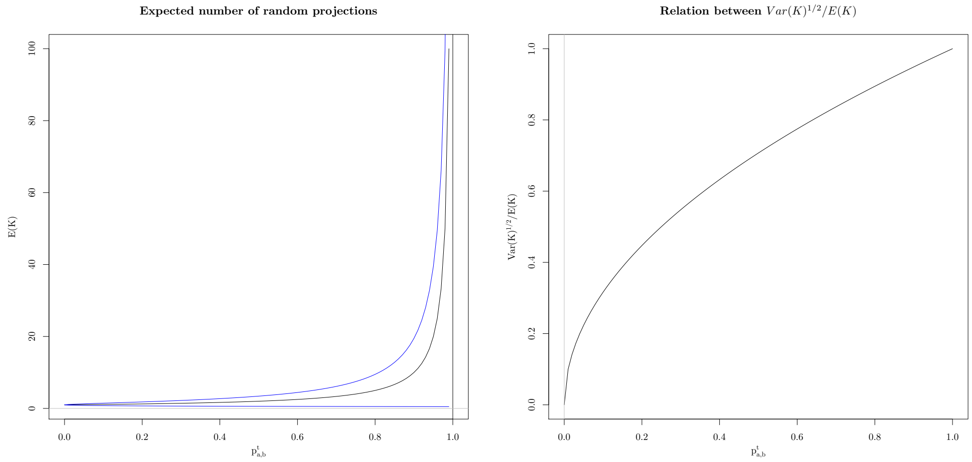


**Figure 3.2.:** The left panel shows the curves $\mathrm{E}(K|\,\|\mathbf{X}\| = t)$ (black) and $\mathrm{E}(K|\,\|\mathbf{X}\| = t) \pm (\mathrm{Var}(K|\,\|\mathbf{X}\| = t))^{1/2}$ (blue). The right panel shows the coefficient of variation of the required number of projections depending on the value of $p^t_{a,b}$ when $\Sigma = I_d$.

**Lemma 3.4.1** (Laurent and Massart)**.** *Let $d \geq 1$ and $s \in (0,1)$, then*

$$F^{-1}_{\chi^2_d}(s) \leq d + \log\left(\frac{1}{1-s}\right) + 2\sqrt{d\log\left(\frac{1}{1-s}\right)}.$$

**Lemma 3.4.2.** *Let $f$ and $g$ be functions such that $\lim_{t\to\infty} f(t) = \lim_{t\to\infty} g(t) = 0$ and $\lim_{t\to\infty} \frac{f'(t)}{g'(t)} = c \in \mathbb{R}$. Then*

$$\lim_{t\to\infty} \frac{\log(f(t))}{\log(g(t))} = 1.$$

*Proof.* By applying L'Hôpital's rule twice, we have:

$$\lim_{t\to\infty} \frac{\log(f(t))}{\log(g(t))} = \lim_{t\to\infty} \frac{g(t)}{f(t)} \lim_{t\to\infty} \frac{f'(t)}{g'(t)} = 1. \qquad \square$$

**Theorem 3.4.3.** *Let $\delta \in (0,1)$ and $C^d_n(\delta)$ defined in (1.1). Then $C^d_n(\delta) \to \infty$ as $n \to \infty$ or $d \to \infty$ while the other parameter remain fixed with rates $\log(n)$ and $d^{1/2}$, respectively.*

*Proof.* Firstly, we obtain the limit when $n \to \infty$ and $d$ is fixed. From (1.1) we have that if $q_{n,d} := F^{-1}_{\chi^2_d}((1-\delta)^{1/n})$, then $q_{n,d} = (C^d_n(\delta))^2$.

Lower bound: By definition we have $(1-\delta)^{1/n} = \mathcal{P}(d/2, q_{n,d}/2)$, where $\mathcal{P}(\cdot,\cdot)$ is the regularized lower gamma function. For $d = 1$,

$$(1-\delta)^{1/n} = \mathcal{P}\left(\frac{1}{2}, \frac{q_{n,1}}{2}\right) = \int_0^{q_{n,1}/2} t^{-1/2} e^{-t}\,\mathrm{d}t = \operatorname{erf}\left(\sqrt{\frac{q_{n,1}}{2}}\right).$$

Take $\beta = 2$ and $\alpha = \sqrt{e/(2\pi)}$ in Theorem 2 in Chang et al. [29] to obtain

$$(1-\delta)^{1/n} < 1 - \sqrt{e/(2\pi)}\exp\{-q_{n,1}\},$$

from where, taking into account that the quantiles of a $\chi^2_d$ increase with $d$, we have that for any $d \geq 1$:

$$q_{n,d} > -\log(1-(1-\delta)^{1/n}) + \log(\sqrt{e/(2\pi)}).$$

Upper bound: By Lemma 3.4.1, for any $n$ and $d$ we have that

$$q_{n,d} \leq d - 2\log\left(1-(1-\delta)^{1/n}\right) + 2\sqrt{-d\log\left(1-(1-\delta)^{1/n}\right)}.$$

As the third term of the above inequality has a lower order than the second one when $n \to \infty$ and $d$ is fixed, we have that both upper and lower bounds have the same order. Take $f(n) := 1-(1-\delta)^{1/n}$ and $g(n) := n^{-1}$. Both functions $f(n)$ and $g(n)$ trivially go to zero when $n \to \infty$. Furthermore

$$\lim_{n\to\infty} \frac{f'(n)}{g'(n)} = -\log(1-\delta).$$

Hence Lemma 3.4.2 gives that $q_{d,n}$ has the same order than $\log(n)$ when $n \to \infty$.

Secondly, we analyse the limit of $C^d_n$ when $d \to \infty$ while $n$ is fixed. Let $Y_d$ be a rv with distribution $\chi^2_d$. Thus, $Y_d$ is the sum of $d$ iid rv's with distribution $\chi^2_1$ whose mean is 1 and whose variance is 2. Then, by the Central Limit Theorem, for $a \in \mathbb{R}$,

$$F_{Y^\star_d}(a) \to \Phi(a), \tag{3.13}$$

where $F_{Y^\star_d}$ denotes the cdf of $Y^\star_d := (Y_d - d)/\sqrt{2d}$. Instead of in a fixed $a$, we are interested in computing this limit on $a_d := ((C^d_n)^2 - d)/\sqrt{2d}$.

Suppose for a contradiction that $\{a_d\}$ is unbounded. Then there exists a subsequence $\{a_{d_k}\}$ such that $\lim_{k\to\infty} a_{d_k} = \infty$. By (3.13) since $F_{Y^\star_d}$ is increasing, for any $a > 0$, we have that

$$1 \geq \overline{\lim}\, F_{Y^\star_{d_k}}(a_{d_k}) \geq \underline{\lim}\, F_{Y^\star_{d_k}}(a_{d_k}) \geq \lim F_{Y^\star_{d_k}}(a) = \Phi(a).$$

On the other hand, since $\lim_{t\to\infty} \Phi(t) = 1$, we would have that $\lim_{k\to\infty} F_{Y^\star_{d_k}}(a_{d_k}) = 1$. This is a contradiction because by definition $F_{Y^\star_{d_k}}\left(\left(C_n^{d_k}\right)^2\right) = \mathbf{P}\left(Y^\star_{d_k} \leq \left(C_n^{d_k}\right)^2\right) = (1 - \delta)^{1/n} \neq 1$ (remember that $n$ is fixed now). Similarly we handle the case $a_{d_k} \to -\infty$. Thus $\{a_d\}$ is bounded.

Suppose now that $\{a_d\}$ does not converge, i.e. suppose that there exist two subsequences $\{d_k^1\}$ and $\{d_k^2\}$ such that $a_{d_k^1} \to a_1$ and $a_{d_k^2} \to a_2$, with $a_1 < a_2$. Let $a_1 < x_1 < x_2 < a_2$. From an index $k$ onward:

$$(1-\delta)^{1/n} = \mathbf{P}\left(\chi^2_{d_k^1} \leq \left(C_n^{d_k^1}\right)^2\right) = \mathbf{P}\left(Y^\star_{d_k^1} \leq a_{d_k^1}\right) \leq \mathbf{P}\left(Y^\star_{d_k^1} \leq x_1\right) \to \Phi(x_1), \tag{3.14}$$

$$(1-\delta)^{1/n} = \mathbf{P}\left(\chi^2_{d_k^2} \leq \left(C_n^{d_k^2}\right)^2\right) = \mathbf{P}\left(Y_{d_k^2} \leq a_{d_k^2}\right) \geq \mathbf{P}\left(Y^\star_{d_k^2} \leq x_2\right) \to \Phi(x_2), \tag{3.15}$$

where the convergence follow from (3.13). (3.14) and (3.15) are simultaneously impossible because for $x_1 < x_2$, $\Phi(x_1) \neq \Phi(x_2)$. Hence $\{a_d\}$ does converge.

Let $a := \lim_{d\to\infty} a_d$, then $0 < (1-\delta)^{1/n} = \lim_{d\to\infty} F_{Y^\star_d}(a_d) = \Phi(a)$ and we have that $a \neq 0$. Then, the result follows from the fact that for any $\epsilon > 0$, from an index onward

$$(a-\epsilon)\sqrt{2d} + d \leq (C_n^d)^2 \leq (a+\epsilon)\sqrt{2d} + d. \tag{3.16}$$

□

**Table 3.1.:** Values of $C_n^d$ for different dimensions, sample sizes and $\delta = 0.05, 0.007$.

| | $C_n^d(0.05)$ | | | | | $C_n^d(0.007)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d = 10$ | $d = 50$ | $d = 200$ | $d = 500$ | $d = 1000$ | $d = 10$ | $d = 50$ | $d = 200$ | $d = 500$ | $d = 1000$ |
| 10 | 5.01 | 8.91 | 15.97 | 24.19 | 33.44 | 5.53 | 9.39 | 16.43 | 24.64 | 33.90 |
| 20 | 5.20 | 9.09 | 16.14 | 24.35 | 33.61 | 5.69 | 9.54 | 16.58 | 24.78 | 34.04 |
| 50 | 5.43 | 9.30 | 16.35 | 24.56 | 33.82 | 5.89 | 9.73 | 16.76 | 24.96 | 34.21 |
| 100 | 5.60 | 9.46 | 16.50 | 24.71 | 33.96 | 6.04 | 9.87 | 16.89 | 25.09 | 34.34 |
| 200 | 5.76 | 9.79 | 16.64 | 24.85 | 34.10 | 6.18 | 10.00 | 17.02 | 25.21 | 34.46 |
| 1000 | 6.10 | 9.93 | 16.95 | 25.15 | 34.40 | 6.49 | 10.29 | 17.29 | 25.48 | 34.73 |

From (3.16), we obtain the following corollary.

**Corollary 3.4.4.** *Let $C_n^d(\delta)$ defined in (1.1) and $n \in \mathbb{N}$, then*

$$\lim_{d\to\infty} \frac{C_n^d(\delta)}{\sqrt{d}} = 1.$$

Some illustrations of Theorem 3.4.3 appear in Table 3.1 and Figure 3.3 which show, for different values of $d$ and $n$, the values of $C_n^d(\delta)$ for $\delta = 0.05$ and $\delta = 0.007$. In both cases, it is evident that, even for small sizes, the value of $C_n^d$ grows faster on the dimension than on the sample size. However, the difference between tha values for $\delta = 0.05$ and $\delta = 0.007$ is not too large and decreases with $n$ and $d$.

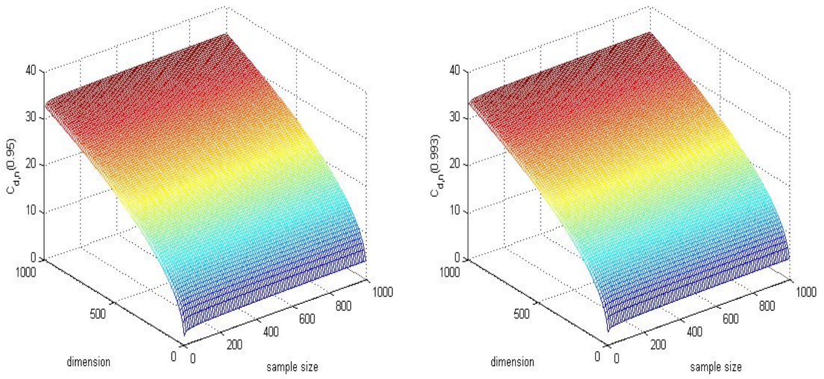


**Figure 3.3.:** Values of $C_n^d(0.05)$ (left panel) and of $C_n^d(0.007)$ (right panel).

## 3.5 Computation of the constants $a$ and $b$

In this section we seek the $a$ and $b$ solutions of (3.11). An obvious answer is $a = b = a_\alpha$, the quantile $(1 - \alpha)$ of $Y^{\mathbf{V}}$, however this obviously produces the worst error of type II. Therefore, we will search for another solutions taking into account the power of the test under the alternative hypothesis. We obtain explicit formulae of the solutions of (3.11) for the identity and approximated numerical solutions for $b$ given a general covariance matrix and $a$ computed with the identity (see Algorithm 3). Note that in the case of $\Sigma \neq I_d$ since $\Sigma$ is known, making use of the representation $\mathbf{X} = \Sigma^{1/2}\mathbf{X}_0$, we could transform the problem into the case $\Sigma = I_d$. Nevertheless we do not do it because these results will be helpful in Chapter 4 where the parameters are unknown. The main difference between the scenarios where $\Sigma = I_d$ and $\Sigma \neq I_d$ is the dependency of the projections given $\|\mathbf{X}\|_\Sigma$ as Proposition 3.2.10 showed. Because of this, we split the computation of $a$ and $b$ in the cases $\Sigma = I_d$ (Subsection 3.5.1) and $\Sigma \neq I_d$ (Subsection 3.5.2).

According to Proposition 3.5.1 below, for every $a \in (0, a_\alpha)$, there exists a unique $b_a$ such that the pair $(a, b_a)$ gives a test at the level $\alpha$ for the covariance matrix under consideration. Moreover, the lower the $a$, the larger the number of required projections, what increases the chances of taking the right decision (at the price of a higher computational time). Later, we will take advantage of this proposition to compute $a$ and $b$ in some specific simulations. This will be crucial when the parameters are unknown.

**Proposition 3.5.1.** *Let $\mathbf{x} \in \mathbb{R}^d$ be such that $0 < t = \|\mathbf{x}\|_\Sigma$. Given $a > 0$ such that $\mathbf{P}(|y^{\mathbf{V}}| < a) \leq \alpha$, there exists $b_a^t$ such that $F_\Sigma(a, b_a^t, t) = \alpha$. Moreover, for every $t > C_n^d$ the map $a \mapsto b_a^t$ is strictly decreasing on $a$.*

*Proof.* We know it must be $a \leq b$. If $a = b$, then

$$\alpha = \int_{\Omega_\Sigma^{d-1}(t)} \mathbf{P}\left(|y^{\mathbf{V}}| > b\right) f_{\hat{\mu}}(\mathbf{m}) f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{3.17}$$

This condition determines $b$ because the function $b \mapsto \mathbf{P}\left(|y^{\mathbf{V}}| > b\right)$ is strictly decreasing and continuous for any $\mathbf{x}$. Let $b_0^t$ be the unique solution of (3.17). If $a < b_0^t$, there exists a unique $b_a^t$ such that

$$\alpha = F_\Sigma(a, b_a^t, t),$$

because the integrand which implicitly appears in $F_\Sigma(a, b, t)$ (see Proposition 3.2.11) is strictly increasing on $b$ and continuous.

If $a_1 < a_2$, denote $h_a^b(\mathbf{x}) := \mathbf{P}(|y^{\mathbf{V}}| > b)/(1-\mathbf{P}(|y^{\mathbf{V}}| \in (a, b)))$, then $h_{a_2}^{b_{a_1}^t}(\mathbf{x}) < h_{a_1}^{b_{a_1}^t}(\mathbf{x})$ since $\mathbf{P}\left(|y^{\mathbf{V}}| < a_1\right) < \mathbf{P}\left(|y^{\mathbf{V}}| < a_2\right)$. Hence $F_\Sigma(a_2, b_{a_1}^t, t) > F_\Sigma(a_1, b_{a_1}^t, t)$, then $b_{a_2}^t < b_{a_1}^t$. □

Proposition 3.5.2 describes the relation between the expected number of projections required to reach a decision when $\Sigma \neq I_d$ and $\Sigma = I_d$, what leads us to use the solution of the identity.

**Proposition 3.5.2.** *Suppose that $0 < a \leq b < t$, and consider $K$ as in (3.3). Under assumptions (A1) and (A2), if $\Sigma \neq I_d$, then*

$$E(K | \|\mathbf{X}\|_\Sigma = t) > E(K |\, \|\mathbf{X}\| = t).$$

*Proof.* By Propositions 3.2.7 and 3.3.1

$$\begin{aligned} \mathrm{E}(K |\, \|\mathbf{X}\| = t) &= \frac{1}{1 - \mathbf{P}(|Y^{\mathbf{V}}| \in (a, b) | \|\mathbf{X}\| = t)} \\ &= \frac{1}{\displaystyle\int_{\Omega_\Sigma^{d-1}(t)} \left(1 - \mathbf{P}(|y^{\mathbf{V}}| \in (a, b))\right) f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}} \\ &< \int_{\Omega_\Sigma^{d-1}(t)} \frac{1}{1 - \mathbf{P}(|y^{\mathbf{V}}| \in (a, b))} f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}, \end{aligned}$$

where the inequality comes from the fact that the map $\mathbf{x} \mapsto \mathbf{P}(|y^{\mathbf{V}}| > b) + \mathbf{P}(|y^{\mathbf{V}}| < a)$ is continuous and not constant on $\mathbf{x}$ a.s. and Jensen's inequality. □

**Remark 3.5.3.** *Let us denote $Z$ the rv whose value is 1 if the point under consideration is declared outlier and zero otherwise. Let us assume that, in order to accelerate the computations, we fix $k$ equal to the maximum number of projections to be taken. If this number is reached, we decide with probability $\gamma \in (0,1)$ that the point under consideration is an outlier. Let $Z_\gamma^k$ be defined as $Z$ with this accelerated procedure. It is clear that*

$$Z_0^k \leq Z \text{ and } \mathbf{P}(Z_0^k < Z) > 0 \ \left(\text{resp. } Z_1^k \geq Z \text{ and } \mathbf{P}(Z_1^k > Z) > 0\right).$$

*Therefore, the probability to declare the point as an outlier when $\gamma = 0$ (resp. $\gamma = 1$) is strictly less (resp. greater) than $\alpha$.*

*Intermediate values of $\gamma$ are considered in Section 3.7. However we leave an in-depth study of the variation on $\gamma$ of those probabilities because the selection of its proper value giving a probability equal to $\alpha$ does not seem trivial. The strict monotonicity of the map $\gamma \to \mathbf{P}(Z_\gamma^k < Z)$ and its continuity make this value unique.*

### 3.5.1 Computation of $(a, b)$ when $\Sigma = I_d$

Let us assume that we want to compute the constants $a$ and $b$ giving a power $\alpha$ test with a given finite value $k \geq 1$ for $\mathrm{E}(K)$, when $\Sigma = I_d$. Taking into account the expressions in Corollaries 3.2.12 and 3.3.2, we could solve the system

$$\begin{aligned} k &= (1 - v + u)^{-1} \\ \alpha &= (1 - v)(1 - v + u)^{-1}, \end{aligned} \tag{3.18}$$

and then look for $a$, $b$ such that $u := \mathrm{I}_{a^2/(C_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ and $v := \mathrm{I}_{b^2/(C_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)$. Thus, Proposition 3.5.4, whose proof is trivial, gives explicit expressions of the values $a$ and $b$ satisfying (3.18).

**Proposition 3.5.4.** *Let $\Sigma = I_d$ and $k \in \mathbb{N}$. The pair $(a, b)$ determining the solution of (3.11) with $E(K) = k$ satisfies that $a = C_n^d B_d(u)$, $b = C_n^d B_d(v)$, where $v = 1 - \alpha/k$, $u = (1 - \alpha)/k$ with $v - u \neq 1$ and*

$$B_d(w) := \sqrt{\mathrm{B}^{-1}\left(w\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right); \frac{1}{2}, \frac{d-1}{2}\right)}, \text{ for } w = u, v.$$

For the sake of brevity, we omit the dependency on $n$, $d$, $u$ and $v$ of the values $(a, b)$ obtained in Proposition 3.5.4 unless needed. Note that in such proposition, we have implicitly excluded the values $a = 0$ and $b \geq C_n^d$, which corresponds to $v - u = 1$; because in this case we would have $k = \infty$ in (3.18).

Proposition 3.5.4 allows to study the behaviour of the values $a$ and $b$ when the dimension or the sample size increase. From Proposition 3.5.4, it is clear that, if we fiz $d$, then $a$ and $b$ become proportional to $C_n^d$. However, Proposition 3.5.7 shows that $a$ and $b$ converge to a strictly positive finite value when $d$ goes to infinity and $n$ is fixed and, it provides the rate of convergence. This result was not evident because, according to Theorem 3.4.3, $\lim_{d\to\infty} C_n^d = \infty$ and $a$ and $b$ are clearly related to $C_n^d$. To ease the proof of such proposition, we state some previous lemmas.

**Lemma 3.5.5.** *Consider* $h_d = d^{1/2}\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right)$, *then there exists* $\lim_{d\to\infty} h_d = h \in (0, \infty)$.

*Proof.* It follows from Stirling's approximation, since $\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right) = \Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{1}{2}\right)/\Gamma\left(\frac{d}{2}\right)$. □

**Lemma 3.5.6.** *Given* $u \in (0,1)$, *then there exists* $S_u := \lim\limits_{d\to\infty} \sqrt{d}B_d(u)$ *and* $S_u \in (0, \infty)$.

*Proof.* Let $x_d(u) := (B_d(u))^2$ and $q_d(u) := u\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right)$. Two claims are proved: (i) The function $x_d(u)$ converges to zero with rate of convergence $d^{-1}$ and (ii) The limit of $dx_d(u)$ exists. For (i), by definition,

$$\begin{aligned} q_d(u) &= \int_0^{x_d(u)} t^{-1/2}(1-t)^{(d-3)/2}\,\mathrm{d}t \\ &\geq (1 - x_d(u))^{(d-3)/2} \int_0^{x_d(u)} t^{-1/2}\,\mathrm{d}t \\ &= 2(x_d(u))^{1/2}(1 - x_d(u))^{(d-3)/2}. \end{aligned}$$

Thus, from Lemma 3.5.5 we have that $x_d(u) \to 0$ with rate at most $n^{-1}$. However

$$q_d(u) \leq \int_0^{x_d(u)} t^{-1/2}\,\mathrm{d}t = 2(x_d(u))^{1/2},$$

and consequently, Lemma 3.5.5 again allows us to conclude (i). For (ii), by (i), there exists a function $g_d(u)$ such that $x_d(u) = g_d(u)d^{-1}$ and for any $u \in (0,1)$

$$0 < \underline{\lim}_{d\to\infty}\, g_d(u) \leq \overline{\lim}_{d\to\infty}\, g_d(u) < \infty.$$

Fix $u$ and omit it in the notation. Suppose that there exist two sequences $\{d_k^1\}$ and $\{d_k^2\}$ such that both diverge and

$$0 < \lim_{k\to\infty} g_{d_k^2} = g_2 < g_1 = \lim_{k\to\infty} g_{d_k^1} < \infty.$$

Take a subsequence $\{d^2_{k^\star}\}$ of $\{d^2_k\}$ such that $d^2_{k^\star} > kd^1_k$ and that for any $k$,

$$g_{d^1_k} > g_{d^2_k}. \tag{3.19}$$

Let $h_d$ be defined in Lemma 3.5.5, thus $h_d \to h \in (0, \infty)$. Then

$$\lim_k \left( q_{d^1_k} \sqrt{d^1_k} - q_{d^2_{k^\star}} \sqrt{d^2_{k^\star}} \right) = 0. \tag{3.20}$$

To ease the notation, we call $\Lambda_k := q_{d^1_k}\sqrt{d^1_k} - q_{d^2_{k^\star}}\sqrt{d^2_{k^\star}}$, $\tilde{x}^1_k := x_{d^1_k}$, $\tilde{x}^2_k := x_{d^2_{k^\star}}$ and $\varphi(t,d) := d^{1/2}(1-t)^{\frac{d-3}{2}}$. Since $d^2_{k^\star} > d^1_k$, by (3.19), we have that $\tilde{x}^1_k > \tilde{x}^2_k$ and

$$\begin{aligned} \Lambda_k &= \int_0^{\tilde{x}_1} t^{-\frac{1}{2}} \varphi(t, d^1_k)\,\mathrm{d}t - \int_0^{\tilde{x}_2} t^{-\frac{1}{2}} \varphi(t, d^2_{k^\star})\,\mathrm{d}t \\ &= \int_{\tilde{x}_2}^{\tilde{x}_1} t^{-\frac{1}{2}} \varphi(t, d^1_k)\,\mathrm{d}t + \int_0^{\tilde{x}_2} t^{-\frac{1}{2}} \left( \varphi(t, d^1_k) - \varphi(t, d^2_{k^\star}) \right) \mathrm{d}t. \end{aligned}$$

Note that the second term of the above equality is strictly positive from an index onward because $d^2_{k^\star} > d^1_k$ and the function $\varphi(t,d)$ is decreasing in $d$ from an index forward, since the derivative of its square is

$$\frac{\partial \varphi^2(t,d)}{\partial d} = (1-t)^{d-3} + d\log(1-t)(1-t)^{d-3}.$$

Therefore,

$$\Lambda_k \geq \int_{\tilde{x}^2_k}^{\tilde{x}^1_k} t^{-\frac{1}{2}} \varphi(t, d^1_k)\,\mathrm{d}t \geq \varphi(\tilde{x}^1_k, d^1_k) \int_{\tilde{x}^2_k}^{\tilde{x}^1_k} t^{-\frac{1}{2}}\,\mathrm{d}t = (1-\tilde{x}^1_k)^{\frac{d^1_k - 3}{2}} 2 \left( \sqrt{g_{d^1_k}} - \sqrt{g_{d^2_{k^\star}} \frac{d^1_k}{d^2_{k^\star}}} \right).$$

Take $k \to \infty$, as $d^2_{k^\star} > d^1_k$, we have $\Lambda_k \to e^{-g_1/2} 2\sqrt{g_1} > 0$, which is a contradiction with (3.20) and the Claim (ii) is proved. Thus there exists $S_u = \lim\limits_{d\to\infty} dx_d(u)$ and $S_u \neq 0$ by Claim (i). □

**Proposition 3.5.7.** *Let $\alpha \in (0,1)$ and $k \geq 1$. The pair $(a_d, b_d)$ determining the solution of (3.18) satisfies that there exist $a_\infty = \lim\limits_{d\longrightarrow\infty} a_d$ and $b_\infty = \lim\limits_{d\longrightarrow\infty} b_d$, when $n$ is kept fixed. Furthermore, both limits are finite and strictly positive.*

*Proof.* From Corollary 3.4.4, we have that $\lim\limits_{d\to\infty} \frac{C^d_n}{\sqrt{d}} = 1$ and the result follows from Lemma 3.5.6. □

Table 3.2 shows the values of $a$ and $b$ (approximated taking four decimals) with $\delta = 0.05$ and $\alpha = 0.05$. Those values have been computed using the expression of Proposition

3.5.4 for several values of $n$, $d$ and $k$. To do this, we have used the function `Ibeta.inv` of the `zipfR` package in R which allows to compute the inverse of an incomplete beta function. The choice of the values of the dimension and the sample size (here and through this chapter and the next one) attempts to represent the scenarios when the dimension is higher (smaller) than the sample size. From this table, the bigger $\mathrm{E}(K)$ and/or $n$, the wider the interval $(a, b)$ according to Corollary 3.3.2. However, the larger the dimension, the narrower the interval $(a, b)$. In addition, $a$ and $b$ converge to a different limit as the dimension goes to infinity, which is consistent with Proposition 3.5.7.

Table 3.2.: Values of $(a, b)$ for $\Sigma = I_d$, $C_n^d \equiv C_n^d(0.05)$ and different values of $n, d$ and $k = \mathrm{E}\left(K \,|\, \|\mathbf{X}\| = C_n^d\right)$.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k = 50$ | | $k = 100$ | | $k = 50$ | | $k = 100$ | | $k = 50$ | | $k = 100$ | |
| $d$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ |
| 5 | 0.0573 | 4.4057 | 0.0287 | 4.4401 | 0.0595 | 4.5737 | 0.0297 | 4.6095 | 0.0642 | 4.9369 | 0.0321 | 4.9755 |
| 50 | 0.0318 | 4.1611 | 0.0159 | 4.3736 | 0.0323 | 4.2302 | 0.0162 | 4.4463 | 0.0335 | 4.3793 | 0.0167 | 4.6029 |
| $10^2$ | 0.0293 | 3.9424 | 0.0147 | 4.1569 | 0.0297 | 3.9912 | 0.0148 | 4.2084 | 0.0305 | 4.0962 | 0.0152 | 4.3192 |
| 500 | 0.0262 | 3.6000 | 0.0131 | 3.8057 | 0.0264 | 3.6214 | 0.0132 | 3.8283 | 0.0267 | 3.6675 | 0.0133 | 3.8770 |
| $10^3$ | 0.0255 | 3.5119 | 0.0127 | 3.7137 | 0.0256 | 3.5270 | 0.0128 | 3.7297 | 0.0258 | 3.5593 | 0.0129 | 3.7638 |

### 3.5.2 Computation of $(a, b)$ when $\Sigma \neq I_d$

Taking into account the expressions in Propositions 3.2.11 and 3.3.1, in order to compute the constants $a$ and $b$ giving an $\alpha$-level test with a given value $k \geq 1$ for $\mathrm{E}(K)$, when $\Sigma \neq I_d$, we should solve the system

$$
\begin{aligned}
k &= \int_{\Omega_\Sigma^{d-1}(C_n^d)} \frac{1}{1 - \mathbf{P}(|y^{\mathbf{V}}| < b) + \mathbf{P}(|y^{\mathbf{V}}| < a)} f_{C_n^d}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
\alpha &= \int_{\Omega_\Sigma^{d-1}(C_n^d)} \frac{1 - \mathbf{P}(|y^{\mathbf{V}}| < b)}{1 - \mathbf{P}(|y^{\mathbf{V}}| < b) + \mathbf{P}(|y^{\mathbf{V}}| < a)} f_{C_n^d}(\mathbf{x}) \, \mathrm{d}\mathbf{x}.
\end{aligned}
$$

However, due to the difficulty in solving this system (see Proposition 3.2.7) when $k > 1$ (remember that in the not too interesting case $k = 1$, those expressions do not depend on $\Sigma$, see Proposition 3.2.1, and we take the solution corresponding to $\Sigma = I_d$), we have decided, basing our selves on Proposition 3.5.3, to use the values obtained for $\Sigma = I_d$ to manage any covariance matrix. Before taking this decision we need to check if those values are adequate for general matrices. As stated, we are not aware of any specific determination of $a$ and $b$ for general matrices and to make this check, we have resorted to carry out the comparison in four specific families of covariance matrices. In addition, since we think that the worst situation with those constants could occur in matrices with sparse eigenvalues, we have chosen three families with large variation among them and one with little variation. The considered families are:

- $\Sigma_1^d$ are matrices with the half of their eigenvalues 1's and the others $d^2$.
- $\Sigma_2^d$ are matrices with equally spaced eigenvalues from 1 to $d^2$.
- $\Sigma_3^d$ are matrices whose eigenvalues are 1's $d-1$ times and one is $d^2$.
- $\Sigma_4^d$ are matrices with eigenvalues varying between 1 and 2. They are the result of the ratio between two equispaced sequences between $d^2$ and 2 and between $d^2$ and 1 respectively.

In those computations and in those that follows $\mathrm{E}(K | \|\mathbf{X}\|_\Sigma = rC_n^d(\delta))$, with $\delta = 0.05$, will be denoted as $k_I^r$ and $k_i^r$ when $\Sigma = I_d$ or $\Sigma = \Sigma_i^d$, $i = 1, \ldots, 4$ respectively. The super-index $r$ will be omitted when $r = 1$.

The comparison goes as follows: For each combination of dimension and sample size, we have computed a pair $(a_I, b_I)$ giving an $\alpha$-level test for the identity matrix with the expression in Proposition 3.5.4. We have kept $a_I$ and, for every $\Sigma_i^d$, $i = 1, \ldots, 4$, we have computed the value $b_\Sigma$ such that the pair $(a_I, b_{\Sigma_i})$ is an $\alpha$-level test using Algorithm 3 with $N = 10^4$ simulations.

```
Algorithm 3: Computation of $b$ when $\Sigma$ is known
1) We fix $N$ large and for $j = 1, \ldots, N$:
    1.1) Generate $\mathbf{X}_0^j, \mathbf{X}_1^j, \ldots, \mathbf{X}_n^j$ iid rv's with distribution $N_d(\mathbf{0}, \Sigma)$ and $\mathbf{V}^j$
         independent from the rest with distribution $N_d(\mathbf{0}, I_d)$
    1.2) Consider $\mathbf{X}^j = \frac{C_n^d(\delta)\mathbf{X}_0^j}{\|\mathbf{X}_0^j\|}$
    1.3) Compute $Y^j = \frac{|(\mathbf{X}^j)'\mathbf{V}^j|}{\sigma_{\mathbf{V}^j}}$
2) Take $b$ equal to the quantile $v$ of the sample $Y^1, \ldots, Y^N$
```

Regrettably, some simulations have shown that the test associated to the obtained pair has generally power lower than $\alpha$, because the values $b_\Sigma$ are lower than desired. To fix this point we recalculate $b_{\Sigma_i}$, keeping $a_I$ fixed, by simulations with the bisection method. This procedure has proved to give tests at the right level.

The great news is that in all cases we have considered, we have found that the upper bound of the interval which covers the $95\%$ of the values of the rv $Y^{\mathbf{V}}$ with an endpoint in $a_I$ is very similar to $b_I$ when $d \geq 50$. Moreover, excepting if $\Sigma = \Sigma_3^d$, the sample means of the number of projections $K$ are also very similar (see Table A.1 in the Appendix). All obtained $b_\Sigma$'s are in Table 3.3. For each pair of sample size and dimension, Table 3.4 shows the $b_\Sigma$ maximizing the difference $|b_I - b_\Sigma|$ along the four covariance matrices and the matrix producing it. Those values of $b_\Sigma$ should be compared with the $b_I$'s shown in Table 3.2.

**Table 3.3.:** Approximated values of $b$ for $\Sigma = \Sigma_i^d$ with $i = 1, \dots, 4$ and different values of $n$, $d$, and $a$'s are the values obtained in Table 3.2 for $k_I^1 = 50, 100$.

| | | $\Sigma_1^d$ | | $\Sigma_2^d$ | | $\Sigma_3^d$ | | $\Sigma_4^d$ | |
|---|---|---|---|---|---|---|---|---|---|
| $d$ | $n$ | $k_I^1 = 50$ | $k_I^1 = 100$ | $k_I^1 = 50$ | $k_I^1 = 100$ | $k_I^1 = 50$ | $k_I^1 = 100$ | $k_I^1 = 50$ | $k_I^1 = 100$ |
| 5 | 50 | 4.4118 | 4.4418 | 4.4139 | 4.4440 | 4.4358 | 4.4593 | 4.4074 | 4.4399 |
| | 100 | 4.5766 | 4.6135 | 4.5794 | 4.6132 | 4.6021 | 4.6317 | 4.5778 | 4.6066 |
| | 500 | 4.9412 | 4.9749 | 4.9412 | 4.9800 | 4.9651 | 4.9971 | 4.9412 | 4.9754 |
| 50 | 50 | 4.1832 | 4.3740 | 4.1606 | 4.3694 | 4.1606 | 4.3695 | 4.1731 | 4.3693 |
| | 100 | 4.2482 | 4.4237 | 4.2390 | 4.4415 | 4.2482 | 4.4421 | 4.2298 | 4.4363 |
| | 500 | 4.3978 | 4.6081 | 4.3788 | 4.6177 | 4.3692 | 4.5795 | 4.3788 | 4.5938 |
| 100 | 50 | 3.9429 | 4.1683 | 3.9577 | 4.1743 | 3.9339 | 4.1385 | 3.9339 | 4.1623 |
| | 100 | 4.0098 | 4.2267 | 3.9947 | 4.1988 | 3.9827 | 4.1705 | 4.0068 | 4.2079 |
| | 500 | 4.0875 | 4.3371 | 4.0998 | 4.3310 | 4.0875 | 4.2753 | 4.0906 | 4.3310 |
| 500 | 50 | 3.5960 | 3.8125 | 3.6028 | 3.8088 | 3.5893 | 3.7526 | 3.5891 | 3.8006 |
| | 100 | 3.6190 | 3.8232 | 3.6340 | 3.8202 | 3.5933 | 3.8184 | 3.6314 | 3.8190 |
| | 500 | 3.6726 | 3.8902 | 3.6573 | 3.8718 | 3.6489 | 3.8629 | 3.6513 | 3.8707 |
| 1000 | 50 | 3.5069 | 3.7183 | 3.5069 | 3.7180 | 3.4978 | 3.7075 | 3.5057 | 3.7165 |
| | 100 | 3.5385 | 3.7297 | 3.5328 | 3.7295 | 3.5317 | 3.7218 | 3.5328 | 3.7295 |
| | 500 | 3.5593 | 3.7578 | 3.5472 | 3.7349 | 3.5439 | 3.7317 | 3.5469 | 3.7350 |

**Table 3.4.:** Values of $b_\Sigma$ giving the greatest difference $|b_I - b_\Sigma|$ for $\Sigma = \Sigma_i^d, i = 1, \dots, 4$, and different values of $d$ and $n$. $a$'s are taken from Table 3.2. Column $\Sigma$ tells the matrix in which $b_\Sigma$ was obtained.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k_I^1$=50 | | $k_I^1$=100 | | $k_I^1$=50 | | $k_I^1$=100 | | $k_I^1$=50 | | $k_I^1$=100 | |
| $d$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ |
| 5 | 4.4074 | $\Sigma_4^d$ | 4.4399 | $\Sigma_4^d$ | 4.6021 | $\Sigma_3^d$ | 4.6317 | $\Sigma_3^d$ | 4.9651 | $\Sigma_3^d$ | 4.9971 | $\Sigma_3^d$ |
| 50 | 4.1832 | $\Sigma_1^d$ | 4.3693 | $\Sigma_4^d$ | 4.2482 | $\Sigma_3^d$ | 4.4237 | $\Sigma_1^d$ | 4.3978 | $\Sigma_1^d$ | 4.5795 | $\Sigma_3^d$ |
| 100 | 3.9577 | $\Sigma_2^d$ | 4.1743 | $\Sigma_2^d$ | 4.0098 | $\Sigma_1^d$ | 4.1705 | $\Sigma_3^d$ | 4.0875 | $\Sigma_3^d$ | 4.2753 | $\Sigma_3^d$ |
| 500 | 3.5891 | $\Sigma_4^d$ | 3.7526 | $\Sigma_3^d$ | 3.5933 | $\Sigma_3^d$ | 3.8184 | $\Sigma_3^d$ | 3.6489 | $\Sigma_3^d$ | 3.8629 | $\Sigma_3^d$ |
| 1000 | 3.4978 | $\Sigma_3^d$ | 3.7075 | $\Sigma_3^d$ | 3.5317 | $\Sigma_3^d$ | 3.7218 | $\Sigma_3^d$ | 3.5439 | $\Sigma_3^d$ | 3.7317 | $\Sigma_3^d$ |

## 3.6 On the error of type II

Apart from the number of expected number of projections, the error of type II is another issue that we have to take into account in order to decide which values of $a$ and $b$ are preferable. For that, we take points in the alternative, i.e. points with Mahalanobis norm greater than $C_n^d$, and compute the probability of declaring them as outliers.

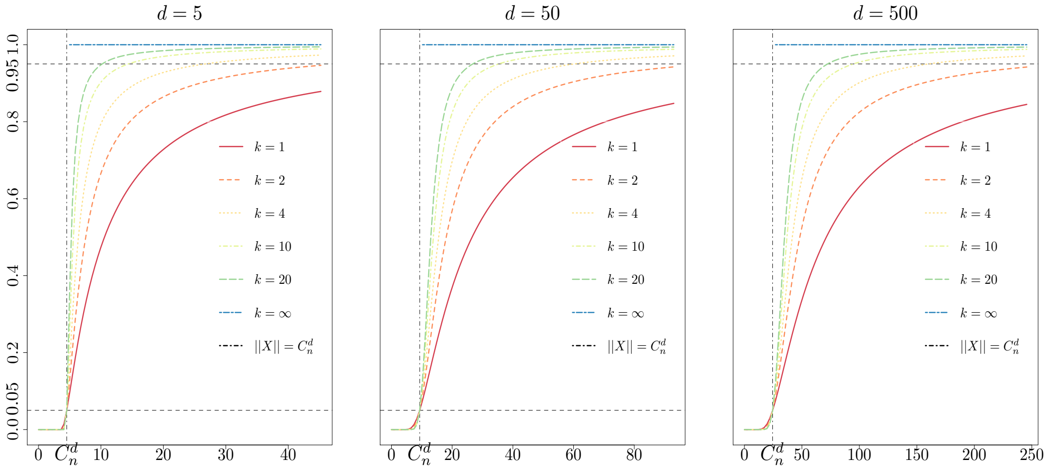


**Figure 3.4.:** Probability of declaring a point as outlier (vertical axis) given its Mahalanobis norm (horizontal axis) for different values of $d$, $n = 50$ and $\Sigma = I_d$. Here $k := \mathrm{E}(K \,|\, \|\mathbf{X}\| = C_n^d)$.

Figure 3.4 shows a graphical representation of the function of Corollary 3.2.12 for $a$ and $b$ which satisfy equation (3.11). We see how the probability of declaring a point as an outlier increases when the norm does so for $n = 50$, $d = 5, 50, 500$, $\Sigma = I_d$ and for different values of $k_I := \mathrm{E}(K \,|\, \|\mathbf{X}\| = C_n^d)$, i.e. for different values of $a$ and $b$. We also see in dashed horizontal lines the probabilities 0.05 and 0.95 and in the vertical line the norm $C_n^d$. This vertical line divides the graph in two parts, the left part corresponds to the points whose Mahalanobis norm is less than $C_n^d$, that is, points which are no outliers, the right part corresponds to points with Mahalanobis norm greater than $C_n^d$, that means points which are outliers. When we are in the left part, the probability of declaring the point $\mathbf{X}$ as an outlier is always less than $\alpha = 0.05$ for every value of $k$, according to the fact that we are handling tests at the level 0.05. Meanwhile, when $\mathbf{X}$ has a Mahalanobis norm greater than $C_n^d$, the probability of declaring the point as an outlier depends clearly upon the value of $k$.

The reason of choosing only a particular value of $n$ in Figure 3.4 is that the results show that this probability is hardly dependent on the sample size. Additionally, we see in the figure that all the curves basically have the same shape independently on $d$, and even the probability to declare as outliers points with norm $hC_n^d$ seems similar for the different values of $d$ and $h \geq 5$. On the other hand, when $k = 1$ ($a = b$), we are in the worst scenario because the area under the curve is the lowest and when $k = \infty$ ($a = 0$ and $b = C_n^d$), it is the best one because the area under the curve is the highest possible. In other words, the larger the number of expected number of projections, the lower the error of type II. Consequently, a desirable feature is to find a balance between the expected number of vectors and the error of type II. Proposition 3.6.1 helps us to it because it allows to compute $a$ and $b$ in order to have a desired power.

**Proposition 3.6.1.** *Let* $\mathbf{X}$ *be a rv with distribution* $N_d(\mathbf{0}, I_d)$. *Under assumption (A2), assume* $h > 1$ *and* $0 < p_0 < p_h < 1$. *Then, the system*

$$\begin{cases} \mathbf{P}(\mathbf{X} \textit{ declared as outlier} \mid \|\mathbf{X}\| = C_n^d) & = p_0 \\ \mathbf{P}(\mathbf{X} \textit{ declared as outlier} \mid \|\mathbf{X}\| = hC_n^d) & = p_h, \end{cases} \tag{3.21}$$

*has solutions* $a$ *and* $b$ *such that* $a = C_n^d B_d(u)$, $b = C_n^d B_d(v)$ *where* $v = 1 - up_0/(1-p_0)$ *and* $u$ *satisfies the following system, where* $u^* \in (0, 1-p_h)$:

$$\begin{cases} \mathrm{B}^{-1}\left(u\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right); \frac{1}{2}, \frac{d-1}{2}\right) & = h^2\mathrm{B}^{-1}\left(u^\star\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right); \frac{1}{2}, \frac{d-1}{2}\right) \\ \mathrm{B}^{-1}\left(v\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right); \frac{1}{2}, \frac{d-1}{2}\right) & = h^2\mathrm{B}^{-1}\left(\left(1 - u^\star\frac{p_h}{1-p_h}\right)\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right); \frac{1}{2}, \frac{d-1}{2}\right). \end{cases} \tag{3.22}$$

*Proof.* By Corollary 3.2.12, we rewrite the system (3.21) as

$$\begin{cases} p_0 = \left(1 - I_{b^2/(C_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right)\Big/\left(1 + I_{a^2/(C_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) - I_{b^2/(C_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right) \\ p_h = \left(1 - I_{b^2/(hC_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right)\Big/\left(1 + I_{a^2/(hC_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right) - I_{b^2/(hC_n^d)^2}\left(\frac{1}{2}, \frac{d-1}{2}\right)\right). \end{cases} \tag{3.23}$$

Taking $\alpha = p_0$ and $k = 1/(1-p_0)$ in Proposition 3.5.4, the first equation of system (3.23) becomes

$$a = C_n^d B_d(u), b = C_n^d B_d(v), \text{ with } v = 1 - up_0/(1-p_0) \text{ and } u \in (0, 1-p_0),$$

and taking $\alpha = p_h$ and $k = 1/(1-p_h)$ in Proposition 3.5.4, the second one is

$$a = hC_n^d B_d(u^\star), b = hC_n^d B_d(v^\star), \text{ with } v^\star = 1 - u^\star p_h/(1-p_h) \text{ and } u^\star \in (0, 1-p_h),$$

where $B_d(t)$ is defined in Proposition 3.5.4. The result is deduced since the values of $a$ and $b$ have to be the same in both equations. □

Table 3.5 shows the expected number of projections, calculated with Corollary 3.3.2, and using the values of $a$ and $b$ of Proposition 3.6.1 for $p_0 = 0.1$ and $p_h = 0.95$, $n = 100$ and different values of $h$ and $d$. Note that we have computed the values of $a$ and $b$ of Proposition 3.6.1 in a numerical way since we are not able to obtain an explicit formula from system (3.22). In the table we see that $k$ considerably increases when $d$ does so and $h \leq 4$, while it has resembling values when $h = 5$.

Table 3.5.: Values of $k$ and $k^h$ for $n = 100$ and different values of $d$ using $a$ and $b$ of Proposition 3.6.1 with $p_0 = 0.1$ and $p_h = 0.95$.

| | | | | $d = 5$ | | | |
|---|---|---|---|---|---|---|---|
| | $h = 1.1$ | $h = 1.2$ | $h = 1.3$ | $h = 1.4$ | $h = 1.5$ | $h = 3$ | $h = 5$ |
| $k_I$ | 51 | 40 | 33 | 28 | 24 | 8 | 4 |
| $k_I^h$ | 3 | 3 | 2 | 2 | 2 | 1 | 1 |
| | | | | $d = 50$ | | | |
| | $h = 1.1$ | $h = 1.2$ | $h = 1.3$ | $h = 1.4$ | $h = 1.5$ | $h = 3$ | $h = 5$ |
| $k_I$ | $> 10^{16}$ | 625000 | 26938 | 4608 | 1465 | 21 | 6 |
| $k_I^h$ | $> 10^{16}$ | 41667 | 1946 | 358 | 122 | 3 | 2 |
| | | | | $d = 500$ | | | |
| | $h = 1.1$ | $h = 1.2$ | $h = 1.3$ | $h = 1.4$ | $h = 1.5$ | $h = 3$ | $h = 5$ |
| $k_I$ | $> 10^{16}$ | 7500000 | 97508 | 10352 | 2576 | 22 | 6 |
| $k_I^h$ | $> 10^{16}$ | 500000 | 7042 | 805 | 215 | 4 | 2 |

## 3.7 Numerical studies

This section explores how the use of the constants of $a$ and $b$ computed with Proposition 3.5.4, i.e. for $\Sigma = I_d$, affects the probability of declaring a point as an outlier and the required number of projections to reach this decision when we handle other matrices.

This choice is not very relevant in this chapter but it becomes crucial in Chapter 4 when we will face unknown $\boldsymbol{\mu}$ and $\Sigma$. The selection of these constants is justified by Proposition 3.5.2 and since we have seen that the values of $b$'s for the covariance matrices that we handle are resembling (see Table 3.3), thus we expect not too big differences on the $k_\Sigma$'s. All the results are attained with $5000$ replicated simulations. The estimates of $\mathrm{E}(K \;|\|\mathbf{X}\|_\Sigma = rC_n^d)$ (which will be the sample means along the simulations we do) will be denoted $\hat{k}_\Sigma^r$. To ease the notation, we will write $k_I^r, \hat{k}_I^r, k_i^r$ and $\hat{k}_i^r$ when $\Sigma = I_d$ or $\Sigma = \Sigma_i^d, i = 1, \ldots, 4$ for the matrices defined in Section 3.5.2.

Table 3.6 shows the estimate of the probability of declaring as an outlier a rv such that $\|\mathbf{X}\|_\Sigma = C_n^d$ with $\Sigma \neq I_d$ when we use $a$ and $b$ computed for the identity matrix for the case $n = 50$ and different values of $d$ (the complete cases are in Table A.1 in the Appendix). We see that they are reasonable for $d \geq 50$, i.e. around 0.05. However, for $d = 5$ the estimations are up to 1.5 times the expected value 0.05. In any case, the values more distant from 0.05 are those corresponding to $\Sigma = \Sigma_3^5$.

We also see that the sample mean of the number of required projections $\hat{k}_1, \ldots, \hat{k}_4$ are always greater or very close to $k_I$, which illustrates the result shown in Proposition 3.5.3.

Table 3.6.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$, for $n = 50$ and different values of $d$ and $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$ when we use the values of $a$ and $b$ computed with Proposition 3.5.4. We also show the sample means of $K$.

| $d$ | $k_I^1$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 49 | 0.0544 | 51 | 0.0528 | 190 | 0.0492 | 50 | 0.0448 |
| | 100 | 101 | 0.0504 | 103 | 0.0534 | 379 | 0.0510 | 100 | 0.0508 |
| 100 | 50 | 51 | 0.0500 | 50 | 0.0460 | 262 | 0.0474 | 50 | 0.0470 |
| | 100 | 101 | 0.0498 | 99 | 0.0478 | 516 | 0.0466 | 101 | 0.0542 |
| 500 | 50 | 51 | 0.0476 | 49 | 0.0498 | 548 | 0.0416 | 51 | 0.0498 |
| | 100 | 99 | 0.0556 | 98 | 0.0514 | 1184 | 0.0484 | 102 | 0.0552 |
| 1000 | 50 | 50 | 0.0494 | 50 | 0.0520 | 846 | 0.0500 | 50 | 0.0510 |
| | 100 | 101 | 0.0530 | 101 | 0.0515 | 1610 | 0.0415 | 96 | 0.0525 |

Table 3.7.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ for $n = 50$ when we use the values of $a$ and $b$ computed with Proposition 3.5.4. We also show the sample mean of $K$.

| $d$ | $\|X\|_\Sigma$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | $1.2C_n^d$ | 50 | 43 | 0.3124 | 44 | 0.2816 | 43 | 0.3020 | 148 | 0.2070 | 44 | 0.3188 |
| | | 100 | 82 | 0.3638 | 84 | 0.3112 | 82 | 0.3460 | 287 | 0.2122 | 79 | 0.3590 |
| | $2C_n^d$ | 50 | 8 | 0.9220 | 9 | 0.9146 | 8 | 0.9230 | 34 | 0.7224 | 8 | 0.9226 |
| | | 100 | 10 | 0.9463 | 11 | 0.9470 | 10 | 0.9526 | 49 | 0.7786 | 10 | 0.9576 |
| 100 | $1.2C_n^d$ | 50 | 44 | 0.3018 | 46 | 0.2854 | 45 | 0.2890 | 218 | 0.1910 | 45 | 0.2956 |
| | | 100 | 84 | 0.3354 | 87 | 0.3138 | 83 | 0.3454 | 418 | 0.2034 | 83 | 0.3492 |
| | $2C_n^d$ | 50 | 9 | 0.9163 | 9 | 0.9152 | 9 | 0.9136 | 49 | 0.6960 | 9 | 0.9156 |
| | | 100 | 11 | 0.9513 | 12 | 0.9440 | 11 | 0.9452 | 73 | 0.7376 | 11 | 0.9420 |
| 500 | $1.2C_n^d$ | 50 | 46 | 0.2824 | 45 | 0.2818 | 45 | 0.2790 | 488 | 0.1958 | 45 | 0.2808 |
| | | 100 | 86 | 0.3324 | 88 | 0.3106 | 87 | 0.3174 | 932 | 0.1912 | 88 | 0.3294 |
| | $2C_n^d$ | 50 | 9 | 0.9007 | 9 | 0.9124 | 9 | 0.9106 | 106 | 0.6764 | 9 | 0.9142 |
| | | 100 | 11 | 0.9493 | 11 | 0.9426 | 11 | 0.9520 | 164 | 0.7196 | 11 | 0.9488 |
| 1000 | $1.2C_n^d$ | 50 | 45 | 0.2753 | 45 | 0.2840 | 45 | 0.2815 | 660 | 0.1810 | 46 | 0.2755 |
| | | 100 | 86 | 0.3140 | 84 | 0.3095 | 86 | 0.3255 | 1367 | 0.1875 | 83 | 0.3215 |
| | $2C_n^d$ | 50 | 9 | 0.9127 | 9 | 0.9147 | 9 | 0.9193 | 152 | 0.6867 | 9 | 0.9123 |
| | | 100 | 11 | 0.9430 | 11 | 0.9447 | 11 | 0.9443 | 237 | 0.7043 | 12 | 0.9463 |

Table 3.7 shows the estimation of the probability of declaring a point as outlier when its norm is $1.2C_n^d$ and $2C_n^d$ for $n = 50$ and different values of $d$ (complete cases are in Table A.2 in the Appendix). The values of the column $I_d$ are the highest, although with the exception of the case $\Sigma = \Sigma_3^d$, the differences in power are small. Obviously, when $\mathrm{E}(K)$ increases, the probability also increases. Analogously to Table A.1, there is no difference when $n$ increases because the only parameter which depends on it is $C_n^d$. We also see that the case $\Sigma_4^d$ is the most resembling to the identity matrix case. Moreover, with the

exception of $\Sigma = \Sigma_3^d$, the sample means of the number of projections $k_i$ and $k_I$ are also very similar.

Table 3.8.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$, for $n = 50$ using $a$ and $b$ from Proposition 3.5.4 and $k_{\max} = 1/(1 - F(b, C_n^d) + F(a, C_n^d))$. We also show the sample mean of the required projections.

| $d$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 32 | 0.0300 | 33 | 0.0310 | 32 | 0.0306 | 38 | 0.0130 | 33 | 0.0326 |
|  | 100 | 65 | 0.0306 | 64 | 0.0364 | 63 | 0.0344 | 75 | 0.0106 | 64 | 0.0358 |
| 100 | 50 | 32 | 0.0310 | 32 | 0.0326 | 32 | 0.0318 | 40 | 0.0092 | 32 | 0.0268 |
|  | 100 | 64 | 0.0316 | 65 | 0.0280 | 64 | 0.0338 | 78 | 0.0066 | 64 | 0.0326 |
| 500 | 50 | 32 | 0.0316 | 32 | 0.0342 | 32 | 0.0304 | 43 | 0.0044 | 32 | 0.0350 |
|  | 100 | 64 | 0.0362 | 64 | 0.0324 | 64 | 0.0326 | 85 | 0.0054 | 64 | 0.0282 |
| 1000 | 50 | 32 | 0.0337 | 33 | 0.0340 | 33 | 0.0373 | 43 | 0.0053 | 32 | 0.0320 |
|  | 100 | 63 | 0.0322 | 63 | 0.0377 | 64 | 0.0337 | 86 | 0.0027 | 64 | 0.0247 |

Table 3.9.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ for $n = 50$ when we use $a$ and $b$ from Proposition 3.5.4 and $k_{\max} = 1/(1 - F(b, C_n^d) + F(a, C_n^d))$. We also show the sample mean of the required projections.

| $d$ | $\|\mathbf{X}\|_\Sigma$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | $1.2C_n^d$ | 50 | 30 | 0.2126 | 31 | 0.2050 | 31 | 0.2104 | 38 | 0.0628 | 30 | 0.2118 |
|  |  | 100 | 57 | 0.2714 | 59 | 0.2338 | 59 | 0.2488 | 75 | 0.0690 | 59 | 0.2574 |
|  | $2C_n^d$ | 50 | 8 | 0.9186 | 9 | 0.9042 | 8 | 0.9182 | 23 | 0.5714 | 8 | 0.9198 |
|  |  | 100 | 10 | 0.9566 | 11 | 0.9434 | 10 | 0.9540 | 40 | 0.6504 | 10 | 0.9574 |
| 100 | $1.2C_n^d$ | 50 | 30 | 0.2014 | 31 | 0.2010 | 30 | 0.1986 | 39 | 0.0504 | 31 | 0.1956 |
|  |  | 100 | 58 | 0.2466 | 60 | 0.2256 | 59 | 0.2374 | 78 | 0.0466 | 59 | 0.2398 |
|  | $2C_n^d$ | 50 | 8 | 0.9184 | 9 | 0.9106 | 9 | 0.9106 | 27 | 0.4804 | 8 | 0.9172 |
|  |  | 100 | 11 | 0.9468 | 12 | 0.9368 | 11 | 0.9560 | 46 | 0.5752 | 11 | 0.9432 |
| 500 | $1.2C_n^d$ | 50 | 31 | 0.1898 | 31 | 0.1844 | 31 | 0.1878 | 43 | 0.0260 | 31 | 0.1904 |
|  |  | 100 | 60 | 0.2198 | 60 | 0.2210 | 59 | 0.2134 | 84 | 0.0268 | 60 | 0.2244 |
|  | $2C_n^d$ | 50 | 9 | 0.9100 | 9 | 0.9116 | 9 | 0.9128 | 34 | 0.3262 | 9 | 0.9098 |
|  |  | 100 | 11 | 0.9430 | 12 | 0.9442 | 11 | 0.9496 | 64 | 0.3834 | 11 | 0.9468 |
| 1000 | $1.2C_n^d$ | 50 | 31 | 0.1950 | 31 | 0.1880 | 31 | 0.1797 | 44 | 0.0230 | 31 | 0.1993 |
|  |  | 100 | 58 | 0.2381 | 60 | 0.2307 | 60 | 0.2110 | 87 | 0.0227 | 59 | 0.2153 |
|  | $2C_n^d$ | 50 | 9 | 0.9160 | 9 | 0.9130 | 9 | 0.9120 | 36 | 0.2693 | 9 | 0.9097 |
|  |  | 100 | 11 | 0.9450 | 12 | 0.9380 | 11 | 0.9477 | 68 | 0.3363 | 11 | 0.9477 |

Tables 3.8 and 3.9 show the probability of declaring a point with Mahalanobis norm $C_n^d$, $1.2C_n^d$ and $2C_n^d$ respectively as outlier for $n = 50$ and different values of $d$ and the sample mean of the number of projections when we truncate the number of projections by $1/(1 - F(b, C_n^d) + F(a, C_n^d))$ and, if no decision has been taken when the limit is reached, we declare the points as non-outlier. Complete cases are in Tables A.3 and A.4 in the Appendix. For $\Sigma_1^d$, the values of Table 3.8 are around 0.03 and for the rest of $\Sigma_i^d$ those values are similar, with the exception of $\Sigma = \Sigma_3^d$ where they go down to 0.011. This

indicates that the truncation is too conservative, what obviously implies a loss of power as we see when we compare the values of Table 3.9 with Table 3.7. In fact this loss is more noticeable when $d$ increases.

Table 3.10.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$, for $n = 50$ using $a$ and $b$ from Proposition 3.5.4 when we truncate by $k_{\max} = 1/(1 - F(b, C_n^d) + F(a, C_n^d))$ and if no decision is taken, we decide with probability $\gamma = 0.05$. We also show the sample mean of the required projections.

| $d$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 32 | 0.0730 | 32 | 0.0842 | 33 | 0.0816 | 38 | 0.0582 | 33 | 0.0776 |
|  | 100 | 64 | 0.0796 | 64 | 0.0766 | 64 | 0.0842 | 75 | 0.0556 | 65 | 0.0812 |
| 100 | 50 | 32 | 0.0894 | 32 | 0.0796 | 32 | 0.0892 | 40 | 0.0580 | 32 | 0.0790 |
|  | 100 | 64 | 0.0818 | 64 | 0.0760 | 63 | 0.0840 | 78 | 0.0550 | 63 | 0.0804 |
| 500 | 50 | 32 | 0.0744 | 32 | 0.0874 | 32 | 0.0778 | 43 | 0.0562 | 32 | 0.0784 |
|  | 100 | 64 | 0.0902 | 63 | 0.0844 | 63 | 0.0792 | 84 | 0.0502 | 64 | 0.0802 |
| 1000 | 50 | 32 | 0.0786 | 33 | 0.0798 | 32 | 0.0812 | 44 | 0.0538 | 32 | 0.0858 |
|  | 100 | 63 | 0.0796 | 64 | 0.0818 | 64 | 0.0848 | 87 | 0.0572 | 64 | 0.0828 |

Table 3.11.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$, $r = 1.2, 2$ for $n = 50$ using $a$ and $b$ from Proposition 3.5.4, when we truncate by $k_{\max} = 1/(1 - F(b, C_n^d) + F(a, C_n^d))$ and if no decision is taken, we decide with probability $\gamma = 0.05$. We also show the sample mean of the required projections.

| $d$ | $\|\mathbf{X}\|_\Sigma$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | $1.2C_n^d$ | 50 | 30 | 0.2600 | 30 | 0.2448 | 30 | 0.2534 | 38 | 0.1104 | 30 | 0.2596 |
|  |  | 100 | 58 | 0.2958 | 59 | 0.2742 | 59 | 0.2784 | 73 | 0.1152 | 57 | 0.2916 |
|  | $2C_n^d$ | 50 | 8 | 0.9252 | 9 | 0.9106 | 8 | 0.9186 | 24 | 0.585 | 8 | 0.9338 |
|  |  | 100 | 10 | 0.9552 | 11 | 0.9510 | 10 | 0.9542 | 38 | 0.691 | 10 | 0.9514 |
| 100 | $1.2C_n^d$ | 50 | 31 | 0.2322 | 30 | 0.2440 | 31 | 0.2372 | 40 | 0.0900 | 30 | 0.2426 |
|  |  | 100 | 59 | 0.2784 | 60 | 0.2484 | 59 | 0.2600 | 78 | 0.0998 | 58 | 0.2676 |
|  | $2C_n^d$ | 50 | 8 | 0.9226 | 9 | 0.9204 | 9 | 0.917 | 27 | 0.5092 | 8 | 0.9162 |
|  |  | 100 | 11 | 0.9482 | 11 | 0.9438 | 11 | 0.948 | 47 | 0.5996 | 11 | 0.9524 |
| 500 | $1.2C_n^d$ | 50 | 30 | 0.2320 | 31 | 0.2308 | 31 | 0.2226 | 42 | 0.0804 | 31 | 0.2280 |
|  |  | 100 | 59 | 0.2716 | 60 | 0.2610 | 59 | 0.2648 | 84 | 0.0738 | 59 | 0.2628 |
|  | $2C_n^d$ | 50 | 9 | 0.9152 | 9 | 0.9176 | 9 | 0.916 | 33 | 0.3704 | 9 | 0.916 |
|  |  | 100 | 11 | 0.9448 | 12 | 0.9492 | 12 | 0.950 | 64 | 0.4150 | 11 | 0.946 |
| 1000 | $1.2C_n^d$ | 50 | 31 | 0.2250 | 30 | 0.2374 | 31 | 0.2394 | 44 | 0.0746 | 31 | 0.2304 |
|  |  | 100 | 59 | 0.2572 | 60 | 0.2610 | 59 | 0.2578 | 87 | 0.0740 | 59 | 0.2710 |
|  | $2C_n^d$ | 50 | 9 | 0.9204 | 9 | 0.9172 | 9 | 0.9224 | 36 | 0.3230 | 9 | 0.9128 |
|  |  | 100 | 11 | 0.9506 | 12 | 0.9452 | 11 | 0.9486 | 67 | 0.3786 | 12 | 0.9480 |

As we commented in Section 3.5, another possibility when we truncate the number of required projections is that, if no decision has been taken when the limit is reached, we declare the points as outliers with a determined probability $\gamma$. Under these conditions and taking $\gamma = 0.05$, Tables 3.10 and 3.11 give the probability of declaring a point with Mahalanobis norm $C_n^d$, $1.2C_n^d$ and $2C_n^d$ respectively as outlier for $n = 50$ and different values of $d$. It is noticeable that a value as small as this one, provides powers well above $0.05$ under the null for all cases except $\Sigma = \Sigma_3^d$; but even in the last situation, all powers are above $0.05$. In fact, the mean increments of power are about $135\%$, $15\%$ and $20\%$ when $\|\mathbf{X}\|_\Sigma = rC_n^d$ for $r = 1, 1.2, 2$ respectively.

# 4 Outliers detection: unknown parameters

*Nothing in life is to be feared, it's only to be understood.*

— **Marie Curie**

This chapter explores the proposed method when the parameters $\boldsymbol{\mu}$ and $\Sigma$ are unknown (which is the only situation interesting in practice), and therefore, it is not possible to be completely sure if the Definition 1.1.1 holds. Basically we follow the same scheme as in Chapter 3.

## 4.1 Introduction to the method and additional notation

To test (3.1), we follow Algorithm 2. We will consider two assumptions and some notation:

(A1.e) $\mathbf{X}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid rv's with distribution $N_d(\boldsymbol{\mu}, \Sigma)$.

(A2.e) $\mathbf{V}$ and $\mathbf{V}_1, \ldots, \mathbf{V}_n$ are iid rv's with distribution $N_d(\mathbf{0}, I_d)$ which also are independent from the rv's in (A1.e)

The sample mean and the covariance matrix will be denoted by $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$. Initially the centre and the dispersion of the projected sample $\mathbf{X}_1'\mathbf{V}, \ldots, \mathbf{X}_n'\mathbf{V}$ will be estimated by its sample mean, $\hat{\mu}_{\mathbf{V}}$, and standard deviation $\hat{\sigma}_{\mathbf{V}}$. Afterwards, we will use the sample median and MADN, denoted by $\hat{m}_{\mathbf{V}}$ and $\hat{M}_{\mathbf{V}}$ respectively. Furthermore, since both of them may not be unique, the notation $\hat{m}_{\mathbf{v}}$ and $\hat{M}_{\mathbf{v}}$ refers to the choice of any one of them.

We use the same notation as in Chapter 3 but with the estimated quantities, which are denoted adding the sub-index $n$. For instance:

$$Y_n^{\mathbf{V}} := \frac{\mathbf{X}'\mathbf{V} - \hat{\mu}_{\mathbf{V}}}{\hat{\sigma}_{\mathbf{V}}}, \tag{4.1}$$

$$K_n^{a,b}(\Sigma) := \inf\left\{k : |Y_n^k| < a \text{ or } |Y_n^k| > b\right\}. \tag{4.2}$$

The denominator in (4.1) can be zero for some $\mathbf{V}$'s, however the subset of $\mathbf{V}$'s satisfying this has null Lebesgue measure even in the case $d > n$.

As it happened with $Y^{\mathbf{V}}$, the distribution of $Y_n^{\mathbf{V}}$ does not depend on $\boldsymbol{\mu}$ nor on scale. Since our method relies on $Y_n^{\mathbf{V}}$, we can assume w.l.o.g. that $\boldsymbol{\mu} = \mathbf{0}$.

For $\mathbf{m}$, $\mathbf{x} \in \mathbb{R}^d$, $S$ a $d \times d$ semi-positive definite matrix and $\mathbf{V}$ with distribution $N_d(\mathbf{0}, I_d)$, we denote

$$y_{\mathbf{m},S}^{\mathbf{V}} := \frac{(\mathbf{x} - \mathbf{m})' \mathbf{V}}{(\mathbf{V}' S \mathbf{V})^{1/2}},$$

and $\|\mathbf{x}\|_S := \| \left(S^+\right)^{1/2} \mathbf{x}\|$ with $S^+$ is the Moore-Penrose inverse of $S$.

## 4.2 Some properties of the distribution of $Y_n^{\mathbf{V}}$

We begin obtaining explicit expressions for the conditional distribution of $Y_n^{\mathbf{V}}$ given $\|\mathbf{X}\|_\Sigma$ in Proposition 4.2.2. Then, Proposition 4.2.3 gives an expression of the cdf of the standardized random projection of a given $d$-dimensional vector. In this proposition we suppose that $S$ is diagonal, which entails no loss of generality, since a rotation of the coordinates axes allows us to obtain this kind of matrix. Note that the only assumption on the number of non-null eigenvalues of $S$ is the existence of at least two strictly positive ones.

To prove Proposition 4.2.2, the equality $Y_n^{\mathbf{V}} = \left(\frac{\mathbf{X}'\mathbf{V}}{\sigma_{\mathbf{V}}} - \frac{\hat{\mu}_{\mathbf{V}}}{\sigma_{\mathbf{V}}}\right)\frac{\sigma_{\mathbf{V}}}{\hat{\sigma}_{\mathbf{V}}}$ leads us to consider the following rv's:

$$Y_1 := \frac{\mathbf{X}'\mathbf{V}}{\sigma_{\mathbf{V}}}, \quad Y_2 := \frac{\hat{\mu}_{\mathbf{V}}}{\sigma_{\mathbf{V}}}, \quad Y_3 := \frac{\sigma_{\mathbf{V}}}{\hat{\sigma}_{\mathbf{V}}}. \tag{4.3}$$

We next obtain the pdf's of those rv's given that $\|\mathbf{X}\|_\Sigma = t$, with $t > 0$. Since we compute the conditional pdf's given the norm of the point, the rv $Y_1$ does not follow a standard normal distribution. Recall also that the sample mean and the sample variance are calculated using only the sample and therefore, $Y_1$ is the only rv which depends on $\mathbf{X}$, the point we want to classify as outlier or not. Consequently, the distribution of $Y_2$ and $Y_3$ do not depend on $t$.

**Lemma 4.2.1.** *Under assumptions* (A1.e) *and* (A2.e)*, the pdf's of the rv's* $Y_1$*,* $Y_2$*,* $Y_3$*, defined in* (4.3) *given that* $\|\mathbf{X}\|_\Sigma = t$ *with* $t > 0$*, are*

$$f^t_{Y_1}(u) = \left(\mathrm{B}\left(\tfrac{d-1}{2},\tfrac{1}{2}\right)\right)^{-1} t^{2-d}(t^2-u^2)^{(d-3)/2}, \textit{ if } u \in [-t,t] \textit{ and null otherwise,}$$

$$f_{Y_2}(u) = \left(\frac{n}{2\pi}\right)^{1/2} \exp\{-nu^2/2\}, u \in \mathbb{R},$$

$$f_{Y_3}(u) = \frac{(n-1)^{(n-1)/2}}{2^{(n-3)/2}\Gamma\left(\frac{n-1}{2}\right)} u^{-n}\exp\left\{-\tfrac{n-1}{2u^2}\right\}, \textit{ for } u \in [0,\infty) \textit{ and null otherwise.}$$

*Proof.* Firstly, fix $\mathbf{V} = \mathbf{v}$. Using Lemma 2.5.4, it is easily seen that the pdf of $Y_1$ given $\mathbf{v}$ and that $\|\mathbf{X}\|_\Sigma = t$ coincides with the expression we propose for $f^t_{Y_1}$. Secondly, for $Y_2$, since $\hat{\mu}_\mathbf{v}$ follows a $N_1(0,\sigma^2_\mathbf{v}/n)$ distribution, then the rv $\hat{\mu}_\mathbf{v}/\sigma_\mathbf{v}$ follows a $N_1(0,1/n)$ distribution. For $Y_3$, it is known that $\hat{\sigma}^2_\mathbf{v}(n-1)/\sigma^2_\mathbf{v}$ follows a $\chi^2_{n-1}$ distribution. Then, a change of variable gives that the pdf of $Y_3$ given $\mathbf{v}$ is $f_{\chi^2_{n-1}}((n-1)u^{-2})2(n-1)u^{-3}$, which writing the expression of $f_{\chi^2_{n-1}}$ gives the function we propose for $f_{Y_3}$. The result follows because none of those distributions depend on the chosen $\mathbf{v}$. □

**Proposition 4.2.2.** *Under assumptions* (A1.e) *and* (A2.e)*, the cdf of* $Y^\mathbf{V}_n$ *given that* $\|\mathbf{X}\|_\Sigma = t$*, with* $t > 0$*, does not depend on* $\Sigma$ *and its value is*

$$\mathbf{P}\left(Y^\mathbf{V}_n < r \,\middle|\, \|\mathbf{X}\|_\Sigma = t\right) = \begin{cases} -\tau \displaystyle\int_{-\infty}^{r}\int_{-\infty}^{0}\int_{-t}^{t} g_t(s,x,z)\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}z, & r < 0,\\ \frac{1}{2} + \tau \displaystyle\int_{0}^{r}\int_{0}^{\infty}\int_{-t}^{t} g_t(s,x,z)\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}z, & r > 0,\end{cases}$$

*where* $g_t(s,x,z) := \frac{x^{n-1}}{z^n}\exp\left\{\frac{-(n-1)x^2}{2z^2}\right\}(t^2-s^2)^{d^\star}\exp\left\{\frac{-n}{2}(s-x)^2\right\}$*,* $d^\star := (d-3)/2$ *and* $\tau := t^{2-d}\sqrt{\frac{n}{2\pi}}(n-1)^{\frac{n-1}{2}}/\left(2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)\mathrm{B}\left(\frac{d-1}{2},\frac{1}{2}\right)\right)$*.*

*Proof.* The rv's $Y_1$, $Y_2$, $Y_3$ defined in (4.3) are conditionally independent given $\mathbf{V}$. If $r < 0$, then the pdf of the rv $Y^\mathbf{V}_n$ given $\|\mathbf{X}\|_\Sigma = t$ is:

$$\begin{aligned} f^t_{Y^\mathbf{V}}(r) &= f^t_{(Y_1-Y_2)Y_3}(r)\\ &= \int_\mathbb{R} f^t_{Y_1-Y_2}(x) f_{Y_3}(r/x)|x|^{-1}\,\mathrm{d}x\\ &= \int_\mathbb{R} f_{Y_3}(r/x)|x|^{-1}\left(\int_\mathbb{R} f^t_{Y_1}(s) f_{Y_2}(s-x)\,\mathrm{d}s\right)\mathrm{d}x. \end{aligned}$$

It suffices then to write the expressions of the pdf's of the rv's $Y_1$, $Y_2$, $Y_3$, given by Lemma 4.2.1, to obtain the first equality of this proposition. The reasoning when $r$ is positive is identical. □

We omit the proof of the following proposition because it is the same to that in Proposition 3.2.7 replacing $h_\pm(\cdot)$, $\psi_{\mathbf{V}}$ and $\phi_{\mathbf{V}}$ by $\hat{h}_\pm(\cdot)$, $\hat{\psi}_{\mathbf{V}}$ and $\hat{\phi}_{\mathbf{V}}$ respectively.

**Proposition 4.2.3.** *Let $\mathbf{x} = (x_1, \dots, x_d)' \in \mathbb{R}^d$. Assume that $\mathbf{m} = (m_1, \dots, m_d)' \in \mathbb{R}^d$, $S$ is diagonal with eigenvalues $0 < s_1^2 \leq \dots \leq s_\ell^2$ and $0 = s_{\ell+1}^2 = \dots = s_d^2$ with $2 \leq \ell \leq d$ and that $t := \|\mathbf{x} - \mathbf{m}\|_S > 0$. If $\mathbf{V}$ is uniformly distributed on $\Omega^{d-1}$, then the distribution of $y^{\mathbf{V}}_{\mathbf{m},S}$ is supported by $[-t, t]$ and*

$$\mathbf{P}(y^{\mathbf{V}}_{\mathbf{m},S} \leq z) = \begin{cases} \tau \int_{A_-^{\mathbf{v}}} \Delta(z) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2} \, \mathrm{d}\mathbf{v}_{-1}, & -t < z < -\frac{|u_1|}{s_1}, \\ \frac{1}{2} - \operatorname{sign}(z)\tau \int_{A_+^{\mathbf{v}}} \Delta(z) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2} \, \mathrm{d}\mathbf{v}_{-1}, & -\frac{|u_1|}{s_1} \leq z \leq \frac{|u_1|}{s_1}, \\ 1 - \tau \int_{A_+^{\mathbf{v}}} \Delta(z) e^{-\frac{1}{2}\sum_{i=2}^d v_i^2} \, \mathrm{d}\mathbf{v}_{-1}, & \frac{|u_1|}{s_1} < z < t, \end{cases}$$

*with $\Delta(z) := \operatorname{erf}\left(\hat{h}_+(z)/\sqrt{2}\right) - \operatorname{erf}\left(\hat{h}_-(z)/\sqrt{2}\right)$ where $\operatorname{erf}(\cdot)$ is the error function, $\hat{h}_\pm(z) = \left(u_1\hat{\psi}_{\mathbf{v}} \pm |z|\sqrt{(u_1)^2\hat{\varphi}_{\mathbf{v}} + s_1^2\hat{\psi}_{\mathbf{v}}^2 - s_1^2 z^2 \hat{\varphi}_{\mathbf{v}}}\right) / \left(s_1^2 z^2 - (u_1)^2\right)$, $\tau := (2^{\frac{d+3}{2}}\pi^{\frac{d-1}{2}})^{-1}$, $A_+^{\mathbf{v}} := \{\mathbf{v}_{-1} : \hat{\psi}_{\mathbf{v}} > 0\}$, $A_-^{\mathbf{v}} := \{\mathbf{v}_{-1} : \hat{\psi}_{\mathbf{v}} < 0\}$ and $\mathbf{v}_{-1} := (v_2, \dots, v_d)'$, with $\hat{\psi}_{\mathbf{v}} := u_2 v_2 + \dots + u_d v_d$, $\hat{\varphi}_{\mathbf{v}} := s_2^2 v_2^2 + \dots + s_\ell^2 v_\ell^2$, and $u_i = x_i - m_i$ for $i = 1, \dots, d$.*

Propositions 4.2.5 and 4.2.11 respectively give an expression and some properties of $F_{n,\Sigma}(a, b, t)$ which will help to determine $a$ and $b$ in practice. As in Chapter 3, given $\alpha \in (0, 1)$, the intended error of type I, we want to obtain $0 < a \leq b$ such that $\mathbf{P}(K_n < \infty) = 1$, and

$$\sup_{t \leq C_n^d} F_{n,\Sigma}(a, b, t) = \alpha. \tag{4.4}$$

Similar reasons given in Chapter 3 lead us to exclude the case $a = 0$.

We include no proof of Proposition 4.2.5 because it is analogous to that in Proposition 3.2.11 but conditioning to the point $\mathbf{X}$ and the sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ instead of conditioning only to the point and taking into account Lemma 4.2.4, which is obvious.

**Lemma 4.2.4.** *Let $\mathbf{V}_1, \dots, \mathbf{V}_k$ be iid rv's, then $Y_n^1, \dots, Y_n^k$ defined in (4.1) are conditionally iid given the $d$-dimensional vectors $\mathbf{X}$ and $\mathbf{X}_1, \dots, \mathbf{X}_n$.*

**Proposition 4.2.5.** *Under assumptions $(\mathrm{A1.e})$ and $(\mathrm{A2.e})$, suppose that $a, b, t$ are strictly positive constants such that $a \leq b$, then*

$$F_{n,\Sigma}(a, b, t) = \int_{\Omega_\Sigma^{d-1}(t)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^{d^2}} g_a^b(\mathbf{x}, \mathbf{m}, S)(S) f_t(\mathbf{x}) \, \mathbf{P}_{\hat{\Sigma}}(\mathrm{d}S) \mathbf{P}_{\hat{\mu}}(\mathrm{d}\mathbf{m}) \, \mathrm{d}\mathbf{x},$$

*where $g_a^b(\mathbf{x}, \mathbf{m}, S) := \mathbf{P}\left(|y^{\mathbf{V}}_{\mathbf{m},S}| > b\right) / \left(\mathbf{P}\left(|y^{\mathbf{V}}_{\mathbf{m},S}| > b\right) + \mathbf{P}\left(|y^{\mathbf{V}}_{\mathbf{m},S}| < a\right)\right)$, $\mathbf{P}_{\hat{\Sigma}}$ is the Wishart distribution with parameters $n$ and $\Sigma$, and $\mathbf{P}_{\hat{\mu}}$ is the $N_d\left(\mathbf{0}, n^{-1}\Sigma\right)$.*

From Proposition 4.2.5, it is clear the following corollary.

**Corollary 4.2.6.** *Under the assumptions in Proposition 4.2.5, we have that*

$$\mathbf{P}\left(\left|Y_n^{K_n}\right| > b \mid \|\mathbf{X}\|_\Sigma = t, \mathbf{X}_1, \dots, \mathbf{X}_n\right) = \int_{\Omega_\Sigma^{d-1}(t)} g_a^b(\mathbf{x}, \hat{\boldsymbol{\mu}}, \hat{\Sigma}) f_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Proposition 4.2.7 leads to an easier expression of Proposition 4.2.5 for $\Sigma = I_d$, provided in Corollary 4.2.8. The quantities in such corollary can be computed from Proposition 4.2.2. We exclude the proof of Proposition 4.2.7 because it is analogous to that in Proposition 3.2.10 conditioning to $(\mathbf{X}, \hat{\boldsymbol{\mu}}, \hat{\Sigma})$ instead of only $\mathbf{X}$.

**Proposition 4.2.7.** *Under the assumptions* (A1.e) *and* (A2.e) *the rv's $Y_n^1, \dots, Y_n^k$ defined in* (4.1) *are conditionally independent given $\|\mathbf{X}\|_\Sigma$ if and only if $\Sigma = I_d$.*

**Corollary 4.2.8.** *Under assumptions in Proposition 4.2.5. If $\Sigma = I_d$, then*

$$F_{n,\Sigma}(a,b,t) = \frac{\mathbf{P}\left(|Y_n^{\mathbf{V}}| > b \| \|\mathbf{X}\| = t\right)}{1 - \mathbf{P}\left(|Y_n^{\mathbf{V}}| \in (a,b) \| \|\mathbf{X}\| = t\right)}.$$

Proposition 4.2.11 proves the monotonicity on $t$ of $F_{n,\Sigma}(a,b,t)$. Before proving and stating Proposition 4.2.11, we need some previous results, but notice firstly that such proposition allows us to write (4.4) as

$$F_{n,\Sigma}(a,b,C_n^d) = \alpha. \tag{4.5}$$

**Lemma 4.2.9.** *Let $d > 1$ and let $S$ and $\Sigma$ be $d \times d$ semi-positive symmetric matrices and $\mathbf{Z}$ be a $d$-dimensional rv. The function $r \mapsto f(r)$ is increasing, where*

$$f(r) := \mathbf{P}(\Sigma^{1/2}\mathbf{Z} \textit{ be declared outlier w.r.t. } N_d(\mathbf{0}, S) \mid \|\Sigma^{1/2}\mathbf{Z}\| = r).$$

*Proof.* Let $\mathbf{z} \in \mathbb{R}^d$ and let $r = \|\Sigma^{1/2}\mathbf{z}\|$, the same reasoning as in Theorem 3.2.14 (taking $\Sigma^{1/2}\mathbf{z}$ instead of $\mathbf{x}$) leads us to the fact that $\mathbf{P}(|(\Sigma^{1/2}\mathbf{z})'\mathbf{V}|/\Sigma_\mathbf{V} > b)$ increases with $r$ and it does not depend on the specific value of $\mathbf{z}$. A similar reasoning implies that the map $r \mapsto \mathbf{P}\left(\frac{|(\Sigma^{1/2}\mathbf{z})'\mathbf{V}|}{\Sigma_\mathbf{V}} < a\right)$ decreases on $r$ and the result follows from:

$$\begin{aligned} f(r) &= \int \frac{\mathbf{P}\left(\frac{|(\Sigma^{1/2}\mathbf{z})'\mathbf{V}|}{\Sigma_\mathbf{V}} > b\right)}{\mathbf{P}\left(\frac{|(\Sigma^{1/2}\mathbf{z})'\mathbf{V}|}{\Sigma_\mathbf{V}} > b\right) + \mathbf{P}\left(\frac{|(\Sigma^{1/2}\mathbf{z})'\mathbf{V}|}{\Sigma_\mathbf{V}} < a\right)} \mathbf{P}_{\mathbf{Z} \mid \|\Sigma^{1/2}\mathbf{Z}\|}(d\mathbf{z}) \\ &= \frac{\mathbf{P}\left(\frac{|V_1|}{\Sigma_\mathbf{V}} > \frac{b}{r}\right)}{\mathbf{P}\left(\frac{|V_1|}{\Sigma_\mathbf{V}} > \frac{b}{r}\right) + \mathbf{P}\left(\frac{|V_1|}{\Sigma_\mathbf{V}} < \frac{a}{r}\right)}. \end{aligned}$$

□

**Lemma 4.2.10.** *Let $\delta$, $c \in (0,1)$, $\Sigma$ be a semi positive definite symmetric matrix and $\mathbf{Z}$ be a rv with distribution $N_d(\mathbf{0}, \delta I_d)$. If $\mathbf{x} \neq 0$, then for any $g$ increasing function*

$$E[g(\|\Sigma^{1/2}(\mathbf{Z}+\mathbf{x})\|)] \geq E[g(\|\Sigma^{1/2}(\mathbf{Z}+c\mathbf{x})\|)].$$

*Proof.* Taking $h(w) = w$, the second part in Corollary 2 in [6] gives that if $\mathbf{x} \neq \mathbf{0}$ then

$$\mathbf{P}(\|\Sigma^{1/2}(\mathbf{Z}+c\mathbf{x})\| \leq r) \geq \mathbf{P}(\|\Sigma^{1/2}(\mathbf{Z}+\mathbf{x})\| \leq r).$$

From here, the lemma trivially follows. □

**Proposition 4.2.11.** *Under assumptions* (A1.e) *and* (A2.e)*, if $a$, $b$ and $t$ are strictly positive constants such that $a \leq b$, then the function $F_{n,\Sigma}(a,b,t)$ is strictly increasing in $t$.*

*Proof.* Given $\mathbf{z} \in \Omega^{d-1}$ and $S \in \mathbb{R}^{d^2}$, let us consider

$$G_{\mathbf{z},S}(t) = \int_{\mathbb{R}^d} \frac{\mathbf{P}\left(\frac{|(t\Sigma^{1/2}\mathbf{z}-\mathbf{y})'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} > b\right)}{\mathbf{P}\left(\frac{|(t\Sigma^{1/2}\mathbf{z}-\mathbf{y})'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} > b\right) + \mathbf{P}\left(\frac{|(t\Sigma^{1/2}\mathbf{z}-\mathbf{y})'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} < a\right)} f_{\hat{\boldsymbol{\mu}}}(\mathbf{y})\,\mathrm{d}\mathbf{y},$$

where $\hat{\boldsymbol{\mu}}$ follows a $N_d(\mathbf{0}, \Sigma/n)$ distribution. The proposition will be proved if we show that $G_{\mathbf{z},S}(t)$ is increasing because

$$F_{n,\Sigma}(a,b,t) = \frac{1}{\omega_1^d} \int_{\Omega_1^{d-1}} \int_{\mathbb{R}^{d^2}} G_{\mathbf{z},S}(t)\, \mathbf{P}_{\hat{\Sigma}}(\mathrm{d}S)\,\mathrm{d}\mathbf{z}.$$

Given the rv $\mathbf{Z}$, let $\{\mathbf{Z}$ outlier wrt $N_d(\mathbf{0},S)\}$ denote the set where $\mathbf{Z}$ is declared outlier with respect to $N_d(\mathbf{0},S)$. Then

$$\mathbf{P}\{\mathbf{Z} \text{ outlier wrt } N_d(\mathbf{0},S)\} = \int \frac{\mathbf{P}\left(\frac{|\mathbf{z}'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} > b\right)}{\mathbf{P}\left(\frac{|\mathbf{z}'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} > b\right) + \mathbf{P}\left(\frac{|\mathbf{z}'\mathbf{V}|}{\|S^{1/2}\mathbf{V}\|} < a\right)} \mathbf{P}_{\mathbf{Z}}(\mathrm{d}\mathbf{z}).$$

If we take $\hat{\boldsymbol{\mu}}_n^{I_d} = \Sigma^{-1/2}\hat{\boldsymbol{\mu}}$ and $f$ is the function defined in Lemma 4.2.9,

$$\begin{aligned}
G_{\mathbf{z},S}(t) &= \mathbf{P}(\Sigma^{1/2}(\hat{\boldsymbol{\mu}}_n^{I_d} + t\mathbf{z}) \text{ outlier wrt } N_d(\mathbf{0},S)) \\
&= \int_0^\infty \mathbf{P}(\Sigma^{1/2}(\hat{\boldsymbol{\mu}}_n^{I_d} + t\mathbf{z}) \text{ outlier wrt } N_d(\mathbf{0},S) | \|\Sigma^{1/2}(\hat{\boldsymbol{\mu}}_n^{I_d} + t\mathbf{z})\| = r)\mathbf{P}(\mathrm{d}r) \\
&= \mathrm{E}[f(\|\Sigma^{1/2}(\hat{\boldsymbol{\mu}}_n^{I_d} + t\mathbf{z})\|)],
\end{aligned}$$

and the result is deduced from Lemmas 4.2.9 and 4.2.10. □

## 4.3 Moments of $K_n$

In this section we state Proposition 4.3.1 which is the version of Proposition 3.3.1 for $\Sigma$ unknown. Its proof is not included because it is similar to that of Proposition 4.2.5.

**Proposition 4.3.1.** *Under assumptions* (A1.e) *and* (A2.e)*, assume that* $a, b$ *and* $t$ *are strictly positive numbers such that* $a \leq b$ *and consider* $K_n$ *defined as in* (4.2)*, then*

$$E\left(K_n | \, \|\mathbf{X}\|_\Sigma = t\right) = \iiint\limits_{\mathcal{D}} \bar{g}_a^b(\mathbf{x}, \mathbf{m}, S) f_t(\mathbf{x})\, \mathbf{P}_{\hat{\Sigma}}(\mathrm{d}S) \mathbf{P}_{\hat{\boldsymbol{\mu}}}(\mathrm{d}\mathbf{m}) \mathrm{d}\mathbf{x},$$

$$\begin{aligned}\mathrm{Var}\left(K_n | \, \|\mathbf{X}\|_\Sigma = t\right) = &\iiint\limits_{\mathcal{D}} \bar{g}_a^b(\mathbf{x}, \mathbf{m}, S)(2\bar{g}_a^b(\mathbf{x}, \mathbf{m}, S) - 1) f_t(\mathbf{x})\, \mathbf{P}_{\hat{\Sigma}}(\mathrm{d}S) \mathbf{P}_{\hat{\boldsymbol{\mu}}}(\mathrm{d}\mathbf{m})\, \mathrm{d}\mathbf{x} \\ &- \left( \iiint\limits_{\mathcal{D}} \bar{g}_a^b(\mathbf{x}, \mathbf{m}, S) f_t(\mathbf{x})\, \mathbf{P}_{\hat{\Sigma}}(\mathrm{d}S) \mathbf{P}_{\hat{\boldsymbol{\mu}}}(\mathrm{d}\mathbf{m})\, \mathrm{d}\mathbf{x} \right)^2,\end{aligned}$$

*where* $\bar{g}_a^b(\mathbf{x}, \mathbf{m}, S) = 1/\left(\mathbf{P}\left(|y^{\mathbf{V}}_{\mathbf{m},S}| > b\right) + \mathbf{P}\left(|y^{\mathbf{V}}_{\mathbf{m},S}| < a\right)\right)$*,* $\mathcal{D} := \Omega_\Sigma^{d-1}(t) \times \mathbb{R}^d \times \mathbb{R}^{d^2}$*, and* $\mathbf{P}_{\hat{\Sigma}}$ *and* $\mathbf{P}_{\hat{\boldsymbol{\mu}}}$ *are the Wishart distribution with parameters* $n$ *and* $\Sigma$*, and the* $N_d(\mathbf{0}, \Sigma/n)$*, respectively.*

Propositions 4.2.7 and 4.3.1 allow to obtain Corollary 4.3.2.

**Corollary 4.3.2.** *Under assumptions in Proposition 4.3.1, if* $\Sigma = I_d$*, then*

$$E\left(K_n | \, \|\mathbf{X}\| = t\right) = \frac{1}{1 - \mathbf{P}\left(|Y_n^{\mathbf{V}}| \in (a,b) | \|\mathbf{X}\| = t\right)},$$

$$\mathrm{Var}\left(K_n | \, \|\mathbf{X}\| = t\right) = \frac{\mathbf{P}\left(|Y_n^{\mathbf{V}}| < b | \|\mathbf{X}\| = t\right) - \mathbf{P}\left(|Y_n^{\mathbf{V}}| < a | \|\mathbf{X}\| = t\right)}{1 - \left(\mathbf{P}\left(|Y_n^{\mathbf{V}}| \in (a,b) | \|\mathbf{X}\| = t\right)\right)^2}.$$

It is clear from Corollary 4.3.2 that $\mathrm{E}(K_n | \|\mathbf{X}\| = t)$ and $\mathrm{Var}(K_n | \|\mathbf{X}\| = t)$ do not depend on either the specific value of $t$ or the dimension, but rather, only on the probability $\mathbf{P}\left(|Y_n^{\mathbf{V}}| \in (a,b) | \|\mathbf{X}\| = t\right)$. We do not include their graphical representation because of their similarity with those in Figure 3.2.

## 4.4 Robust versions of $K_n$ and $Y_n^{\mathbf{V}}$. Additional notation

The results in Sections 4.2 and 4.3 fix the problem when we have a clean sample and we want to decide whether a point which is not in the sample is an outlier or not. Regrettably this setting is unrealistic and often the sample at hand will contain outliers which may considerably affect the sample mean and standard deviation. Because of this, we propose to replace $\hat{\mu}_{\mathbf{V}}$ and $\hat{\sigma}_{\mathbf{V}}$ in (4.1) by some robust counterparts. As stated before, our selections are the median, $m_{\mathbf{V}}$, and the MADN, $M_{\mathbf{V}}$.

To reflect the change, we replace $Y_n^{\mathbf{V}}$ and $K_n$ by $\tilde{Y}_n^{\mathbf{V}}$ and $\tilde{K}_n$, respectively. Notice that in this case (4.5) becomes

$$\mathbf{P}\left(\left|\tilde{Y}_n^{\tilde{K}_n}\right| > b \,\Big|\, \|\mathbf{X}\|_\Sigma = C_n^d\right) = \alpha. \tag{4.6}$$

As it usually occurs with robust estimators (see, for instance, Cerioli *et al.* [28] or Becker and Gather [15]), it is difficult to obtain the conditional exact distribution of $\tilde{Y}_n^{\tilde{K}_n}$. Because of this we prove, in Theorem 4.4.6, that asymptotically on $n$ this distribution coincides with that of $Y_n^{K_n}$. Afterwards, in Section 4.7, we will present simulations suggesting that this approximation gives acceptable results in many cases for small sample sizes and arbitrary values of the dimension.

Theorem 4.4.5 is an auxiliary result to obtain Theorem 4.4.6. However, we consider that it could have some independent interest. We first state some additional notation.

**Notation.** Under assumptions (A1.e) and (A2.e), denote $\mathbf{Q}_{\mathbf{V}}$ and $\bar{\mathbf{Q}}_{\mathbf{V}}$ the probability distribution of $\mathbf{X}'\mathbf{V}$ and of $|\mathbf{X}'\mathbf{V}|$, respectively, and let us consider the following sets:

$$\begin{aligned}
R_{\mathbf{V}}^n &:= \{\mathbf{X}_1'\mathbf{V}, \dots, \mathbf{X}_n'\mathbf{V}\}\\
T_{\mathbf{V}}^n &:= \{|\mathbf{X}_1'\mathbf{V}|, \dots, |\mathbf{X}_n'\mathbf{V}|\}\\
S_{\mathbf{V}}^n &:= \{|\mathbf{X}_1'\mathbf{V} - m_{\mathbf{V}}|, \dots, |\mathbf{X}_n'\mathbf{V} - m_{\mathbf{V}}|\}\\
\hat{S}_{\mathbf{V}}^n &:= \{|\mathbf{X}_1'\mathbf{V} - \hat{m}_{\mathbf{V}}|, \dots, |\mathbf{X}_n'\mathbf{V} - \hat{m}_{\mathbf{V}}|\}.
\end{aligned}$$

Given $S \subset \mathbb{R}$ finite (resp. the real rv $X$) and $\alpha \in (0,1)$, $m(S)$ and $M(S)$ (resp. $m(X)$ and $M(X)$) denote the sets of its medians and MADNs; $[\underline{q}_\alpha(S), \bar{q}_\alpha(S)]$ (resp. $[\underline{q}_\alpha(X), \bar{q}_\alpha(X)]$) is the interval of the $\alpha$-quantiles of $S$ (resp. $X$). We define the interval $[\underline{M}_\alpha(S), \bar{M}_\alpha(S)] := \cup_{m\in m(S)} [\underline{q}_\alpha(|S-m|), \bar{q}_\alpha(|S-m|)]$, similarly for $[\underline{M}_\alpha(X), \bar{M}_\alpha(X)]$. Thus, $m(S) = [\underline{q}_{\frac{1}{2}}(S), \bar{q}_{\frac{1}{2}}(S)]$ and $M(S) = [\underline{M}_{\frac{1}{2}}(S), \bar{M}_{\frac{1}{2}}(S)]$.

According to the assumptions in Section 2.1, all random quantities we handle are defined on $(\Upsilon, \mathcal{A}, \mathbf{P})$. Therefore all of them depend on some $\omega \in \Upsilon$. Here we will be sometimes interested in making this dependence explicit; in those cases, $\omega$ will appear as super-index as in $\hat{m}_{\mathbf{V}}^\omega$, or in $S_{\mathbf{V}}^{n,\omega}$.

**Lemma 4.4.1.** *Let $U$ and $V$ be two real rv's such that there exist $\delta$ and $\gamma$ with $\mathbf{P}\{|U-V| \le \delta\} \ge 1-\gamma$. Then for every $\alpha \in [\gamma, 1-\gamma]$,*

$$[\underline{q}_\alpha(U), \bar{q}_\alpha(U)] \subset [\underline{q}_{\alpha-\gamma}(V) - \delta, \bar{q}_{\alpha+\gamma}(V) + \delta] \tag{4.7}$$

$$[\underline{M}_\alpha(U), \bar{M}_\alpha(U)] \subset [\underline{M}_{\alpha-\gamma}(V) - (2\delta + \delta^*_\gamma), \bar{M}_{\alpha+\gamma}(V) + (2\delta + \delta^*_\gamma)], \tag{4.8}$$

*where* $\delta^*_\gamma = \max\{\underline{q}_{\frac{1}{2}}(V) - \underline{q}_{\frac{1}{2}-\gamma}(V), \bar{q}_{\frac{1}{2}+\gamma}(V) - \bar{q}_{\frac{1}{2}}(V)\}$.

*Proof.* Let $q \in [\underline{q}_\alpha(U), \bar{q}_\alpha(U)]$. Then, by definition of quantile:

$$\alpha \le \mathbf{P}\{U \le q\} \le \mathbf{P}[|U-V| \le \delta, U \le q] + \mathbf{P}\{|U-V| > \delta\} \le \mathbf{P}\{V \le q+\delta\} + \gamma.$$

Hence $\alpha - \gamma \le \mathbf{P}\{V \le q+\delta\}$, which implies $q+\delta \ge \underline{q}_{\alpha-\gamma}(V)$. And then $\underline{q}_\alpha(U) \ge \underline{q}_{\alpha-\gamma}(V) - \delta$. Analogously, we can prove $\bar{q}_\alpha(U) \le \bar{q}_{\alpha+\gamma}(V) + \delta$ and (4.7) is shown.

To prove (4.8), consider $m^U \in m(U)$. Take $\alpha = 1/2$ in (4.7). There exits $m^V \in m(V)$ such that $|m^U - m^V| \le \delta + \delta^*_\gamma$. Hence, if $|U - V| \le \delta$, then

$$\Big| |U - m^U| - |V - m^V| \Big| \le |U - V| + |m^U - m^V| \le 2\delta + \delta^*_\gamma,$$

and (4.8) follows from the definition of MAD and (4.7). □

**Corollary 4.4.2.** *Under the hypotheses in Lemma 4.4.1,*

$$m(U) \subset \left[\underline{q}_{\frac{1}{2}-\gamma}(V) - \delta, \bar{q}_{\frac{1}{2}+\gamma}(V) + \delta\right].$$

If we apply Lemma 4.4.1 to rv's uniformly distributed on finite sets with the same cardinal, we obtain the following corollary.

**Corollary 4.4.3.** *If $S = \{s_1, \ldots, s_n\} \subset \mathbb{R}$ and $R = \{r_1, \ldots, r_n\} \subset \mathbb{R}$ satisfy that there exist $\delta, \gamma$ such that $\#\{i : |s_i - r_i| \le \delta\} \ge n(1-\gamma)$, then for every $\alpha \in [\gamma, 1-\gamma]$,*

$$\begin{aligned} [\underline{q}_\alpha(S), \bar{q}_\alpha(S)] &\subset [\underline{q}_{\alpha-\gamma}(R) - \delta, \bar{q}_{\alpha+\gamma}(R) + \delta] \\ [\underline{M}_\alpha(S), \bar{M}_\alpha(S)] &\subset [\underline{M}_{\alpha-\gamma}(R) - (2\delta + \delta^*_\gamma), \bar{M}_{\alpha+\gamma}(R) + (2\delta + \delta^*_\gamma)], \end{aligned} \tag{4.9}$$

*where* $\delta^*_\gamma = \max\left\{\underline{q}_{\frac{1}{2}}(R) - \underline{q}_{\frac{1}{2}-\gamma}(R), \bar{q}_{\frac{1}{2}+\gamma}(R) - \bar{q}_{\frac{1}{2}}(R)\right\}$.

**Lemma 4.4.4.** *For every $\mathbf{v} \in \Omega^{d-1}$, there exists a probability one set $A \in \mathcal{A}$ such that for every $\omega \in A$ and $\gamma \in (0, 1/2)$*

$$\sup_{\alpha\in(\gamma,1-\gamma)} \left(\max\left\{\left|\underline{q}_\alpha(R_{\mathbf{v}}^{n,\omega}) - q_\alpha(\mathbf{Q}_{\mathbf{v}})\right|, \left|\bar{q}_\alpha(R_{\mathbf{v}}^{n,\omega}) - q_\alpha(\mathbf{Q}_{\mathbf{v}})\right|\right\}\right) \to 0,$$

$$\sup_{\alpha\in(\gamma,1-\gamma)} \left(\max\left\{\left|\underline{q}_\alpha(T_{\mathbf{v}}^{n,\omega}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}})\right|, \left|\bar{q}_\alpha(T_{\mathbf{v}}^{n,\omega}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}})\right|\right\}\right) \to 0.$$

*Proof.* Since $\mathbf{X}'\mathbf{v}$ is a normal rv, then the assumptions in Corollary 1.4.3 in Csörgő [31] are satisfied. Therefore, (1.4.24) in Csörgő [31] holds and the first statement is verified. A similar reasoning leads to the second one. □

**Theorem 4.4.5.** *Under assumptions* (A1.e) *and* (A2.e)*, there exists* $A_0 \in \mathcal{A}$ *with* $\mathbf{P}(A_0) = 1$ *such that if* $\omega \in A_0$*, then, as* $n \to \infty$*,*

$$\sup_{\mathbf{v}\in\Omega^{d-1}} |\hat{m}_{\mathbf{v}}^{\omega} - m_{\mathbf{v}}| \to 0 \quad \textit{and} \quad \sup_{\mathbf{v}\in\Omega^{d-1}} |\hat{M}_{\mathbf{v}}^{\omega} - M_{\mathbf{v}}| \to 0. \tag{4.10}$$

*Proof.* We first apply the Glivenko-Cantelli Theorem to the iid rv's $\{\|\mathbf{X}_i\|\}$ and we have that a.s., as $n \to \infty$,

$$\sup_{r>0} \left| \frac{\#\{i \le n : \|\mathbf{X}_i\| \le r\}}{n} - \mathbf{P}(\|\mathbf{X}_1\| \le r) \right| \to 0. \tag{4.11}$$

Given $h \in \mathbb{N}$, since $\Omega^{d-1}$ is compact, there exist $\mathbf{v}_1^h, \dots, \mathbf{v}_{J_h}^h \in \Omega^{d-1}$ such that for every $\mathbf{v} \in \Omega^{d-1}$ there exists $i_{\mathbf{v}} \in \{1, \dots, J_h\}$ such that $\|\mathbf{v} - \mathbf{v}_{i_h}^h\| \le h^{-1}$. From Lemma 4.4.4, we have that there exists $A_h \in \mathcal{A}$ such that $\mathbf{P}(A_h) = 1$ and for every $\omega \in A_h$, (4.11) is satisfied and for every $\gamma \in (0, 1/2)$,

$$\begin{aligned} &\sup_{\alpha\in(\gamma,1-\gamma)} \left(\max_{i\le J_h}\left\{\left|\underline{q}_\alpha(R_{\mathbf{v}_i^h}^{n,\omega}) - q_\alpha(\mathbf{Q}_{\mathbf{v}_i^h})\right|, \left|\bar{q}_\alpha(R_{\mathbf{v}_i^h}^{n,\omega}) - q_\alpha(\mathbf{Q}_{\mathbf{v}_i^h})\right|\right\}\right) \to 0, \\ &\sup_{\alpha\in(\gamma,1-\gamma)} \left(\max_{i\le J_h}\left\{\left|\underline{q}_\alpha(T_{\mathbf{v}_i^h}^{n,\omega}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}_i^h})\right|, \left|\bar{q}_\alpha(T_{\mathbf{v}_i^h}^{n,\omega}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}_i^h})\right|\right\}\right) \to 0. \end{aligned} \tag{4.12}$$

Denote $A_0 = \cap_{h\in\mathbb{N}} A_h$. Obviously $A_0 \in \mathcal{A}$ and $\mathbf{P}(A_0) = 1$. Let $\omega \in A_0$ be a point which will remain fixed along the proof. We begin proving the first statement in (4.10). Let $\varepsilon > 0$. Let $\lambda_d$ be the largest eigenvalue of $\Sigma$. Given $\mathbf{v} \in \Omega^{d-1}$ and $\gamma \in (0, 1/2)$, we have that

$$\begin{aligned} q_{\frac{1}{2}+\gamma}(\mathbf{Q}_{\mathbf{v}}) - q_{\frac{1}{2}-\gamma}(\mathbf{Q}_{\mathbf{v}}) &= (\mathbf{v}'\Sigma\mathbf{v})\left(q_{\frac{1}{2}+\gamma}(N_1(0,1)) - q_{\frac{1}{2}-\gamma}(N_1(0,1))\right) \\ &\le \lambda_d\left(q_{\frac{1}{2}+\gamma}(N_1(0,1)) - q_{\frac{1}{2}-\gamma}(N_1(0,1))\right). \end{aligned}$$

Therefore, there exists $\gamma_1 \in (0, 1/2)$ such that

$$\sup_{\mathbf{v}\in\Omega^{d-1}} \left(q_{\frac{1}{2}+\gamma_1}(\mathbf{Q}_{\mathbf{v}}) - q_{\frac{1}{2}-\gamma_1}(\mathbf{Q}_{\mathbf{v}})\right) < \frac{\varepsilon}{3}. \tag{4.13}$$

Analogously, we can prove that there exits $\gamma_2 \in (0, 1/2)$ such that,

$$\sup_{\mathbf{v}\in\Omega^{d-1}} \left( q_{\frac{1}{2}+\gamma_2}(\bar{\mathbf{Q}}_{\mathbf{v}}) - q_{\frac{1}{2}-\gamma_2}(\bar{\mathbf{Q}}_{\mathbf{v}}) \right) < \frac{\varepsilon}{3}. \tag{4.14}$$

Take $\gamma = \inf\{\gamma_1, \gamma_2, \varepsilon\}$. Let $r > 0$ be such that $\mathbf{P}(\|\mathbf{X}_1\| \leq r) > 1 - \gamma$ and $h \in \mathbb{N}$ such that $r/h < \varepsilon/3$ and $2d\lambda_d M_1/h < \varepsilon/3$, where $M_1$ is the MADN of a $N_1(0,1)$.

By (4.11) and (4.12), there exists $N^\omega$ such that if $n \geq N^\omega$, then $\#\{i \leq n : \|\mathbf{X}_i(\omega)\| < r\} > n(1-\gamma)$ and

$$\begin{aligned}
\sup_{\alpha\in\left(\frac{1}{2}-\gamma,\frac{1}{2}+\gamma\right)} \left( \max_{i\leq J_h} \left\{ \left| \underline{q}_\alpha(R^{n,\omega}_{\mathbf{V}^h_i}) - q_\alpha(\mathbf{Q}_{\mathbf{v}^h_i}) \right|, \left| \bar{q}_\alpha(R^{n,\omega}_{\mathbf{V}^h_i}) - q_\alpha(\mathbf{Q}_{\mathbf{v}^h_i}) \right| \right\} \right) &< \frac{\varepsilon}{3} \\
\sup_{\alpha\in\left(\frac{1}{2}-\gamma,\frac{1}{2}+\gamma\right)} \left( \max_{i\leq J_h} \left\{ \left| \underline{q}_\alpha(T^{n,\omega}_{\mathbf{v}^h_i}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}^h_i}) \right|, \left| \bar{q}_\alpha(T^{n,\omega}_{\mathbf{v}^h_i}) - q_\alpha(\bar{\mathbf{Q}}_{\mathbf{v}^h_i}) \right| \right\} \right) &< \frac{\varepsilon}{3}.
\end{aligned} \tag{4.15}$$

Let $\mathbf{v} \in \Omega^{d-1}$, if $\|\mathbf{X}_j(\omega)\| \leq r$,

$$\left|(\mathbf{X}_j(\omega))'\mathbf{v} - (\mathbf{X}_j(\omega))'\mathbf{v}^h_{i_\mathbf{v}}\right| \leq \|\mathbf{X}_j(\omega)\| \|\mathbf{v} - \mathbf{v}^h_{i_\mathbf{v}}\| \leq rh^{-1} < \frac{\varepsilon}{3}, \tag{4.16}$$

and therefore, by Corollary 4.4.3 with $\alpha = 1/2$, we have that

$$\hat{m}^\omega_\mathbf{v} \in \left[ \underline{q}_{\frac{1}{2}-\gamma}\left(R^{n,\omega}_{\mathbf{V}^h_{i_\mathbf{V}}}\right) - \frac{\varepsilon}{3}, \bar{q}_{\frac{1}{2}+\gamma}\left(R^{n,\omega}_{\mathbf{V}^h_{i_\mathbf{V}}}\right) + \frac{\varepsilon}{3} \right],$$

and (4.15) gives

$$\hat{m}^\omega_\mathbf{v} \in \left[ q_{\frac{1}{2}-\gamma}\left(\mathbf{Q}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) - \frac{2\varepsilon}{3}, q_{\frac{1}{2}+\gamma}\left(\mathbf{Q}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) + \frac{2\varepsilon}{3} \right]. \tag{4.17}$$

On the other hand, we have that

$$|\hat{m}^\omega_\mathbf{v} - m_\mathbf{v}| \leq |\hat{m}^\omega_\mathbf{v} - m_{\mathbf{v}^h_{i_\mathbf{v}}}| + |m_{\mathbf{v}^h_{i_\mathbf{v}}} - m_\mathbf{v}|. \tag{4.18}$$

Moreover, $\mathbf{m}_\mathbf{v} = 0$ for every $\mathbf{v}$ because all probabilities $\mathbf{Q}_\mathbf{v}$ are normal with mean zero. Thus, the second addend in (4.18) is null. However, (4.17) and (4.13) imply

$$|\hat{m}^\omega_\mathbf{v} - m_{\mathbf{v}^h_{i_\mathbf{v}}}| < \varepsilon. \tag{4.19}$$

Then, the first item in (4.10) is proved because by (4.18) and (4.19) we have that, if $n > N^\omega$, then

$$\sup_\mathbf{v} |\hat{m}^\omega_\mathbf{v} - m_\mathbf{v}| < \varepsilon. \tag{4.20}$$

For the second item in (4.10), notice that if $\mathbf{v} \in \Omega^{d-1}$ and $h \in \mathbb{N}$, then

$$|\hat{M}^{\omega}_{\mathbf{v}} - M_{\mathbf{v}}| \leq |\hat{M}^{\omega}_{\mathbf{v}} - \hat{M}^{\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}}| + |\hat{M}^{\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}} - M_{\mathbf{v}^h_{i_{\mathbf{v}}}}| + |M_{\mathbf{v}^h_{i_{\mathbf{v}}}} - M_{\mathbf{v}}|. \tag{4.21}$$

If $n \geq N^{\omega}$, $i = 1, \ldots, n$, and $\mathbf{v} \in \Omega^{d-1}$, from (4.20) (remember that $\mathbf{m} = \mathbf{0}$) we have

$$\left| \left|(\mathbf{X}_i(\omega))' \, \mathbf{v} - \hat{m}_{\mathbf{v}}\right| - \left|(\mathbf{X}_i(\omega))' \, \mathbf{v}\right| \right| \leq |\hat{m}_{\mathbf{v}}| < \varepsilon. \tag{4.22}$$

Therefore, we can apply Corollary 4.4.3 with $\alpha = 1/2$, $\delta = \varepsilon$ and $\gamma = 0$ to obtain that

$$\begin{aligned}\left[\underline{M}_{\frac{1}{2}}\left(R^{n,\omega}_{\mathbf{v}}\right), \bar{M}_{\frac{1}{2}}\left(R^{n,\omega}_{\mathbf{v}}\right)\right] &= \left[\underline{q}_{\frac{1}{2}}\left(\hat{S}^{n,\omega}_{\mathbf{v}}\right), \bar{q}_{\frac{1}{2}}\left(\hat{S}^{n,\omega}_{\mathbf{v}}\right)\right] \\ &\subset \left[\underline{q}_{\frac{1}{2}}\left(T^{n,\omega}_{\mathbf{v}}\right) - 2\varepsilon, \bar{q}_{\frac{1}{2}}\left(T^{n,\omega}_{\mathbf{v}}\right) + 2\varepsilon\right],\end{aligned}$$

which joined to (4.15) and the fact that $M_{\mathbf{v}} = m(\bar{Q}_{\mathbf{v}})$ gives that if $n \geq N^{\omega}$,

$$\left|\hat{M}^{\omega}_{\mathbf{V}^h_{i_{\mathbf{V}}}} - M_{\mathbf{V}^h_{i_{\mathbf{V}}}}\right| < 2\varepsilon + \frac{\varepsilon}{3}.$$

Concerning the third addend in (4.21), notice that $M_{\mathbf{v}} = m(\bar{\mathbf{Q}}_{\mathbf{v}})$ coincides with $\mathbf{v}'\Sigma\mathbf{v}M_1$. Thus, if we write $\mathbf{v} = (v^1, \ldots, v^d)'$, then

$$\begin{aligned}|M_{\mathbf{v}^h_{i_{\mathbf{v}}}} - M_{\mathbf{v}}| &= \left|\mathbf{v}'\Sigma\mathbf{v} - (\mathbf{v}^h_{i_{\mathbf{v}}})'\Sigma\mathbf{v}^h_{i_{\mathbf{v}}}\right| M_1 \\ &= \left|\sum_{j=1}^{d} \left(v^j\right)^2 \lambda_j - \sum_{j=1}^{d} \left((\mathbf{v}^h_{i_{\mathbf{v}}})^j\right)^2 \lambda_j\right| M_1 \\ &\leq \lambda_d M_1 \sum_{j=1}^{d} \left|\left(v^j\right)^2 - \left((\mathbf{v}^h_{i_{\mathbf{v}}})^j\right)^2\right| \\ &\leq 2\lambda_d M_1 \sum_{j=1}^{d} \left|v^j - (\mathbf{v}^h_{i_{\mathbf{v}}})^j\right| \\ &\leq 2d\lambda_d M_1 \left\|\mathbf{v} - \mathbf{v}^h_{i_j}\right\| \leq \frac{2d\lambda_d M_1}{h} < \frac{\varepsilon}{3}.\end{aligned}$$

Now, let us pay attention to the first addend in (4.21). According to (4.16) and (4.9) in Corollary 4.4.3, we have that

$$|\hat{M}_{\mathbf{v}} - \hat{M}_{\mathbf{v}^h_{i_{\mathbf{v}}}}| \leq \bar{M}_{\frac{1}{2}+\alpha}\left(R^{n,\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}}\right) - \underline{M}_{\frac{1}{2}-\alpha}\left(R^{n,\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}}\right) + \frac{2\varepsilon}{3} + \delta^*_{\gamma}. \tag{4.23}$$

First, (4.15) and (4.13) give that

$$\delta^*_{\gamma} \leq \bar{q}_{\frac{1}{2}+\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}}\right) - \underline{q}_{\frac{1}{2}-\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_{\mathbf{v}}}}\right) \leq q_{\frac{1}{2}+\gamma}\left(\mathbf{Q}_{\mathbf{v}^h_{i_k}}\right) - q_{\frac{1}{2}-\gamma}\left(\mathbf{Q}_{\mathbf{v}^h_{i_k}}\right) + \frac{2\varepsilon}{3} < \varepsilon.$$

For the first addend in (4.23), by Corollary 4.4.3 with $\gamma = 0$, we conclude that

$$\begin{aligned}\bar{M}_{\frac{1}{2}+\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) - \underline{M}_{\frac{1}{2}-\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) &= \bar{q}_{\frac{1}{2}+\gamma}\left(S^{n,\omega}_{i^h_\mathbf{v}}\right) - \underline{q}_{\frac{1}{2}+\gamma}\left(S^{n,\omega}_{i^h_\mathbf{v}}\right)\\ &\leq \bar{q}_{\frac{1}{2}+\gamma}(T^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}) - \underline{q}_{\frac{1}{2}-\gamma}(T^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}) + 2\left|\hat{m}_{\mathbf{v}^h_{i_\mathbf{v}}}\right|,\end{aligned}$$

and from (4.15), (4.14) and (4.20) we have that

$$\bar{M}_{\frac{1}{2}+\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) - \underline{M}_{\frac{1}{2}-\gamma}\left(R^{n,\omega}_{\mathbf{v}^h_{i_\mathbf{v}}}\right) < 3\varepsilon.$$

And the proof ends because (4.23) and the previous inequalities give that if $\omega \in A_0$ and $n \geq N^\omega$, then

$$\left|\hat{M}_\mathbf{v} - \hat{M}\mathbf{v}^h_{i_\mathbf{v}}\right| < 6\varepsilon. \qquad \square$$

**Theorem 4.4.6.** *Assume* (A1.e) *and* (A2.e)*. Suppose that $a, b$ and $t$ are strictly positive constants such that $a \leq b$ and let $\tilde{K}_n$ be as defined above. Then, as $n \to \infty$, a.s.*

$$\mathbf{P}\left(|\tilde{Y}_n^{\tilde{K}_n}| > b|\ \|\mathbf{X}\|_\Sigma = t\right) \to \int_{\Omega^{d-1}_\Sigma(t)} g^b_a(\mathbf{x}, \mathbf{0}, \Sigma) f_t(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

*where $g^b_a(\cdot,\cdot,\cdot)$ was defined in Proposition 4.2.5.*

*Proof.* Let $\omega \in \Upsilon$ and denote

$$g^{n,\omega}_{a,b}(\mathbf{x}) := \frac{\mathbf{P}\left(\mathbf{v} : \frac{\left|\mathbf{x}'\mathbf{v} - \hat{m}^{n,\omega}_\mathbf{v}\right|}{\hat{M}^{n,\omega}_\mathbf{v}} > b\right)}{\mathbf{P}\left(\mathbf{v} : \frac{\left|\mathbf{x}'\mathbf{v} - \hat{m}^{n,\omega}_\mathbf{v}\right|}{\hat{M}^{n,\omega}_\mathbf{v}} > b\right) + \mathbf{P}\left(\mathbf{v} : \frac{\left|\mathbf{x}'\mathbf{v} - \hat{m}^{n,\omega}_\mathbf{v}\right|}{\hat{M}^{n,\omega}_\mathbf{v}} < a\right)}.$$

Notice that the probabilities involved in this expression are conditioned given the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$. It is clear that if we integrate on the samples

$$\mathbf{P}\left(|\tilde{Y}_n^{\tilde{K}_n}| > b\,||\mathbf{X}\|_\Sigma = t\right) = \int_{\Omega^{d-1}_\Sigma(t)} \left(\int_\Upsilon g^{n,\omega}_{a,b}(\mathbf{x})\,\mathrm{d}\mathbf{P}(\omega)\right) f_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Let us prove that the map $g^{n,\omega}_{a,b}$ is well defined. As in Theorem 4.4.5, we denote by $M_1$ the MADN of the $N_1(0,1)$. If $\mathbf{v} \in \Omega^{d-1}$, then $M_\mathbf{v} = (\mathbf{v}'\Sigma\mathbf{v})^{1/2} M_1 \geq \lambda_1 M_1$.

According to Theorem 4.4.5, there exists a set $A_0 \in \mathcal{A}$ with $\mathbf{P}(A_0) = 1$ such that for every $\omega \in A_0$, there exits $N^\omega$ such that if $n \geq N^\omega$ then for every $\mathbf{v} \in \Omega^{d-1}$, $\hat{M}^{n,\omega}_\mathbf{v} > \lambda_1 M_1/2$ and $|\hat{m}^{n,\omega}_\mathbf{v}| < a\lambda_1 M_1/4$. Then, if $\mathbf{x} \in \Omega^{d-1}_\Sigma(t)$

$$
\begin{aligned}
\mathbf{P}\left(\mathbf{v} : \tfrac{|\mathbf{x}'\mathbf{v} - \hat{m}^{n,\omega}_{\mathbf{v}}|}{\hat{M}^{n,\omega}_{\mathbf{v}}} < a\right) &= \mathbf{P}\left(\mathbf{v} : \mathbf{x}'\mathbf{v} \in \left(\hat{m}^{n,\omega}_{\mathbf{v}} - a\hat{M}^{n,\omega}_{\mathbf{v}}, \hat{m}^{n,\omega}_{\mathbf{v}} + a\hat{M}^{n,\omega}_{\mathbf{v}}\right)\right) \\
&\geq \mathbf{P}\left(\mathbf{v} : \mathbf{x}'\mathbf{v} \in \left(\hat{m}^{n,\omega}_{\mathbf{v}} - \frac{a\lambda_1 M_1}{2}, \hat{m}^{n,\omega}_{\mathbf{v}} + \frac{a\lambda_1 M_1}{2}\right)\right) \\
&\geq \mathbf{P}\left(\mathbf{v} : \mathbf{x}'\mathbf{v} \in \left(-\frac{a\lambda_1 M_1}{4}, \frac{a\lambda_1 M_1}{4}\right)\right) > 0,
\end{aligned}
$$

where the last inequality follows from the fact that $\{\mathbf{v} : |\mathbf{x}'\mathbf{v}| < a\lambda_1 M_1/4\} \neq \varnothing$.

Additionally, for every $\omega \in A_0$, $\mathbf{v} \in \Omega^{d-1}$ and $\mathbf{x} \in \Omega^{d-1}_{\Sigma}(t)$

$$
1_{\left\{\frac{|\mathbf{x}'\mathbf{v} - \hat{m}^{n,\omega}_{\mathbf{v}}|}{\hat{M}^{n,\omega}_{\mathbf{v}}} > b\right\}} \to 1_{\left\{\frac{|\mathbf{x}'\mathbf{v}|}{M_{\mathbf{v}}} > b\right\}} \tag{4.24}
$$

unless $\mathbf{v}$ satisfies that $|\mathbf{x}'\mathbf{v}| = bM_{\mathbf{v}}$, but this equality only happens for $\mathbf{v}$ in a set (depending on $\mathbf{x}$) with Lebesgue measure equal to zero. Consequently, for every $\omega \in A_0$ and $\mathbf{x} \in \Omega^{d-1}_{\Sigma}(t)$, the convergence in (4.24) holds for almost every $\mathbf{v} \in \Omega^{d-1}$. Since the involved functions are bounded, (4.24) gives that, for every $\omega \in A_0$,

$$
g^{n,\omega}_{a,b}(\mathbf{x}) \to \tilde{g}_{a,b}(\mathbf{x}).
$$

The fact that $0 \leq g^{n,\omega}_{a,b}(\mathbf{x}) \leq 1$ for every $\mathbf{x}$, allows to apply the dominated convergence theorem and the result is proven. □

Next proposition gives the asymptotic behaviour of the first two moments of $\tilde{K}_n$. Its proof is similar to that one of Theorem 4.4.6 and we do not include it.

**Proposition 4.4.7.** *Under assumptions* (A1.e) *and* (A2.e)*, assume that* $a, b, t$ *are strictly positive constants such that* $a \leq b$*. Then, as* $n \to \infty$*, a.s.*

$$
\begin{aligned}
E\left(\tilde{K}_n \mid \|\mathbf{X}\|_{\Sigma} = t\right) &\to \int_{\Omega^{d-1}_{\Sigma}(t)} \bar{g}^{b}_{a}\left(\mathbf{x}, \mathbf{0}, \Sigma\right) f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}, \\
\mathrm{Var}(\tilde{K}_n \mid \|\mathbf{X}\|_{\Sigma} = t) &\to \int_{\Omega^{d-1}_{\Sigma}(t)} \bar{g}^{b}_{a}\left(\mathbf{x}, \mathbf{0}, \Sigma\right) \left(2\bar{g}^{b}_{a}\left(\mathbf{x}, \mathbf{0}, \Sigma\right) - 1\right) f_t(\mathbf{x})\, \mathrm{d}\mathbf{x} \\
&\quad - \left(\int_{\Omega^{d-1}_{\Sigma}(t)} \bar{g}^{b}_{a}\left(\mathbf{x}, \mathbf{0}, \Sigma\right) f_t(\mathbf{x})\, \mathrm{d}\mathbf{x}\right)^2,
\end{aligned}
$$

*where* $\bar{g}^{b}_{a}(\cdot,\cdot,\cdot)$ *was defined in Proposition 4.3.1.*

The expressions of Theorem 4.4.6 and Proposition 4.4.7 get simplified in the case $\Sigma = I_d$ as shown in the following corollary.

**Corollary 4.4.8.** *With the assumptions and the notation in Theorem 4.4.6, consider $F(\cdot, t)$ defined as in (3.4). If $\Sigma = I_d$, then, as $n \to \infty$, a.s.,*

$$\begin{aligned}
\mathbf{P}\left(\left|\tilde{Y}_n^{\tilde{K}_n}\right| > b \,|\, \|\mathbf{X}\| = t\right) &\to \frac{1 - F(b,t)}{1 - F(b,t) + F(a,t)},\\
E(\tilde{K}_n \,|\, \|\mathbf{X}\| = t) &\to \frac{1}{1 - F(b,t) + F(a,t)},\\
\mathrm{Var}(\tilde{K}_n \,|\, \|\mathbf{X}\| = t) &\to \frac{F(b,t) - F(a,t))}{(1 - F(b,t) + F(a,t))^2}.
\end{aligned}$$

**Remark 4.4.9.** *Denote $\mathbb{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. From the proofs of Theorem 4.4.6 and Proposition 4.4.7 it is clear that a.s.*

$$\begin{aligned}
\mathbf{P}\left(|\tilde{Y}_n^{\tilde{K}_n}| > b \,|\, \|\mathbf{X}\|_\Sigma = t, \mathbb{X}_n\right) &\to \int_{\Omega_\Sigma^{d-1}(t)} g_a^b(\mathbf{x}, \mathbf{0}, \Sigma) f_t(\mathbf{x})\,\mathrm{d}\mathbf{x},\\
E\left(\tilde{K}_n \,|\, \|\mathbf{X}\|_\Sigma = t, \mathbb{X}_n\right) &\to \int_{\Omega_\Sigma^{d-1}(t)} \bar{g}_a^b\,(\mathbf{x}, \mathbf{0}, \Sigma)\, f_t(\mathbf{x})\,\mathrm{d}\mathbf{x},\\
\mathrm{Var}(\tilde{K}_n \,|\, \|\mathbf{X}\|_\Sigma = t, \mathbb{X}_n) &\to \int_{\Omega_\Sigma^{d-1}(t)} \bar{g}_a^b\,(\mathbf{x}, \mathbf{0}, \Sigma)\left(2\bar{g}_a^b\,(\mathbf{x}, \mathbf{0}, \Sigma) - 1\right) f_t(\mathbf{x})\,\mathrm{d}\mathbf{x}\\
&\quad - \left(\int_{\Omega_\Sigma^{d-1}(t)} \bar{g}_a^b\,(\mathbf{x}, \mathbf{0}, \Sigma)\, f_t(\mathbf{x})\,\mathrm{d}\mathbf{x}\right)^2.
\end{aligned}$$

## 4.5 Computation of the constants $a$ and $b$

The explicit computation of $a$ and $b$ requires to find a solution of (4.6) satisfying that $\mathrm{E}\left(\tilde{K}_n | \|\mathbf{X}\|_\Sigma = C_n^d\right)$ equals a pre-specified value. This problem has been impossible for us to tackle even in the non-robust version (4.5) which handles the function $F_{n,\Sigma}(a, b, t)$.

Proposition 4.2.5 gives an explicit expression for $F_{n,\Sigma}(a, b, t)$; the problem being that the integrand in this expression is so involved that, with the exception of $a = b$, we have not been able to compute the integral even when $\Sigma = I_d$ (note that $a$ and $b$ depend on the covariance matrix). In addition, the complexity increases when $\Sigma \neq I_d$, because of the dependency of the projections given $\|\mathbf{X}\|_\Sigma$ as Proposition 4.2.7 showed.

Analogously to Chapter 4, a possibility to solve (4.5) as an approximation to (4.6) would be to take $a = b = a_\alpha$, the conditional $(1 - \alpha)$-quantile of $Y_n^{\mathbf{V}}$ given that $\|\mathbf{X}\|_\Sigma = C_n^d$; but again this does not look very sensible because this means taking the decision based on one single random projection. Similarly to Section 3.5, we have now Proposition 4.5.1 which helps us to compute $a$ and $b$ in some situations. We do not include the proof of such proposition because it is analogous to that in Proposition 3.5.1.

**Proposition 4.5.1.** *Let* $\mathbf{x} \in \mathbb{R}^d$ *such that* $0 < t = \|\mathbf{X}\|_\Sigma$*. Given* $a > 0$ *such that* $\mathbf{P}(|\tilde{y}_n^{\mathbf{V}}| < a) \leq \alpha$*, there exists* $b_a^t$ *such that* $F_{n,\Sigma}(a, b_a^t, C_n^d) = \alpha$*. Moreover, for every* $t > C_n^d$ *the map* $a \mapsto b_a^t$ *is strictly decreasing on* $a$*.*

Similarly to Proposition 3.5.2, Proposition 4.5.2 somehow eases the computation of $a$ and $b$ because it states that if we use the constants of the identity for a general covariance matrix, then we will make the decision using the pre-specified number of projections or more. We include no proof of such proposition because it is similar to that of Proposition 3.5.2.

**Proposition 4.5.2.** *Let us assume* (A1.e) *and* (A2.e) *and let* $t > 0$ *and* $0 < a \leq b$*. Let* $\Sigma \neq I_d$ *be a positive definite matrix. Let* $\mathbb{X}^\Sigma := \{\mathbf{X}_n^\Sigma\}$ *and* $\mathbb{X}^{I_d} := \{\mathbf{X}_n^{I_d}\}$ *be two random samples taken from the* $N_d(\mathbf{0}, \Sigma)$ *and* $N_d(\mathbf{0}, I_d)$ *respectively. Then, almost surely,*

$$\lim_n E\left(\tilde{K}_n^{a,b} \,\middle|\, \|\mathbf{X}\|_\Sigma = t, \mathbb{X}^\Sigma\right) > \lim_n E\left(\tilde{K}_n^{a,b} \,\middle|\, \|\mathbf{X}\| = t, \mathbb{X}^{I_d}\right).$$

Corollary 4.5.3 is directly deduced from Proposition 4.5.2 and Corollary 4.4.8.

**Corollary 4.5.3.** *With the assumptions and notation in Proposition 4.5.2, then a.s.*

$$\lim_{n\to\infty} E\left(\tilde{K}_n^{a,b} \,||\mathbf{X}\|_\Sigma = t, \mathbb{X}_n^\Sigma\right) > \frac{1}{1 - F(b,t) + F(a,t)}.$$

After Proposition 4.5.2, our proposal consists of using for any sample, the constants $a$ and $b$ computed for samples taken from $N_d(\mathbf{0}, I_d)$. However, Proposition 4.5.2 leaves two open points: the level of the test obtained when using constants $a$ and $b$ computed for the identity with $\Sigma \neq I_d$; and some hints on the expected number of observations when $n$ is low, mostly, when $\Sigma \neq I_d$. We have obtained no theoretical result fixing those points, but we have produced practical evidence suggesting that the situation is reasonably good. Specifically, we have used the same covariance matrices as in Subsection 3.5.2 and we have conducted numerical experiments using pairs $(a, b)$ computed for the identity with the following results:

1) The obtained rejection rates with $\Sigma \neq I_d$ are close to the rates of the identity.

2) The results obtained for sample sizes as low as $n = 50$ are similar to those predicted by Proposition 4.5.2. I.e., for sample sizes $n \geq 50$ and covariance matrices $\Sigma \neq I_d$, the mean of the obtained values for $\tilde{K}_n$ are mostly larger than the expected for the identity and they are seldom only slightly lower.

3) The mean of the values obtained for $\tilde{K}_n$ when $\Sigma \neq I_d$ are generally similar to those obtained when $\Sigma = I_d$ but sometimes they are much higher.

Detailed information on those points as well as some details on the computation of $a$ and $b$ when $\Sigma = I_d$ are included in Subsections 4.5.1 and 4.5.2.

### 4.5.1 Computation of $(a, b)$ when $\Sigma = I_d$

In this subsection, given $n \in \mathbb{N}$, we want to compute the constants $a$ and $b$ giving a power $\alpha$-test with a given value $l \geq 1$ for $\mathrm{E}(\tilde{K}_n \,|\|\mathbf{X}\| = C_n^d)$. For this, analogously to Chapter 3, we could solve the system (3.18) and then look for $a$ and $b$ satisfying that $u = \tilde{F}(a, C_n^d)$ and $v = \tilde{F}(b, C_n^d)$, where $\tilde{F}(y,t) = \mathbf{P}\left(|\tilde{Y}_n^{\mathbf{V}}| < y \,|\|\mathbf{X}\| = t\right)$.

The solution of that system in this case is $u = (1-\alpha)/k$ and $v = 1-\alpha/k$. Obviously, if $k > 1$, then $v > u$ and only remains to find the $u$ and $v$ quantiles of the distribution $\tilde{F}(\cdot, C_n^d)$. Since we have no explicit expressions for them and their numerical computation is not feasible, we have decided to begin computing $a$ and $b$ by the Monte Carlo method. The computation is done as in Algorithm 3 taking $\Sigma = I_d$ and with the following modifications: Step *1.3*). Compute $\tilde{Y}^j = \left|(\mathbf{X}^j)'\mathbf{V}^j - \hat{m}_{\mathbf{V}^j}\right|/\hat{M}_{\mathbf{V}^j}$; Step 2). Take $a$ and $b$ equal to the quantiles $u$ and $v$ of the sample $\tilde{Y}^1, \ldots, \tilde{Y}^N$. By the same reasons adduced in Subsection 3.5.2, we need to employ the bisection method to achieve the desired level of the test.

Table 4.1 shows the values of the constants $a$ and $b$ for different values of $l$ (see Table A.5 in the Appendix for the non-robust version of them). Those values have been computed with the above explained methodology with $N = 10^6$.

From this table, the bigger $l$, the wider the interval $(a, b)$ according to Corollary 4.3.2. However, the larger the sample size, the narrower the interval $(a, b)$. This is due to the fact that the estimation of the parameters is more stable for greater sample sizes. Last fact is in contrast with Table 3.2 where we had that, if $\Sigma$ is known, the larger $n$, the wider the interval $(a, b)$.

Table 4.1.: Obtained values of $(a, b)$ when $\Sigma = I_d$, $C_n^d \equiv C_n^d(0.05)$ and different values of $n, d$ and $l = \mathrm{E}(\tilde{K}_n|\|X\| = C_n^d(\delta))$.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $l = 50$ | | $l = 100$ | | $l = 50$ | | $l = 100$ | | $l = 50$ | | $l = 100$ | |
| $d$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ |
| 5 | 0.0587 | 6.0580 | 0.0289 | 6.3681 | 0.05901 | 5.4914 | 0.0299 | 5.6691 | 0.0645 | 5.1718 | 0.0322 | 5.2655 |
| 50 | 0.0325 | 4.9714 | 0.0163 | 5.3212 | 0.0326 | 4.6374 | 0.0163 | 4.9143 | 0.0336 | 4.4525 | 0.0167 | 4.6989 |
| 100 | 0.0303 | 4.7184 | 0.0150 | 5.0936 | 0.0303 | 4.3539 | 0.0151 | 4.6495 | 0.0304 | 4.1478 | 0.0156 | 4.3910 |
| 500 | 0.0268 | 4.3039 | 0.0133 | 4.6239 | 0.0267 | 3.9230 | 0.0133 | 4.2078 | 0.0266 | 3.7278 | 0.0132 | 3.9520 |
| 1000 | 0.0263 | 4.1916 | 0.0130 | 4.5217 | 0.0261 | 3.8253 | 0.0128 | 4.0909 | 0.0259 | 3.6096 | 0.0130 | 3.8197 |

### 4.5.2 Computation of $(a, b)$ when $\Sigma \neq I_d$

As stated, based on Proposition 4.5.2, our idea is to use the values obtained for $\Sigma = I_d$ to handle any covariance matrix. We firstly check if those values are suitable for general matrices. For this, we will use the matrices $\Sigma_1, \ldots, \Sigma_4$ defined in Subsection 3.5.2 in Chapter 3.

The procedure goes as follows, for each combination of dimension and sample size, we have computed a pair $(a_I, b_I)$ giving an $\alpha$-level test for the identity matrix as explained in Subsection 4.5.1. We have kept $a_I$ and, for every $\Sigma = \Sigma_i^d, i = 1, \ldots, 4$, we have computed (using the same procedure as in Subsection 4.5.1 with $N = 10^4$ simulations) the value $b_\Sigma$ such that the pair $(a_I, b_\Sigma)$ is an $\alpha$-level test.

As in Subsection 3.5.2, the values $b_I$ and $b_\Sigma$ are very resembling in the considered cases (see Table 4.2 which shows the $b_\Sigma$ maximizing the difference $|b_I - b_\Sigma|$). The same happens with the quantities $l_I^r := \mathrm{E}(\tilde{K}_n | \|X\| = rC_n^d(0.05))$ and $l_\Sigma^r := \mathrm{E}(\tilde{K}_n | \|X\|_\Sigma = rC_n^d(0.05))$ when $r = 1$ except for $\Sigma = \Sigma_3^d$ (see Table A.6 in the Appendix, and Table A.7 for their non-robust versions).

Table 4.2.: Values of $b_\Sigma$ giving the greatest difference $|b_I - b_\Sigma|$ for $\Sigma = \Sigma_i^d, i = 1, \ldots, 4$, for several values of $d$, $n$ and $\mathrm{E}(\tilde{K}_n \; |\|\mathbf{X}\|_\Sigma = C_n^d)$. $a$'s are taken from Table 4.1. Column $\Sigma$ tells the matrix in which $b_\Sigma$ was obtained.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $l_I^1$=50 | | $l_I^1$=100 | | $l_I^1$=50 | | $l_I^1$=100 | | $l_I^1$=50 | | $l_I^1$=100 | |
| $d$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ | $b_\Sigma$ | $\Sigma$ |
| 5 | 5.0184 | $\Sigma_1^d$ | 5.1241 | $\Sigma_1^d$ | 4.9185 | $\Sigma_1^d$ | 4.9781 | $\Sigma_1^d$ | 5.0139 | $\Sigma_1^d$ | 5.0735 | $\Sigma_1^d$ |
| 50 | 5.1413 | $\Sigma_3^d$ | 5.4932 | $\Sigma_3^d$ | 4.6194 | $\Sigma_4^d$ | 4.9504 | $\Sigma_3^d$ | 4.4439 | $\Sigma_3^d$ | 4.6858 | $\Sigma_3^d$ |
| 100 | 4.8813 | $\Sigma_3^d$ | 5.1857 | $\Sigma_3^d$ | 4.3497 | $\Sigma_4^d$ | 4.6387 | $\Sigma_2^d$ | 4.1399 | $\Sigma_3^d$ | 4.3691 | $\Sigma_3^d$ |
| 500 | 4.3244 | $\Sigma_2^d$ | 4.6946 | $\Sigma_3^d$ | 4.0248 | $\Sigma_3^d$ | 4.2460 | $\Sigma_3^d$ | 3.7509 | $\Sigma_4^d$ | 3.9143 | $\Sigma_3^d$ |
| 1000 | 4.3129 | $\Sigma_3^d$ | 4.6166 | $\Sigma_3^d$ | 3.9221 | $\Sigma_3^d$ | 4.1094 | $\Sigma_3^d$ | 3.6276 | $\Sigma_1^d$ | 3.8363 | $\Sigma_2^d$ |

## 4.6 Practical implementation

In this section we provide advice on the practical implementation of the method. We pay attention to how to fix the number of expected projections (Subsection 4.6.1) and how many simulated values of $\tilde{Y}_n^{\mathbf{V}}$ we should produce to compute $a$ and $b$ (Subsection 4.6.2). We also include an algorithm to analyse all points in a sample (Algorithm 4 in Subsection 4.6.3) and we finalize with a procedure to reduce the role of the randomness in the process (Subsection 4.6.4)

### 4.6.1 Which value should we choose for $l^r_{\Sigma}$?

In principle, the higher the $I^r_{\Sigma}$ the higher the power under the alternative. However, the computational effort increases. The simulations we present below show a detectable increment in power from $I^r_{\Sigma} = 50$ to $I^r_{\Sigma} = 100$. However, this increment is not too striking and, of course, the improvement slows down for values of $I^r_{\Sigma}$ above 100.

Hence, our advise is to fix this parameter at 50, or, at most at 100. In fact, in the simulations in Subsection 4.7.2 we have used $I^r_{\Sigma} = 50$, while we have chosen $I^r_{\Sigma} = 100$ in Subsection 4.7.3.

### 4.6.2 How many simulated values of $\tilde{Y}^{\mathbf{V}}$ are required to obtain $a$ and $b$?

In Algorithm 3 with the modifications mentioned in Subsection 4.5.1, a large value $N$ of replicas of $\tilde{Y}^{\mathbf{V}}_n$ is required. As stated in Subsection 4.5.1, in this work we have chosen $N = 10^6$, but this is quite time consuming. Some computations suggest that $N = 10^4$ could do it depending on the involved percentiles, but it seems that $N = 10^5$ offers a reasonable trade-off between time and precision. Table 4.3 shows the required computational times for some values of dimension, sample size and $l^1_I$. Those times range from 40 seconds to 33 minutes in a four cores processor 3.2 GHz Intel Core i5. The decrement observed in the case $d = 50$, $n = 100$, $l^1_I = 100$ is due to the fact that those cases required a very sort bisection step.

Table 4.3.: Computation times (in seconds) of $a$, $b$ with $N = 10^5$ simulated values of $\tilde{Y}^{\mathbf{V}}$.

| | $n = 50$ | | $n = 100$ | | $n = 500$ | |
|---|---|---|---|---|---|---|
| $d$ | $l^1_I = 50$ | $l^1_I = 100$ | $l^1_I = 50$ | $l^1_I = 100$ | $l^1_I = 50$ | $l^1_I = 100$ |
| 50 | 37.972 | 142.930 | 74.400 | 55.347 | 92.132 | 321.988 |
| 1000 | 184.780 | 624.712 | 282.090 | 535.981 | 1047.797 | 1982.664 |

The results obtained with $N = 10^5$ were acceptable. To see this, it is enough to compare the results in Tables 4.4 and 4.5 with those in Tables 4.1 and A.9: there are some differences between the parameters (due to greater uncertainty in the estimation of the involved quantiles) but, it seems, they are inside reasonable margins.

### 4.6.3 Algorithm to analyse a sample

The algorithm we propose to analyse all points in a sample is given in Algorithm 4. There, $\mathcal{X}_R$ is the set of points deemed as regular.

Table 4.4.: Obtained values of $(a, b)$ when $\Sigma = I_d$ for different values of $n, d$ and $l_I^1$ and $C_n^d \equiv C_n^d(0.05)$. Only $10^5$ simulated values for $\tilde{Y}^{\mathbf{V}}$ in the first step.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $l_I^1 = 50$ | | $l_I^1 = 100$ | | $l_I^1 = 50$ | | $l_I^1 = 100$ | | $l_I^1 = 50$ | | $l_I^1 = 100$ | |
| $d$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ |
| 5 | 0.0583 | 6.0850 | 0.0290 | 6.4198 | 0.0601 | 5.4816 | 0.0305 | 5.6744 | 0.06450 | 5.1759 | 0.0329 | 5.2607 |
| 50 | 0.0333 | 4.9870 | 0.0168 | 5.3563 | 0.0326 | 4.6579 | 0.0165 | 4.9365 | 0.0340 | 4.4449 | 0.0170 | 4.6927 |
| 100 | 0.0297 | 4.7470 | 0.0149 | 5.1214 | 0.0300 | 4.3772 | 0.0150 | 4.6435 | 0.0312 | 4.1760 | 0.0154 | 4.3896 |
| 500 | 0.0262 | 4.3144 | 0.0131 | 4.6484 | 0.0269 | 3.9731 | 0.0140 | 4.2024 | 0.0275 | 3.7194 | 0.0136 | 3.9522 |
| 1000 | 0.0256 | 4.1863 | 0.0122 | 4.5825 | 0.0262 | 3.8331 | 0.0134 | 4.0953 | 0.0257 | 3.6629 | 0.0123 | 3.8329 |

Table 4.5.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\| = C_n^d$, when $\Sigma = I_d$, for several values of $n, d$ using $a, b$ obtained in Table 4.4.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $l_I^1 = 50$ | | $l_I^1 = 100$ | | $l_I^1 = 50$ | | $l_I^1 = 100$ | | $l_I^1 = 50$ | | $l_I^1 = 100$ | |
| $d$ | Prob. | $\hat{l}_I^1$ | Prob. | $\hat{l}_I^1$ | Prob. | $\hat{l}_I^1$ | Prob. | $\hat{l}_I^1$ | Prob. | $\hat{l}_I^1$ | Prob. | $\hat{l}_I^1$ |
| 5 | 0.0488 | 52 | 0.0462 | 102 | 0.0492 | 51 | 0.0574 | 100 | 0.0494 | 50 | 0.0424 | 98 |
| 50 | 0.0528 | 48 | 0.0440 | 98 | 0.0518 | 51 | 0.0454 | 100 | 0.0466 | 48 | 0.0470 | 97 |
| 100 | 0.0464 | 51 | 0.0474 | 103 | 0.0534 | 50 | 0.0516 | 101 | 0.0460 | 50 | 0.0482 | 100 |
| 500 | 0.0530 | 50 | 0.0480 | 102 | 0.0506 | 48 | 0.0508 | 94 | 0.0492 | 49 | 0.0492 | 99 |
| 1000 | 0.0540 | 50 | 0.0532 | 110 | 0.0482 | 49 | 0.0494 | 96 | 0.0440 | 51 | 0.0524 | 104 |

Notice that Algorithm 4 always ends. Moreover, some points declared regular in initial rounds, could later be declared as outliers, because in Step 2) we make $\mathcal{X}_R = \varnothing$ every time a new outlier is identified. This is done so to reduce the masking effect.

### 4.6.4 How to reduce the role of the randomness in deciding if a point is outlier or not?

The proposed procedure is random. Despite the fact that random procedures are used frequently in Statistics (bootstrap, random forests, optimization algorithms with a random step, ...), some people can feel uncomfortable with this. As stated, the larger $l_\Sigma^r$ the lower the role of the randomness. A possibility to reduce further this role is to repeat Algorithm 4 a not so large number of times, $T$, using a significance level $\alpha$. Thus, since points $\mathbf{x}$ satisfying that, let us say, $\|\mathbf{x} - \boldsymbol{\mu}\|_\Sigma = C_n^d(\delta)$ are declared as outliers a proportion $\alpha$ of times, we could resort to declare as outliers those points which have been identified as outliers more than a proportion $\alpha$ of times along the $T$ repetitions.

This criteria can be strengthened (resp. relaxed) identifying as outliers only the points declared as outliers a number of times higher (resp. lower) than the $0.95$ (resp. $0.05$) quantile of a binomial with parameters $T$ and $\alpha$.

**Algorithm 4:** Procedure to look for outliers in a sample

```
Let $\mathcal{X}$ be the set of all points in a sample.
  0) Let $\mathcal{X}_R = \varnothing$
  1) Take a random projection and analyse all points in $\mathcal{X}$.
  2) If some points have been declared as outliers, delete them from $\mathcal{X}$, set $\mathcal{X}_R = \varnothing$,
     and go to Step 1). Else, add the points declared as non-outliers, if any, to $\mathcal{X}_R$.
  3) If $\mathcal{X}_R \neq \mathcal{X}$ go to Step 1). Else, return $\mathcal{X}_R$.
```

## 4.7 Numerical studies

In this section we analyse the behaviour of the method thorough simulated experiments and real datasets. Here, only the results for $n = 50$ are shown; see Section A.2 in the Appendix for the complete results. We also compare our procedure with some existing methods.

The computations of the constants $a$ and $b$ determining the tests are carried out as described in Subsections 4.5.1 and 4.5.2 with $N = 10^6$ simulated values of $\tilde{Y}_n^{\mathbf{V}}$. Analogously to Chapter 3, the estimate of $l_\Sigma^r$ (which is the mean of the obtained values) will be denoted by $\hat{l}_\Sigma^r$ and use the same short notation as in Section 3.7.

### 4.7.1 Simulations

We use the notation introduced in Subsection 4.4. All the results are obtained from $5000$ replicated simulations.

Table 4.6 (see Table A.8 in the Appendix for its non-robust version) shows the proportion of times we have declared a point with Mahalanobis norm $C_n^d(\delta)$ with $\delta = 0.05$ as an outlier for $n = 50$ and several values of $d$. More results including the cases $n = 100, 500$ are in Table A.9 in the Appendix. The results are acceptable because the proportions are close to the intended: the percentiles $0.025$ and $0.975$ of the rv proportion of rejections with a theoretical proportion equal to $0.05$ in $5000$ trials are $0.044$ and $0.0562$ and the price we pay to achieve robustness seems to produce a slightly conservative test, since we obtain 18 (out of 120) proportions outside the confidence interval, all of them in the upper part, but with the maximum (equal to $0.0668$) being close to the upper boundary of the target.

We also see that the mean number of projections $\hat{l}_1^1, \ldots, \hat{l}_4^1$ are always greater or, if lower, very close to $l_I^1$ (giving support to the fact that the asymptotic result shown in Proposition 4.5.2 also holds for finite sample sizes), being $\hat{l}_3^1$ always the largest one. Moreover $\hat{l}_1^1, \hat{l}_2^1$ and $\hat{l}_4^1$ are always reasonably similar to $\hat{l}_I^1$ and, also, all of them are close to the goal $l_I^1$. The values obtained for $\hat{l}_3^1$ increase with the dimension and, when $d = 500$ or $10^3$, they are an order of magnitude larger than intended.

Table 4.6.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$, for $n = 50$ and several values of $d$, $\Sigma$ and $l_I^1$. We also show the sample means of $\tilde{K}_n$.

| $d$ | $l_I^1$ | $\hat{l}_I^1$ | $I_d$ | $\hat{l}_1^1$ | $\Sigma_1^d$ | $\hat{l}_2^1$ | $\Sigma_2^d$ | $\hat{l}_3^1$ | $\Sigma_3^d$ | $\hat{l}_4^1$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 51 | 0.0528 | 49 | 0.0571 | 49 | 0.0541 | 186 | 0.0668 | 50 | 0.0569 |
| | 100 | 98 | 0.0560 | 99 | 0.0558 | 103 | 0.0553 | 366 | 0.0580 | 99 | 0.0572 |
| 100 | 50 | 49 | 0.0507 | 48 | 0.0496 | 50 | 0.0501 | 249 | 0.0628 | 50 | 0.0489 |
| | 100 | 100 | 0.0538 | 101 | 0.0519 | 100 | 0.0526 | 526 | 0.0603 | 98 | 0.0494 |
| 500 | 50 | 49 | 0.0481 | 50 | 0.0507 | 50 | 0.0518 | 552 | 0.0628 | 50 | 0.0483 |
| | 100 | 100 | 0.0520 | 102 | 0.0509 | 99 | 0.0545 | 1111 | 0.0589 | 101 | 0.0538 |
| 1000 | 50 | 50 | 0.0496 | 50 | 0.0538 | 49 | 0.0534 | 790 | 0.0586 | 50 | 0.0500 |
| | 100 | 100 | 0.0520 | 101 | 0.0476 | 102 | 0.0507 | 1601 | 0.0549 | 99 | 0.0553 |

Table 4.7 (see Table A.10 in the Appendix for its non-robust version) shows the estimations of the probability of declaring a point as an outlier when its Mahalanobis norm is $1.2C_n^d$ or $2C_n^d$ and $n = 50$. Complete results are in Table A.11 in the Appendix. The values corresponding to $\Sigma = I_d, \Sigma_4^d$ are the highest, being those of the identity slightly better. The worst results (and the highest number of required projections) are obtained for $\Sigma = \Sigma_3^d$; the remaining ones being similar to those corresponding to the identity. Obviously when $l_I^1$ increases, so does the probability to detect the outliers. We also see an increase of the power when $n$ becomes larger and a slight decrease when $d$ becomes larger. This makes sense because for larger values of $n$, the estimation of the parameters is more accurate, while the larger $d$, the greater the noise in the sample.

Other features of the procedure such that the proportion of observation wrongly classified as outliers or the effect of the masking are analyzed in Subsection 4.7.2.

### 4.7.2 Comparison with other procedures

In this subsection we compare our procedure (denoted RP) with some existing methods proposed for high-dimensional data: the principal component outlier detection procedure (denoted PCOut) of Filzmoser *et al.* [61] and the minimum diagonal product method (denoted MDP) of Ro *et al.* [116]. The main interest in this subsection is twofold: first one is to check how the dimension and the covariance matrix affect those

Table 4.7.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ with $n = 50$ and several values of $d$, $\Sigma$ and $l_I^1$.. We also show the sample means of $\tilde{K}_n$.

| $d$ | $\|X\|_\Sigma$ | $l_I^r$ | $\hat{l}_I^r$ | $I_d$ | $\hat{l}_1^r$ | $\Sigma_1^d$ | $\hat{l}_2^r$ | $\Sigma_2^d$ | $\hat{l}_3^r$ | $\Sigma_3^d$ | $\hat{l}_4^r$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | $1.2C_n^d$ | 50 | 48 | .2378 | 48 | .2247 | 48 | .2338 | 163 | .1752 | 48 | .2333 |
| | | 100 | 93 | .2729 | 96 | .2412 | 95 | .2617 | 313 | .1867 | 92 | .2639 |
| | $2C_n^d$ | 50 | 12 | .8817 | 13 | .8660 | 12 | .8830 | 47 | .6575 | 12 | .8912 |
| | | 100 | 16 | .9259 | 19 | .9061 | 16 | .9153 | 74 | .6985 | 16 | .9229 |
| 100 | $1.2C_n^d$ | 50 | 48 | .2235 | 49 | .2093 | 48 | .2146 | 223 | .1729 | 49 | .2191 |
| | | 100 | 97 | .2387 | 97 | .2236 | 95 | .2320 | 460 | .1723 | 96 | .2487 |
| | $2C_n^d$ | 50 | 13 | .8829 | 13 | .8678 | 13 | .8734 | 70 | .6289 | 13 | .8743 |
| | | 100 | 18 | .9150 | 19 | .9081 | 18 | .9115 | 113 | .6738 | 18 | .9160 |
| 500 | $1.2C_n^d$ | 50 | 50 | .2160 | 48 | .2132 | 49 | .2168 | 518 | .1711 | 50 | .2198 |
| | | 100 | 97 | .2454 | 99 | .2375 | 96 | .2399 | 973 | .1761 | 97 | .2412 |
| | $2C_n^d$ | 50 | 13 | .8771 | 13 | .8617 | 13 | .8780 | 150 | .6139 | 13 | .8726 |
| | | 100 | 18 | .9166 | 18 | .9185 | 18 | .9075 | 249 | .6513 | 18 | .9090 |
| 1000 | $1.2C_n^d$ | 50 | 49 | .2202 | 51 | .2136 | 49 | .2159 | 700 | .1632 | 49 | .2156 |
| | | 100 | 98 | .2470 | 97 | .2338 | 97 | .2429 | 1383 | .1616 | 96 | .2366 |
| | $2C_n^d$ | 50 | 13 | .8797 | 13 | .8728 | 13 | .8729 | 214 | .6128 | 13 | .8674 |
| | | 100 | 19 | .9116 | 19 | .9134 | 19 | .9093 | 360 | .6551 | 19 | .9124 |

methods, second one is to see the capability of the procedures to detect multiple outliers once the parameters have been fixed to have a similar behaviour where there is no outlier in the sample.

For this, we have chosen two settings: in the first one we handle a clean sample and compute the proportion of points which are declared as outliers. In the second one we handle a sample with 10% outliers and analyze the proportion of them which are detected by the procedures. In both cases, we have employed two different sample sizes $n = 50, 100$, three dimensions $d = 50, 500, 1000$ and seven covariance matrices: first one is the identity, the elements in the second matrix, $S_2 = (s_{i,j})$, are $s_{i,j} = e^{-|i-j|/d}$. To construct the third matrix, $S_3$, we generate a matrix $A$ whose elements are iid $N(0,1)$ and take $S_3 = A'A$. The remaining four matrices are the $\Sigma_i^d$'s we handled in Subsection 4.5.2. We include here the results corresponding to $I_d, S_2$ and $S_3$ and leave for Section A.2 in the Appendix those obtained with the $\Sigma_i^d$'s. Thus, the reported cases here include a case with independent marginals, another one with relatively high correlations and a third one in which the correlations are random. For these three frameworks, we have generated the data with the identity as covariance matrix. Then, we have multiplied these data by the appropriate matrix to obtain the desired covariance; thus, somehow, we handle the same data for the three covariance matrices. We have done 500 simulations. Matrix $A$ varies from each simulation.

The approaches of PCOut and MDP are implemented in the functions pcout and rmdp in the R packages mvoutlier and Rfast, respectively. We have kept the default parameters provided by those functions except for the case when we use rmdp that we fix itertime $= d^{1.5}$ in order to follow the suggestion in the help comment that the number of iterations should be similar to the dimension for sample sizes equal to $50$. From there, we have concluded that for higher sample sizes, the number of iterations should be greater than the sample size. Regrettably, this makes MDP quite slow. Thus, we have not run the procedure when $d = 1000$, since it took 364.18 seconds in the first setting to compute five values when $n = 50$.

Since the default options of the functions pcvout and rmdp lead to a discovery of around 10% of outliers in the clean samples, for each dimension and sample size, we have fixed the parameters $a, b$ for RP in order to obtain approximately this percentage of discoveries in Table 4.8. This goal was achieved by taking $a, b$ such that $E\left(\tilde{K}_n \mid \|\mathbf{X}\|_\Sigma = q_d^n\right) = 50$ and $\mathbf{P}(\mathbf{X}$ declared outlier $|\|\mathbf{X}\| = q_d^n) = 0.1$ where $q_d^n$ is the $0.75$-quantile of the square root of a random sample with size $n$ drawn from a $\chi_d^2$. Those parameters have been used in both settings.

The results obtained when using the covariance matrices $\Sigma_i^d$ are similar to those obtained when $\Sigma = S_3$. Those cases are handled as described before, excepting for the fact that we have used a randomly chosen basis in order to prevent the matrices $\Sigma_i^d$ being diagonal. We did not this before because RP is invariant against those rotations. However, it seems that MDP may depend on whether $\Sigma$ is diagonal or not and, the first step in PCOut is to standardise the data, thus making all cases in which $\Sigma$ is diagonal equivalent to $\Sigma = I_d$.

### Handling a clean sample

Here we generate a sample from a $N_d(\mathbf{0}, \Sigma)$ without outliers and compute the proportion of observations in the sample that each procedure declares as outliers. In this case there are no outliers and, therefore, no observation should be declared as outlier. The proportion of outliers is not interesting here (because you can get the desired proportion tuning appropriately the parameters). We are interested, however, in detecting the stability of the procedures; more precisely in seeing if the dimension and/or the covariance matrix affect to the capacity of the procedures to detect outliers.

Our conclusion from those simulations (see Tables 4.8 and A.12 in the Appendix), is that the behaviour of MDP is very different depending on whether $\Sigma$ is diagonal or not and, whether $\Sigma \neq I_d$. The dimension also affects its behaviour. The increment of the sample size decreases the number of wrongly detected outliers. Procedures PCOut and

Table 4.8.: Proportion of outliers found in a clean data set for several covariance matrices.

| | | | MDP | | | PCOut | | | RP | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d$ | $I_d$ | $S_2$ | $S_3$ | $I_d$ | $S_2$ | $S_3$ | $I_d$ | $S_2$ | $S_3$ |
| 50 | 50 | .1360 | .1190 | .1158 | .1025 | .1377 | .1110 | .1108 | .0909 | .1085 |
| | 500 | .1404 | .0320 | .0585 | .0950 | .1308 | .1003 | .1149 | .1003 | .1101 |
| | 1000 | — | — | — | .1028 | .1317 | .1018 | .1132 | .1000 | .1141 |
| 100 | 50 | .0735 | .0896 | .0739 | .1022 | .1219 | .1086 | .1044 | .0790 | .1020 |
| | 500 | .0827 | .0187 | .0498 | .0829 | .1235 | .0813 | .1104 | .0810 | .1104 |
| | 1000 | — | — | — | .0787 | .1232 | .0808 | .1108 | .0830 | .1098 |

RP are quite stable when the dimension varies, in spite of the fact that PCOut tends to declare more outliers when $n = 50$. This effect is more noticeable in the results in Table A.12. Additionally, PCOut seems to declare less outliers when the dependence is not too strong while the opposite happens with RP. Overall, results from RP are more stable than those from MDP or PCOut.

### Handling a sample with 10% of outliers

In this task, we generate a clean sample, with size $.9n$ from a $N_d(\mathbf{0}, \Sigma)$ and we add $n_{out} = .1n$ outliers with distributions $N_d(\mathbf{0}, \Sigma)$ given that $\|\mathbf{X}\|_\Sigma = p_i, i = 1, \ldots, n_{out}$. Here we take $q_i, i = 1, \ldots, n_{out}$, an equispaced sequence from $.95$ to $.99$ and, then, $p_i, i = 1, \ldots, n_{out}$ are the square roots of the $q_i$-quantiles of the $\chi^2_d$ distribution.

Tables 4.9 and A.13 (last one in the Appendix) show the proportion of outliers which were correctly identified along 500 repetitions; thus, the higher the proportions, the better. MDP does an acceptable work when $\Sigma = I_d$, with better results than PCOut, but, regrettably, its behaviour seems to deteriorate in the other two situations in Table 4.9, mostly when $d$ increases. In the situations handled in Table A.13 this method gives the best results when $\Sigma = \Sigma_3^d$. It is not too bad when $d = 50$ with the remaining matrices, but its behaviour deteriorates noticeably when $d$ increases.

Broadly speaking, we can say that PCOut is the winner when $\Sigma = S_2$ while RP is the choice in the remaining cases with $\Sigma \neq \Sigma_3^d$. Those results suggest that, on highly dependent situations, the user could benefit from using PCOut; while one should use RP in no so dependent ones. The problem with this advice is that in order to decide the kind of situation we need to estimate the covariance matrix.

Table 4.9.: Samples contain 10% of real outliers. Columns show the proportion of them correctly identified.

| | | | MDP | | | PCOut | | | RP | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d$ | $I_d$ | $S_2$ | $S_3$ | $I_d$ | $S_2$ | $S_3$ | $I_d$ | $S_2$ | $S_3$ |
| 50 | 50 | .1959 | .1216 | .1357 | .1564 | .2912 | .1684 | .2856 | .2032 | .2844 |
| | 500 | .1842 | .0340 | .0688 | .1196 | .1756 | .1040 | .1568 | .1056 | .1420 |
| | 1000 | — | — | — | .1060 | .1552 | .1112 | .1576 | .1092 | .1424 |
| 100 | 50 | .1301 | .0902 | .0905 | .2112 | .3120 | .2282 | .3076 | .1816 | .2864 |
| | 500 | .1424 | .0185 | .0610 | .0856 | .1808 | .0928 | .1790 | .1138 | .1642 |
| | 1000 | — | — | — | .0812 | .1598 | .0852 | .1478 | .1064 | .1526 |

### 4.7.3 The procedure in practice: Two real data examples

The practical relevance of the proposed test is illustrated on two well-known real data sets. They have been studied by Hubert *et al.* [80]. Those data are considered there as functional; however, all observations in both sets have been measured on the same values of the independent variable; thus they can be also considered as $d$-dimensional.

We compute $a$ and $b$ as in Section 4.5.1 with $\mathbf{P}\left(\left|\tilde{Y}^{\tilde{K}_n}\right| > b \mid \|\mathbf{X}\|_{\Sigma} = C_n^d\right) = 0.05$, $l_I^1 = 100$ and $N = 10^6$. Consequently, a point $\mathbf{x}$ such that $\|\mathbf{x}\|_{\Sigma} = C_n^d$ should be identified as outlier 5% of times. We have applied the proposed method $T = 100$ times to every point in the sample at hand and we have declared as outliers those points that were identified as outliers 5% of times or more, following the procedure described in Subsection 4.6.4.

In the analysis we show the outliers identified by the methods introduced in this paper (denoted RP), in Hubert *et al.* [80] (denoted Hub), in Filzmoser *et al.* [61], (denoted PCOut), and Ro *et al.* [116] (denoted MDP). PCOut and MDP are handled with their default parameters, excepting that we take `itertime` $= d$ in MDP according to the suggestion that this value should be similar to the dimension when $n = 50$.

#### Wine Data

This dataset contains the proton nuclear magnetic resonance spectra of 40 different wine samples (Larsen *et al.* [94]). As in Hubert *et al.* [80], we select the region between wavelengths 5.37 and 5.62, on which each sample has $d = 397$ measurements.

Table 4.10 shows the data identified as outliers by the considered procedures. Those curves are represented in Figure 4.1 with coloured lines. We see that the curve 37 has two large peaks around wavelength 5.4 and may be considered an isolated outlier.

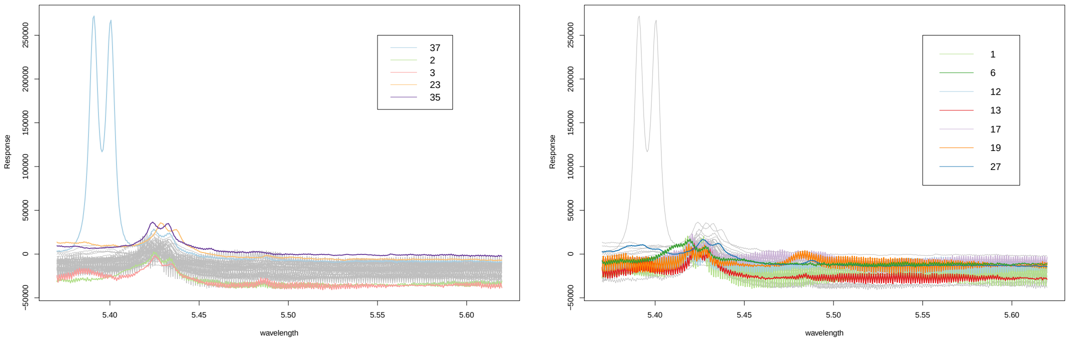


**Figure 4.1.:** The left panel shows the outliers which are detected both in [80] and with the procedure we present here. The right panel shows the outliers detected with the proposed method but not in [80].

The curves 1, 12, 17 and 19 oscillate too much, as shown in Figure 4.2, where the boxplot of the indices of the oscillation $\sum_{j=1}^{d}(X_{j+1}^{i} - X_{j}^{i})^2$ for each point $\mathbf{X}^i, i = 1, \ldots, 40$ appear. Such boxplot employs the fences as in Definition 2.3.1 of Iglewicz and Banerjee [81] with $\delta = 0.05$. Unlike Hub, RP and PCOut declare them as outliers (except for 17 which is not declared by PCOut): RP with probability greater than 0.6 and PCOut with weights 0.04 (weights close to zero indicate potential outliers).

**Table 4.10.:** Wines identified as outliers. Each number in row RP (resp. PCOut) is the proportion of times this wine was declared outlier by RP (resp. the weight of this wine. Low weights identify potential outliers). Marked with an X (resp. white) cells mean the wine was (resp. not) identified by the corresponding procedure.

| | 1 | 2 | 3 | 6 | 12 | 13 | 17 | 18 | 19 | 23 | 27 | 35 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RP | 0.89 | 0.16 | 0.24 | 0.05 | 0.67 | 0.32 | 0.63 | | 0.64 | 0.61 | 0.06 | 0.19 | 1.00 |
| Hub | | X | X | | | | | | | X | | X | X |
| PCOut | 0.04 | 0.06 | 0.16 | 0.20 | 0.04 | 0.08 | | 0.16 | 0.04 | 0.14 | | | 0.04 |
| MDP | | | | | | | | | | | | | X |

RP, PCOut and Hub also declare the wines 2, 3 and 23 as outliers. Wine 35 is only identified by RP and Hub. Figure 4.1 shows that those curves are in the external part of the bulk of the data: 2 and 3 in the bottom and 23 and 35 in the top part. RP and PCOut additionally detect 13 as outlier; this curve is in the bottom part of the data just above of 2 and 3 (see Figure 4.1). RP also detects the curves 6 and 27 (in coloured lines in Figure 4.1), PCOut only the curve 6, and Hub none of them. We see that curve 6 starts to increase before the other curves; while 27 has a similar shape to the curve 3 (which is declared as an outlier by Hub) but in the top part of the data. However, the number of times these curves have been detected by RP (well below the .95-quantile of a binomial with parameters 100 and .05, which is 9) make them doubtful as outliers from the RP point of view. PCOut gives the maximum weight, 0.2, to 6, i.e. among all the outliers that PCOut detects, this curve is the least anomalous.

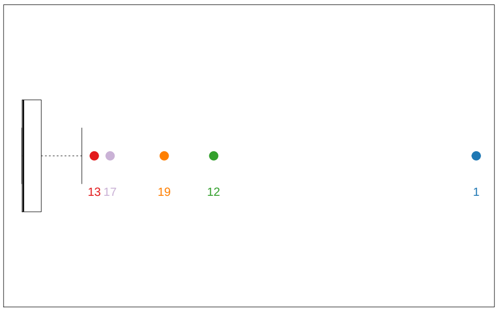


Figure 4.2.: Boxplot using fences of Definition 2.3.1 of Iglewicz and Banerjee [81] with $\delta = 0.05$ for the squared of the differences among the components of each point in the wine data.

The difference between the detected curves by PCOut and RP is that PCOut detects the curve 18 (with the same weight that curve 3), and RP detects the curves 17, 27 and 35. Figure 4.3 shows these curves. It seems the curve 18 has some fluctuation, however this curve does not appear as an outlier in the boxplot of Figure 4.2.

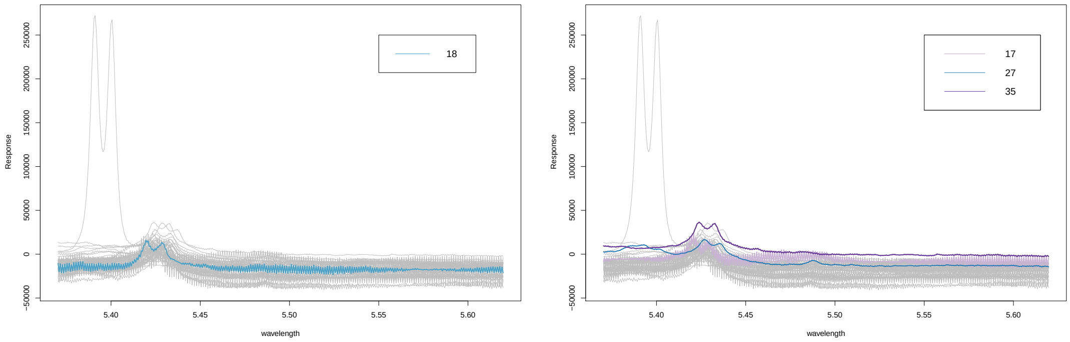


Figure 4.3.: The left panel shows the outlier which was detect by PCOut but not with our method. The right panel shows the outliers detected with our method but not with PCOut.

In conclusion, it seems that PCOut and RP detect better the shape outliers than Hub. RP also detects curves that have not so big peculiarities or those which are in the border of the bulk of the data albeit with lower probability.

### Octane data

This data set consists of 39 near infrared spectra of gasoline samples over $d = 226$ wavelengths ranging from 1102 nm to 1552 nm with measurements every two nm. It is known that samples 25, 26 and 36-39 have a very different spectrum because they contain added ethanol (Esbensen *et al.* [49], Rousseeuw *et al.* [121] and Hubert *et al.* [80]). Table 4.11 shows the data identified as outliers by the considered procedures.

Table 4.11.: Outliers in the gasolines. Each number in row RP (resp. PCOut) is the proportion of times this gasoline was declared outlier by RP (resp. the weight of this wine. Low weights identify potential outliers). Marked with an X (resp. white) cells mean the gasoline was (resp. not) identified by the corresponding procedure.

| | 6 | 23 | 25 | 26 | 34 | 36 | 37 | 38 | 39 |
|---|---|---|---|---|---|---|---|---|---|
| RP | 0.11 | 0.06 | 0.99 | 1.00 | 0.28 | 1.00 | 1.00 | 1 .00 | 0.99 |
| Hub | | | X | X | | X | X | X | X |
| PCOut | | 0.10 | 0.04 | 0.04 | 0.08 | 0.04 | 0.04 | 0.04 | 0.04 |
| MDP | | | X | X | | X | X | X | X |

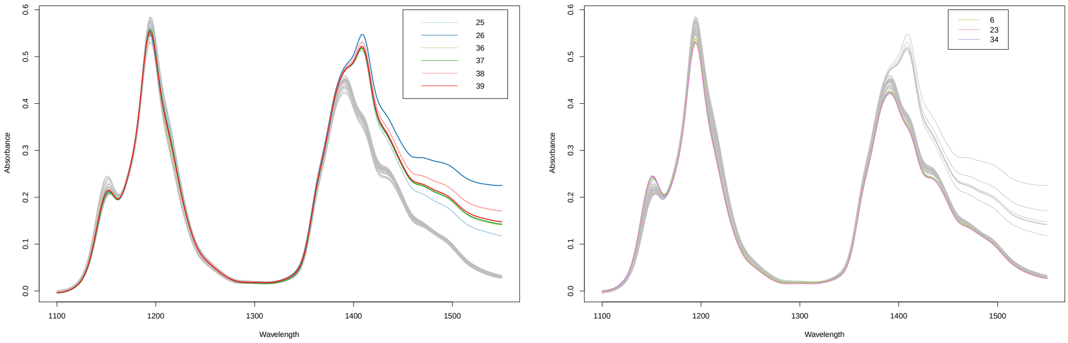


Figure 4.4.: Outliers detected in octane data with RP and PCOut methods: the left panel shows the outliers which are also detected with Hub and MDP methods. The right panel shows the outliers detected only with RP and PCOut methods.

All the curves identified as outliers are plotted with coloured lines in Figure 4.4. Curiously, Hub and MDP (resp. PCOut and RP excepting for the gasoline 6) detect the same curves as outliers. Clearly the curves 25, 26 and 36-39, represented in the left panel, are persistently outlying from wavelength 1390 onward and all procedures detect them. The curves 23 and 34, represented in the right panel, are declared outliers by PCOut and RP but not by Hub and MDP. Additionally, RP detects the curve 6. We see that these three curves are in the border of the bulk of the data and they are slightly separated from the rest on wavelengths around 1150, 1195 and 1390. Anyhow, curve 23 is only declared as outlier 6% of times what makes it doubtful from the point of view of RP. This is the curve with the highest weight, 0.1, when we apply PCOut.

Similarly to the wine dataset, it seems that PCOut and RP detect the outliers which are far away from the bulk of the data (curves 25, 26 and 36 to 39) and those which always are in the border of the data (23 and 34, and additionally RP detects 6).

# 5 On a class of uniformity tests on the hypersphere

*The beauty of Mathematics only shows itself to more patient followers.*

— **Maryam Mirzakhani**

In this chapter we propose a projection-based class of uniformity tests on the hypersphere. We show its relation to the Sobolev class of uniformity tests and, also, to other existing tests such as Watson, Ajne, and Rothman tests, and introduce the first instance of an Anderson–Darling-like test. We finish the chapter with a simulation study which corroborates the theoretical findings and evidences that, for certain scenarios, the new tests are competitive against previous proposals. Real data examples illustrate the usage of the new tests.

## 5.1 Projected statistics: A new approach to testing uniformity

As we mentioned in the Introduction, testing uniformity on the hypersphere $\Omega^{d-1}$ is formalized as the testing of

$$\mathbf{H}_0 : \mathbf{P} = \nu_{d-1} \quad \text{vs.} \quad \mathbf{H}_1 : \mathbf{P} \neq \nu_{d-1},$$

from a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of iid observations of $\mathbf{P}$. Remember that $\nu_{d-1}$ was defined in Subsection 2.5.1.

### 5.1.1 Genesis

Our proposal is inspired by the projection-based test of Cuesta-Albertos *et al.* [37], see Subsection 2.6.2. Within the same approach of Cuesta-Albertos *et al.* [37], an alternative to (2.21) is the well-known weighted quadratic norm by Anderson and Darling [7]:

$$Q^w_{n,d-1,\gamma} := n \int_{-1}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, w(F_{d-1}(x)) \, \mathrm{d}F_{d-1}(x), \tag{5.1}$$

where the specifications $w \equiv 1$ and $w(u) = 1/(u(1-u))$ for the weight $w : [0,1] \to \mathbb{R}$ yield the Cramér–von Mises and Anderson–Darling test statistics, respectively. In addition to the extra flexibility provided by the weight $w$ in (5.1) in comparison with (2.21), the use of quadratic norms instead of sup-norms in goodness-of-fit tests is typically advised in practise, as they tend to provide higher powers (see, e.g., Stephens [127, Section 5] or D'Agostino and Stephens [40, page 110]).

Our class of test statistics is based on $Q^w_{n,d-1,\gamma}$ but, rather than drawing several random directions and aggregating afterwards the outcomes of the associated tests (as in Cuesta-Albertos *et al.* [37]), the statistic itself gathers information from all the directions on $\Omega^{d-1}$: it is defined as the *expectation* of $Q^w_{n,d-1,\gamma}$ with respect to $\gamma \overset{d}{=} \nu_{d-1}$. Therefore, the proposed test rejects $\mathbf{H}_0$ for large values of the test statistic

$$\begin{aligned} P^w_{n,d-1} &:= \mathrm{E}_\gamma \left( Q^w_{n,d-1,\gamma} \right) \\ &= n \int_{\Omega^{d-1}} \left[ \int_{-1}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, w(F_{d-1}(x)) \, \mathrm{d}F_{d-1}(x) \right] \nu_{d-1}(\mathrm{d}\gamma). \end{aligned} \tag{5.2}$$

The choice of $\nu_{d-1}$ as the distribution for $\gamma$ is canonical: it is the only sample-independent distribution that guarantees the invariance of (5.2) against any rotation of the sample, this being the fundamental property that any uniformity test on $\Omega^{d-1}$ must have. In addition, as we mentioned in the Introduction, the idea of integrating along all the unit-norm directions has been already considered.

An obvious generalization of $P^w_{n,d-1}$ follows by substituting the weight function $w$ with the integration with respect to a (positive) $\sigma$-finite Borel measure $W$ on $[0,1]$, giving the statistic

$$P^W_{n,d-1} := n\mathrm{E}_\gamma \left( \int_{-1}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, \mathrm{d}W(F_{d-1}(x)) \right). \tag{5.3}$$

Notice that if $W$ is a probability measure, we integrate with respect to a probability whose cdf is $x \mapsto W\{[0, F_{d-1}(x)]\}$. This generalisation, for instance, allows for weighting schemes that are concentrated on denumerable or finite sets. We will consider this generalized formulation (5.3) henceforth in the chapter, since it unifies many tests statistics (see Sections 5.1.3–5.1.5) under the projection-based view. For the sake of simplicity, $W$ will denote indistinctly a measure and its cdf when there is no possible ambiguity. Furthermore, we will focus only on symmetric measures with respect to $1/2$, as our first result shows that, due to the construction of $P^W_{n,d-1}$, this restriction does not imply a loss in generality.

**Proposition 5.1.1.** *Let $W$ and $\tilde{W}$ be $\sigma$-finite Borel measures on $[0,1]$ such that $\tilde{W}\{(0,t)\} = \frac{1}{2}\left(W\{(0,t)\} + W\{(1-t,1)\}\right)$, $t \in [0,1]$. Then,*

$$P^W_{n,d-1} = P^{\tilde{W}}_{n,d-1}.$$

*Proof.* A simple change of variables gives

$$\begin{aligned}\mathrm{E}_\gamma\left(\int_{-1}^{0} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, \mathrm{d}W(F_{d-1}(x))\right) = \\ \mathrm{E}_\gamma\left(\int_{1}^{0} \{F_{n,\gamma}(x^-) - F_{d-1}(x)\}^2 \, \mathrm{d}W(1 - F_{d-1}(x))\right)\end{aligned}$$

employing the facts that $F_{n,\gamma}(-x) = 1 - F_{n,-\gamma}(x^-)$, $x \in [0,1]$, and that $\gamma \overset{d}{=} \nu_{d-1}$. However, $F_{n,\gamma}(x^-) = F_{n,\gamma}(x)$ except for $x \in \{\boldsymbol{\gamma}'\mathbf{X}_1, \ldots, \boldsymbol{\gamma}'\mathbf{X}_n\}$. Thus,

$$\int_{1}^{0} \{F_{n,\gamma}(x^-) - F_{d-1}(x)\}^2 \mathrm{d}W(1 - F_{d-1}(x)) \neq \int_{1}^{0} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \mathrm{d}W(1 - F_{d-1}(x))$$

only if $W(\{F_{d-1}(\boldsymbol{\gamma}'\mathbf{X}_1), \ldots, F_{d-1}(\boldsymbol{\gamma}'\mathbf{X}_n)\}) > 0$. Now, let $\mathcal{D}_W$ denote the points of discontinuity of $x \mapsto W([0,x])$. This set it at most denumerable, and consequently,

$$\begin{aligned}\mathrm{E}_\gamma\left(\int_{1}^{0} \{F_{n,\gamma}(x^-) - F_{d-1}(x)\}^2 \, \mathrm{d}W(1 - F_{d-1}(x))\right) \neq \\ \mathrm{E}_\gamma\left(\int_{1}^{0} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, \mathrm{d}W(1 - F_{d-1}(x))\right)\end{aligned} \tag{5.4}$$

only if $\mathrm{Prob}\left[\boldsymbol{\gamma} \in \Omega^{d-1} : \{F_{d-1}(\boldsymbol{\gamma}'\mathbf{X}_1), \ldots, F_{d-1}(\boldsymbol{\gamma}'\mathbf{X}_n)\} \cap \mathcal{D}_W \neq \varnothing\right] > 0$. But this probability is bounded by $\sum_{z \in \mathcal{D}_W} \sum_{m=1}^{n} \mathrm{Prob}\left\{\boldsymbol{\gamma} \in \Omega^{d-1} : \boldsymbol{\gamma}'\mathbf{X}_m = F^{-1}_{d-1}(z)\right\}$ and, since each addend represents the probability of $\gamma$ belonging to a particular hyperplane and $\gamma$ has a continuous distribution, the sum equals zero. Therefore, we have that the inequality in (5.4) is impossible.

Observing the implicit sign change in $\mathrm{d}W(1 - F_{d-1}(x))$ and remembering the definition of $P^W_{n,d-1}$, we have that

$$P^W_{n,d-1} = 2n\mathrm{E}_\gamma\left(\int_{0}^{1} \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \, \mathrm{d}\tilde{W}(F_{d-1}(x))\right). \tag{5.5}$$

From (5.5), undoing the previous change of variables and recalling that $\mathrm{d}\tilde{W}(F_{d-1}(x)) = \mathrm{d}\tilde{W}(1 - F_{d-1}(x))$ by construction, we conclude that $P^W_{n,d-1} = P^{\tilde{W}}_{n,d-1}$. □

### 5.1.2 $U$-statistic form of $P^W_{n,d-1}$

Simple computations show that

$$P^W_{n,d-1} = n\int_{-1}^{1}\left\{\mathrm{E}_\gamma\left(F_{n,\gamma}(x)^2\right) - F_{d-1}(x)^2\right\}\mathrm{d}W(F_{d-1}(x)) \tag{5.6}$$

$$= \int_{-1}^{1}\left\{\frac{1}{n}\sum_{i\neq j}A_{ij}(x) + F_{d-1}(x)(1-nF_{d-1}(x))\right\}\mathrm{d}W(F_{d-1}(x)), \tag{5.7}$$

where (5.6) follows from $\mathrm{E}_\gamma\left(F_{n,\gamma}(x)\right) = F_{d-1}(x)$ and (5.7) by $\mathrm{E}_\gamma\left(F_{n,\gamma}(x)^2\right) = n^{-1}F_{d-1}(x) + n^{-2}\sum_{i\neq j}A_{ij}(x)$, where

$$A_{ij}(x) := \int_{\Omega^{d-1}}1_{\{\gamma'\mathbf{X}_i\leq x,\gamma'\mathbf{X}_j\leq x\}}\,\mathrm{d}\nu_{d-1}(\gamma). \tag{5.8}$$

We can not split (5.7) into two addends if $W$ is not a finite measure, since neither of them would be finite. This is exactly the case for the measure associated to the Anderson–Darling weight.

The term in (5.8) is the driver of the $P^W_{n,d-1}$ statistic. Geometrically, it is the proportion of the hypersphere $\Omega^{d-1}$ covered by the intersection of two hyperspherical caps centered, respectively, at $\mathbf{X}_i$ and $\mathbf{X}_j$ with solid angle $\theta_x := \pi - \cos^{-1}(x)$ radians. Evaluating $A_{ij}(x)$, for arbitrary $x\in[-1,1]$, $d\geq 2$, and $i\neq j$, is not straightforward, despite formulae for the area of the intersection of two hyperspherical caps being available in Lee and Kim [97, page 4]. These formulae involve 10 out of the 25 possible cases (precisely, the cases: 1, 2, 4, 5, 6, 8, 14, 15, 23, 25), depending on the values of $\theta_x$ and $\theta_{ij} := \cos^{-1}(\mathbf{X}_i'\mathbf{X}_j)$, and require univariate integrals on $F_{d-1}$.

We simplify the computation of $A_{ij}(x)$, henceforth denoted by $A(\theta_{ij}, x)$ due to its dependence on $\theta_{ij}$, with Proposition 5.1.4. Before we need Lemma 5.1.2, which also is useful for several proofs of the main results, and Lemma 5.1.3.

**Lemma 5.1.2.** *Let* $d\geq 3$. *Then:*

$$\int_0^{\cos(\theta/2)}F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t) = F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right) + \frac{\theta}{4\pi} - \frac{3}{4}. \tag{5.9}$$

*Proof.* Denote $\phi_d(\theta)$ to the left hand side of (5.9). Then:

$$\begin{aligned}\phi_d(\theta) &= F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right) - \frac{1}{2} + \int_0^{\cos(\theta/2)}\left(F_{d-2}\left(t\tan(\theta/2)/(1-t^2)^{1/2}\right) - 1\right)\mathrm{d}F_{d-1}(t)\\ &=: F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right) - \frac{1}{2} + \phi_d^*(\theta).\end{aligned} \tag{5.10}$$

Compute the derivative of $\phi_d^*(\theta)$ with respect to $\theta$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\phi_d^*(\theta) &= -\frac{1}{2}\sin\left(\tfrac{\theta}{2}\right)\left(F_{d-2}(1)-1\right)F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right)\\ &\quad+\frac{1}{2}\frac{\sec^2(\theta/2)}{\mathrm{B}\left(\frac{1}{2},\frac{d-2}{2}\right)\mathrm{B}\left(\frac{1}{2},\frac{d-1}{2}\right)}\int_0^{\cos(\theta/2)} t\left(1-\frac{t^2}{\cos^2(\theta/2)}\right)^{\frac{d-4}{2}}\mathrm{d}t = \frac{1}{4\pi}.\end{aligned} \tag{5.11}$$

The proof is concluded from (5.10) and (5.11), and since $\phi_d^*(\pi) = 0$. □

**Lemma 5.1.3.** *Let $\alpha$, $\theta \in [0,\pi]$. If $d = 2$, then*

$$A(\theta,\cos(\alpha)) := 1 - \frac{2}{\pi}\alpha + A^*(\theta,\cos(\alpha)),$$

*where*

$$A^*(\theta,\cos(\alpha)) := \begin{cases} 0, & 0 \le \alpha \le \frac{\theta_{ij}}{2},\\ \frac{2\alpha-\theta_{ij}}{2\pi}, & \frac{\theta_{ij}}{2} < \alpha < \pi - \frac{\theta_{ij}}{2},\\ \frac{2\alpha}{\pi} - 1, & \pi - \frac{\theta_{ij}}{2} \le \alpha \le \pi.\end{cases} \tag{5.12}$$

*Proof.* By definition of $A(\theta,\cos(\alpha))$, we have

$$A(\theta,\cos(\alpha)) = 1 - 2A(0,\cos(\alpha)) + A^*(\theta,\cos(\alpha)) = 1 - \frac{2\alpha}{\pi} + A^*(\theta,\cos(\alpha)),$$

where

$$A^*(\theta_{ij},\cos(\alpha)) = \frac{1}{2\pi}\int_{\Omega^1} 1_{\{\angle(\gamma,\mathbf{X}_i)\le\alpha,\,\angle(\gamma,\mathbf{X}_j)\le\alpha\}}\,\omega_{d-1}(\mathrm{d}\boldsymbol{\gamma}),$$

and simple geometric arguments give (5.12). □

**Proposition 5.1.4.** *Let $x \ge 0$. Then, $A(\theta,-x) = A(\theta,x) + 1 - 2F_{d-1}(x)$ and*

$$A(\theta,x) = \begin{cases} 2F_1(x) - 1 + \frac{1}{\pi}\left(\cos^{-1}(x) - \frac{\theta}{2}\right)_+, & d = 2,\\ 2\int_{-1}^{x} F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t), & d \ge 3.\end{cases}$$

*Alternatively, for $d \ge 3$, we have the expression*

$$A(\theta,x) = \begin{cases} \frac{1}{2} - \frac{\theta}{2\pi} + 2\int_0^x F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t), & 0 \le \theta < 2\cos^{-1}(x),\\ 2F_{d-1}(x) - 1, & 2\cos^{-1}(x) \le \theta \le \pi.\end{cases} \tag{5.13}$$

*Proof.* The equality $A(\theta,-x) = A(\theta,x) + 1 - 2F_{d-1}(x)$ follows from (5.8) and the symmetry of $\nu_{d-1}$. Assume $x \ge 0$. For $d = 2$, by Lemma 5.1.3

$$A(\theta, x) = 1 - \frac{2\cos^{-1}(x)}{\pi} + A^*(\theta, x),$$

where

$$A^*(\theta, x) = \begin{cases} 0, & 0 \leq \theta < 2\cos^{-1}(x), \\ \frac{2\cos^{-1}(x) - \theta}{2\pi}, & 2\cos^{-1}(x) \leq \theta \leq \pi. \end{cases}$$

If $d \geq 3$, by (5.8) we have that

$$\begin{aligned} A(\theta, 0) &= 2\int_{-1}^{0} F_{d-1}(t) \int_{-1}^{1} 1_{\{u \leq t/(1-t^2)^{1/2} \tan(\theta/2)\}} F_{d-2}(u) \, \mathrm{d}u \, \mathrm{d}t \\ &= 2\int_{-\cos(\theta/2)}^{0} F_{d-1}(t) F_{d-2}\left(\frac{t \tan(\theta/2)}{(1-t^2)^{1/2}}\right) \mathrm{d}t \\ &= \frac{1}{2} - \frac{\theta}{2\pi}, \end{aligned} \tag{5.14}$$

where (5.14) is due to Lemma 5.1.2. The result is deduced from (5.14), Lemma 5.1.2, and the following equality

$$A(\theta, x) = A(\theta, 0) + 2\int_{0}^{x} F_{d-1}(t) \int_{-1}^{1} 1_{\{u \leq t/(1-t^2)^{1/2} \tan(\theta/2)\}} F_{d-2}(u) \, \mathrm{d}u \, \mathrm{d}t,$$

taking into account that $t/(1-t^2)^{1/2} \tan(\theta/2) \leq 1$ when $t \in [0, \cos(\theta/2)]$ and $t/(1-t^2)^{1/2} \tan(\theta/2) \geq 1$ when $t \in [\cos(\theta/2), 1]$. □

Expression (5.7) is not computationally pleasant. For that reason, we provide next alternative forms for $P_{n,d-1}^W$ that expose its $U$-statistic nature. We do so for the fairly natural and general case in which $W$ is a positive *finite* measure on $[0,1]$, standardized to a probability measure on $[0,1]$ without loss of generality. We firstly write our statistic, from (5.7), as

$$P_{n,d-1}^W = \frac{1}{n} \sum_{i \neq j} \psi_{d-1}^W(\theta_{ij}) + \int_{-1}^{1} F_{d-1}(x)(1 - nF_{d-1}(x)) \, \mathrm{d}W(F_{d-1}(x)), \tag{5.15}$$

where

$$\psi_{d-1}^W(\theta) := \int_{-1}^{1} A(\theta, x) \, \mathrm{d}W(F_{d-1}(x)). \tag{5.16}$$

Notice that, since $A(0, x) = F_{d-1}(x)$, an equivalent expression to (5.15) is

$$P_{n,d-1}^W = \frac{1}{n}\sum_{i,j=1}^n \tilde{\psi}_{d-1}^W(\theta_{ij}), \text{ where } \tilde{\psi}_{d-1}^W(\theta) := \int_{-1}^1 \left(A(\theta,x) - F_{d-1}(x)^2\right) \mathrm{d}W(F_{d-1}(x)). \tag{5.17}$$

**Proposition 5.1.5.** *If $W$ is a cdf on $[0,1]$, then, for $\theta \in [0,\pi]$ and $d \geq 3$,*

$$\begin{aligned}
\psi_2^W(\theta) &= \frac{1}{2} - \frac{\theta}{2\pi} + 2\int_0^{\theta/(2\pi)} W(u)\,\mathrm{d}u,\\
\psi_{d-1}^W(\theta) &= -\frac{1}{2} + \frac{\theta}{2\pi} + 2\int_0^{1/2} W(u)\,\mathrm{d}u\\
&\quad + 4\int_0^{\cos(\theta/2)} W(F_{d-1}(t))\left(1 - F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\right)\mathrm{d}F_{d-1}(t).
\end{aligned}$$

*Proof.* From (5.15), we trivially have

$$P_{n,d-1}^W = \frac{2}{n}\sum_{i<j}\psi_{d-1}^W(\theta_{ij}) + \int_0^1 u(1-nu)\,\mathrm{d}W(u). \tag{5.18}$$

We separate the first addend for the cases $d = 2$ and $d \geq 3$. For $d = 2$, direct integration of $\int_0^\pi A(\theta,\cos(\alpha))\,\mathrm{d}W(F_1(\cos(\alpha)))$ using Lemma 5.1.3 gives

$$\begin{aligned}
\psi_1^w(\theta) &= -1 + 2\int_0^1 u\,\mathrm{d}W(u) + \int_{\theta/(2\pi)}^{1-\frac{\theta}{2\pi}} \left(1 - u - \tfrac{\theta}{2\pi}\right)\mathrm{d}W(u) + \int_0^{\frac{\theta}{2\pi}} (1-2u)\,\mathrm{d}W(u)\\
&= 1 - \int_{\frac{\theta}{2\pi}}^1 W(u)\,\mathrm{d}u - \int_{1-\frac{\theta}{2\pi}}^1 W(u)\,\mathrm{d}u,
\end{aligned}$$

where the second equality follows because $W$ is a cdf. The result is deduced because $W(1-t) = 1 - W(t)$ for $t \in [0,1]$ and $\int_0^1 W(u)\,\mathrm{d}u = 1/2$. For $d \geq 3$, simple expressions for $A(\theta_{ij},x)$ are not easy to obtain. We follow thus an alternative approach. First, note that by symmetry

$$\begin{aligned}
&\omega_{d-1}\left(\{\boldsymbol{\gamma} \in \Omega^{d-1} : \boldsymbol{\gamma}'\mathbf{X}_i \leq x, \boldsymbol{\gamma}'\mathbf{X}_j \leq x\}\right)\\
&\quad = \omega_{d-1}\left(\{\boldsymbol{\gamma} \in \Omega^{d-1} : \boldsymbol{\gamma}'\mathbf{X}_i \geq \boldsymbol{\gamma}'\mathbf{X}_j, \boldsymbol{\gamma}'\mathbf{X}_i \leq x\} \cup \{\boldsymbol{\gamma} \in \Omega^{d-1} : \boldsymbol{\gamma}'\mathbf{X}_j \geq \boldsymbol{\gamma}'\mathbf{X}_i, \boldsymbol{\gamma}'\mathbf{X}_j \leq x\}\right)\\
&\quad = 2\omega_{d-1}\left(\{\boldsymbol{\gamma} \in \Omega^{d-1} : \boldsymbol{\gamma}'\mathbf{X}_i \geq \boldsymbol{\gamma}'\mathbf{X}_j, \boldsymbol{\gamma}'\mathbf{X}_i \leq x\}\right)
\end{aligned}$$

and hence

$$\int_{-1}^1 A(\theta_{ij},x)\,\mathrm{d}W(F_{d-1}(x)) = \frac{2}{\omega_{d-1}}\int_{-1}^1\int_{\Omega^{d-1}} 1_{\{\boldsymbol{\gamma}'\mathbf{X}_i \geq \boldsymbol{\gamma}'\mathbf{X}_j, \boldsymbol{\gamma}'\mathbf{X}_i \leq x\}}\,\omega_{d-1}(\mathrm{d}\boldsymbol{\gamma})\,\mathrm{d}W(F_{d-1}(x)). \tag{5.19}$$

Consider now the two successive tangent-normal decompositions:

$$\begin{cases} \boldsymbol{\gamma} = t\mathbf{X}_i + (1-t^2)^{1/2}\mathbf{B}_{\mathbf{X}_i,d}\boldsymbol{\xi}, \\ \omega_{d-1}(\mathrm{d}\boldsymbol{\gamma}) = (1-t^2)^{(d-3)/2}\,\mathrm{d}t\,\omega_{d-2}(\mathrm{d}\boldsymbol{\xi}), \end{cases} \quad t \in [-1,1],\ \boldsymbol{\xi} \in \Omega^{d-2} \tag{5.20}$$

and

$$\begin{cases} \boldsymbol{\xi} = u\boldsymbol{\eta}_{ij} + (1-u^2)^{1/2}\mathbf{B}_{\boldsymbol{\eta}_{ij},d-1}\boldsymbol{\zeta}, \\ \omega_{d-2}(\mathrm{d}\boldsymbol{\xi}) = (1-u^2)^{(d-4)/2}\,\mathrm{d}u\,\omega_{d-3}(\mathrm{d}\boldsymbol{\zeta}), \end{cases} \quad u \in [-1,1],\ \boldsymbol{\zeta} \in \Omega^{d-3}, \tag{5.21}$$

where $\mathbf{B}_{\mathbf{x},d}$ denotes a semi-orthonormal matrix $d \times (d-1)$ such that $\mathbf{B}_{\mathbf{x},d}\mathbf{B}'_{\mathbf{x},d} = \mathbf{I}_d - \mathbf{x}\mathbf{x}'$ and $\mathbf{B}'_{\mathbf{x},d}\mathbf{B}_{\mathbf{x},d} = \mathbf{I}_{d-1}$ for $\mathbf{x} \in \Omega^{d-1}$, and, $\boldsymbol{\eta}_{ij} := \mathbf{B}'_{\mathbf{X}_i,d}\mathbf{X}_j(1-(\mathbf{X}_i'\mathbf{X}_j)^2)^{-1/2} \in \Omega^{d-2}$. Plugging-in (5.20) and (5.21) in (5.19) gives

$$\begin{aligned} &\int_{-1}^{1} A(\theta_{ij}, x)\,\mathrm{d}W(F_{d-1}(x)) \\ &\quad = 2\frac{\omega_{d-3}}{\omega_{d-1}} \int_{-1}^{1}\int_{-1}^{1}\left\{\int_{-1}^{1} 1_{\{t \geq t\cos(\theta_{ij}) + u(1-t^2)^{1/2}\sin(\theta_{ij}), t \leq x\}}\,\mathrm{d}W(F_{d-1}(x))\right\} \\ &\qquad \times (1-t^2)^{\frac{d-3}{2}}(1-u^2)^{\frac{d-4}{2}}\,\mathrm{d}t\,\mathrm{d}u, \\ &\quad = 2\int_{-1}^{1} H_{d-1}(t)\left\{\int_{-1}^{1} 1_{\{t\cos(\theta_{ij}) + u(1-t^2)^{1/2}\sin(\theta_{ij}) \leq t\}}\,\mathrm{d}F_{d-2}(u)\right\}\mathrm{d}F_{d-1}(t) \\ &\quad =: \psi_{d-1}^{W}(\theta_{ij}) \end{aligned} \tag{5.22}$$

since $\int_t^1 \mathrm{d}W(F_{d-1}(x)) = 1 - W(F_{d-1}(t)) =: H_{d-1}(t)$. Now, for $\theta \in [0,\pi]$ and $t, u \in [-1,1]$,

$$t\cos(\theta) + u(1-t^2)^{1/2}\sin(\theta) \leq t \iff u \leq \frac{t\tan\left(\frac{\theta}{2}\right)}{(1-t^2)^{1/2}} =: u(\theta, t).$$

Note that it can occur that $|u(\theta,t)| > 1$, so in the integral limits of (5.22) we rather handle

$$\tilde{u}(\theta,t) := ((u(\theta,t) \vee -1) \wedge 1) = \begin{cases} 1, & \cos\left(\frac{\theta}{2}\right) \leq t \leq 1, \\ u(\theta,t), & -\cos\left(\frac{\theta}{2}\right) < t < \cos\left(\frac{\theta}{2}\right), \\ -1, & -1 \leq t \leq -\cos\left(\frac{\theta}{2}\right). \end{cases}$$

Therefore, denoting $\tau_\theta = \cos(\theta/2)$, (5.22) becomes

$$\begin{aligned} \psi_{d-1}^{W}(\theta) &= 2\int_{-1}^{1} H_{d-1}(t)\left\{\int_{-1}^{\tilde{u}(\theta,t)} \mathrm{d}F_{d-2}(u)\right\}\mathrm{d}F_{d-1}(t) \\ &= 2\int_{\tau_\theta}^{1} H_{d-1}(t)\int_{-1}^{1} \mathrm{d}F_{d-2}(u)\,\mathrm{d}F_{d-1}(t) + 2\int_{-\tau_\theta}^{\tau_\theta} H_{d-1}(t)\int_{-1}^{u(\theta,t)} \mathrm{d}F_{d-2}(u)\,\mathrm{d}F_{d-1}(t) \end{aligned}$$

$$= 2\int_{\tau_\theta}^{1} H_{d-1}(t)\,\mathrm{d}F_{d-1}(t) + 2\int_{-\tau_\theta}^{\tau_\theta} H_{d-1}(t) F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t). \tag{5.23}$$

The proposition is proved by applying Lemma 5.1.2 to the second integral and recalling that $W(1-t) = 1 - W(t)$ for $t \in [0,1]$ and that $\int_0^1 W(u)\,\mathrm{d}u = 1/2$. □

The above exposed ideas open the way for generating new uniformity tests based on particular choices of $W$. Among the many $W$-specific instances of $P_{n,d-1}^W$ that could be constructed, we opt to investigate in-depth in the following subsections some weights that deliver new tests that relate and extend previous proposals of uniformity tests.

### 5.1.3 Extending the Watson test

One of the simplest measures that can be considered in Proposition 5.1.5 is given by the cdf $W(x) = x$, $x \in [0,1]$, which is the associated to the Cramér–von Mises (CvM) weight. As seen in this section, $P_{n,d-1}^W$ yields the celebrated Watson [135] test of circular uniformity when $d = 2$ and connects it with the chordal-based uniformity test by Bakshaev [11] on $\Omega^2$. Therefore, the test statistic $P_{n,d-1}^W$ can be seen as a generalization of the former to $\Omega^d$, $d \geq 3$.

**Proposition 5.1.6** (An extension of the Watson test)**.** *Consider the CvM cdf* $W(x) = x$, $x \in [0,1]$*. Then,*

$$P_{n,d-1}^{\mathrm{CvM}} := \frac{2}{n}\sum_{i<j}\psi_{d-1}^{\mathrm{CvM}}(\theta_{ij}) + \frac{3-2n}{6},$$

*where, for* $\theta \in [0,\pi]$*,*

$$\psi_{d-1}^{\mathrm{CvM}}(\theta) := \begin{cases} \frac{1}{2} + \frac{\theta}{2\pi}\left(\frac{\theta}{2\pi} - 1\right), & d = 2,\\ \frac{1}{2} - \frac{1}{4}\sin\left(\frac{\theta}{2}\right), & d = 3,\\ \psi_1^{\mathrm{CvM}}(\theta) + \frac{1}{4\pi^2}\left((\pi - \theta)\tan\left(\frac{\theta}{2}\right) - 2\sin^2\left(\frac{\theta}{2}\right)\right), & d = 4, \end{cases}$$

*and, if* $d \geq 3$*,*

$$\psi_{d-1}^{\mathrm{CvM}}(\theta) = -\frac{3}{4} + \frac{\theta}{2\pi} + 2F_{d-1}^2\left(\cos\left(\frac{\theta}{2}\right)\right) - 4\int_0^{\cos(\theta/2)} F_{d-1}(t) F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t).$$

*Proof.* The second term in (5.18) and the expression for $\psi_1^{\mathrm{CvM}}$ follow trivially. For $\psi_{d-1}^{\mathrm{CvM}}$, $d \geq 3$, we can write

$$\psi_{d-1}^{\mathrm{CvM}}(\theta) = -\frac{1}{4} + \frac{\theta}{2\pi} + 4\int_0^{\cos(\theta/2)} F_{d-1}(t)\left(1 - F_{d-2}\left(\frac{t\tan\left(\frac{\theta}{2}\right)}{(1-t^2)^{1/2}}\right)\right)\mathrm{d}F_{d-1}(t),$$

from which the desired expression of $\psi_{d-1}^{\mathrm{CvM}}$ for $d \geq 5$. For the case in which $d = 3$:

$$\begin{aligned}\psi_2^{\mathrm{CvM}}(\theta) &= -\frac{1}{4} + \frac{\theta}{2\pi} + \frac{\tan^2\left(\frac{\theta}{2}\right)}{\pi}\int_0^1 \cos^{-1}(y)\left(\frac{y}{\left(y^2+\tan^2\left(\frac{\theta}{2}\right)\right)^{1/2}} + 1\right)\frac{1}{\left(y^2+\tan^2\left(\frac{\theta}{2}\right)\right)^{3/2}}\,\mathrm{d}y\\ &= -\frac{1}{4} + \frac{\theta}{2\pi} + \frac{1}{4\pi}\left(\pi - \pi\tan\left(\frac{\theta}{2}\right)\frac{1}{\left(1+\tan^2\left(\frac{\theta}{2}\right)\right)^{1/2}} + 4\tan^{-1}\left(\frac{1}{\tan\left(\frac{\theta}{2}\right)}\right)\right)\\ &= \frac{1}{2} - \frac{1}{4}\sin\left(\frac{\theta}{2}\right).\end{aligned}$$

If $d = 4$, then

$$\begin{aligned}\psi_3^{\mathrm{CvM}}(\theta) = &-\frac{3}{4} + \frac{\theta}{2\pi} + 2F_3^2\left(\cos\left(\frac{\theta}{2}\right)\right) - 2\int_0^{\cos(\theta/2)} F_3(t)\left(1 + \frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_3(t)\\ = &-\frac{1}{2} + \frac{\theta}{2\pi} + F_3^2\left(\cos\left(\frac{\theta}{2}\right)\right) - \frac{2}{\pi}\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right)\\ &-\frac{4}{\pi^2}\tan\left(\frac{\theta}{2}\right)\int_0^{\cos(\theta/2)}\left(t^2(1-t^2)^{1/2} - t\cos^{-1}(t)\right)\mathrm{d}t,\end{aligned} \tag{5.24}$$

with

$$\begin{aligned}&\int_0^{\cos(\theta/2)}\left(t^2(1-t^2)^{1/2} - t\cos^{-1}(t)\right)\mathrm{d}t\\ &\quad= \frac{1}{4}\left\{\cos^3\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right) - \theta\cos^2\left(\frac{\theta}{2}\right) + \frac{1}{2}\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right) - \frac{1}{2}\sin^{-1}\left(\cos\left(\frac{\theta}{2}\right)\right)\right\}.\end{aligned} \tag{5.25}$$

Since $F_3(\cos(\theta/2)) = 1 + (\cos(\theta/2)\sin(\theta/2) - \theta/2)/\pi$ due to (2.5), the third term in (5.24) results

$$F_3^2\left(\cos\left(\frac{\theta}{2}\right)\right) = 1 + \frac{\theta^2}{4\pi^2} - \frac{\theta}{\pi} + \frac{1}{\pi}(2-\theta)\cos\left(\frac{\theta}{2}\right)\sin\left(\frac{\theta}{2}\right) + \frac{1}{\pi^2}\cos^2\left(\frac{\theta}{2}\right)\sin^2\left(\frac{\theta}{2}\right). \tag{5.26}$$

The expression for $\psi_3^{\mathrm{CvM}}$ arises from combining (5.24), (5.25), and (5.26). □

**Remark 5.1.7.** *The test based on $P_{n,1}^{\mathrm{CvM}}$ is equivalent to the Watson [135] test. Precisely, $P_{n,1}^{\mathrm{CvM}} = \frac{1}{2}U_n^2$, where $U_n^2$ is defined in* (2.16) *in Subsection* 2.6.1*. The relation of $P_{n,1}^{\mathrm{CvM}}$ and $U_n^2$ stems from Proposition* 5.1.6 *and* (2.17)*. The connection has two main implications:*

*(a) It introduces an interpretation of $P^{\mathrm{CvM}}_{n,d-1}$ as a natural extension of the well-known $U^2_n$ to an arbitrary dimension $d$.*

*(b) It adds yet another interesting connection between $U^2_n$ and the CvM test.*

**Remark 5.1.8.** *From the definition of Bakshaev test* (2.20) *in Subsection 2.6.1, it is evident that $P^{\mathrm{CvM}}_{n,2} = \frac{1}{8} N_{n,2}$.*

**Remark 5.1.9.** *We highlight a perhaps intriguing behaviour of $P^{\mathrm{CvM}}_{n,d-1}$: it simultaneously yields as particular cases the Watson test for $d = 2$ and the test by Bakshaev [11] for $d = 3$, despite the test by Bakshaev [11] for $d = 2$ being actually different from the Watson test. This phenomenon is explained by the dimension-dependence of $\psi^W_{d-1}$, a distinctive feature of $P^W_{n,d-1}$ that naturally arises from its construction, and that sharply contrasts with the dimension-independent kernels (that handle $\theta_{ij}$ equally for any dimension $d$) of other uniformity tests, such as Bakshaev [11]'s $N_{n,d}$. As a consequence, $P^W_{n,d-1}$ allows the extension of circular uniformity tests in a dimension-dependent manner.*

### 5.1.4 Extending the Ajne and Rothman test

We turn now to the consideration of a discrete measure $W$ that yields as particular cases the circular uniformity tests by Ajne [4] and Rothman [117], and that delivers extensions of them to $\Omega^{d-1}$, $d \geq 3$. This extension of the Rothman test, given in Proposition 5.1.12, is new, meanwhile that of the Ajne test coincides with that proposed in Prentice [114].

The following result provides a computationally amenable form for (2.18) in Subsection 2.6.1, in the spirit of (2.19), that is required for its comparison with (5.15).

**Proposition 5.1.10** (Computation of the Rothman test)**.** *Let $t_m := \min(t, 1-t)$ for $t \in (0,1)$. The test statistic* (2.18) *can be expressed as*

$$R_{n,t} = t(1-t) + \frac{2}{n}\sum_{i<j} h_t(\theta_{ij}), \text{ where } \quad h_t(\theta) := \left(t_m - \tfrac{\theta}{2\pi}\right)_+ - t_m^2. \tag{5.27}$$

*Proof.* The statistic $R_{n,t}$ admits the representation (2.7) for a certain sequence $\{v_{k,d-1}\}$. Following the arguments in Watson [136], Rothman [117] considered the Fourier expansion of $N(\alpha,t) - nt$ that, adapted to a circular sample $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$, is given by $N(\alpha,t) - nt = \sum_{k=1}^{\infty} a_k \cos(k\alpha) + b_k \sin(k\alpha)$, where

$$a_k = \frac{1}{\pi k}\sum_{i=1}^{n}[\sin(k\Theta_i) - \sin(k\Theta_i - 2\pi kt)], \quad b_k = \frac{1}{\pi k}\sum_{i=1}^{n}[\cos(k\Theta_i - 2\pi kt) - \cos(k\Theta_i)].$$

From the Fourier expansion it readily follows that

$$R_{n,t} = \frac{1}{2n}\sum_{k=1}^{\infty}\left(a_k^2 + b_k^2\right). \tag{5.28}$$

Expanding the squares of (5.28) and using the cosine addition formula gives

$$\begin{aligned} R_{n,t} &= \frac{1}{2n}\sum_{i,j=1}^{n}\sum_{k=1}^{\infty}\frac{1}{(\pi k)^2}\{2\cos(k(\Theta_i - \Theta_j)) - \cos(k(\Theta_i - \Theta_j) + 2\pi k t) \\ &\qquad - \cos(k(\Theta_j - \Theta_i) + 2\pi k t)\} \\ &= \frac{2}{n}\sum_{i,j=1}^{n}\sum_{k=1}^{\infty}\frac{\sin^2(k\pi t)}{(\pi k)^2}\cos(k(\Theta_i - \Theta_j)). \end{aligned}$$

where the last equality follows from the basic trigonometric identities

$$\cos(x+y) + \cos(x-y) = 2\cos(x)\cos(y) \quad \text{and} \quad 1 - \cos(2k\pi t) = 2\sin^2(k\pi t).$$

As a consequence, $v_{k,1} = \sin(k\pi t)/(\pi k)$ and $R_{n,t} = \frac{1}{n}\sum_{i,j=1}^{n} h_t(\theta_{ij})$ with

$$h_t(\theta) = 2\sum_{k=1}^{\infty}\frac{\sin^2(k\pi t)}{(\pi k)^2}\cos(k\theta) = \tilde{h}(\theta) - \tfrac{1}{2}\tilde{h}(\theta - 2\pi t) - \tfrac{1}{2}\tilde{h}(\theta + 2\pi t), \quad \theta \in [0,\pi],$$

where $\tilde{h}(x) := \sum_{k=1}^{\infty}\frac{1}{k^2}\cos(kx) = \frac{1}{4}\{(x \mod 2\pi - \pi)^2 - \frac{\pi^2}{3}\}$, for $x \in \mathbb{R}$. A case-by-case analysis of the possible values of $\theta \in [0,\pi]$ and $t \in (0,1)$ gives

$$h_t(\theta) = \begin{cases} \frac{1}{2\pi}(2\pi t(1-t) - \theta), & \text{if } \theta \in [0, 2\pi t_m) \\ -t_m^2, & \text{if } \theta \in [2\pi t_m, \pi) \end{cases} = -\min\left(\tfrac{\theta}{2\pi} - t(1-t), t_m^2\right),$$

which immediately provides (5.27). □

**Remark 5.1.11.** *Somehow imprecise computational recipes for* (2.18) *seem to have proliferated in the literature. Rothman [117] provided in his equation (35) a computational form for $R_{n,t}$, but relying on an undefined notation and without hints to its derivation. Employing his equation (8) produces a factor 2 difference with respect to the statistic defined in his equation (1), since (8) misses $\frac{1}{2} = \int_0^1 \cos^2(2k\pi x)\,\mathrm{d}x = \int_0^1 \sin^2(2k\pi x)\,\mathrm{d}x$ (that does appear in our equation* (5.28)*). Differently, equation (6.3.63) in [102] states that the sequence $\{v_{k,1}\}$ of $R_{n,t}$ (see Section 2.6.1) is $\sin(k\pi t)/(2k\pi t)$, which does not correspond to the statistic as defined in their equation (6.3.50) (equal to our equation* (2.18)*). These two misprints introduce new statistics that are proportional to the original definition of $R_{n,t}$ and thus yield the same test decision. However, they may induce spurious test outcomes if the asymptotic distribution of one version is employed with the statistic of another.*

**Proposition 5.1.12** (An extension of the Rothman test)**.** *Consider the cdf*

$$W_t(x) := \frac{1}{2}\left(1_{\{t_m \leq x\}} + 1_{\{1-t_m \leq x\}}\right),$$

*with* $x \in [-1,1]$ *and where* $t_m$ *is defined in Proposition 5.1.10. Then,*

$$P_{n,d-1}^{\mathrm{R}_t} = \frac{2}{n}\sum_{i<j}\psi_{d-1}^{\mathrm{R}_t}(\theta_{ij}) + \frac{1-n}{2} + nt(1-t),$$

*where, for* $\theta \in [0,\pi]$ *and* $d = 2$*,*

$$\psi_1^{\mathrm{R}_t}(\theta) = h_t(\theta) + \frac{1}{2} - t(1-t),$$

*if* $d = 3, 4$*,*

$$\psi_2^{\mathrm{R}_t}(\theta) = \begin{cases} -t_m + \frac{1}{2} - \frac{1-2t_m}{\pi}\cos^{-1}\left(\frac{(1/2-t_m)\tan(\theta/2)}{(t_m(1-t_m))^{1/2}}\right) & \\ \quad +\frac{1}{\pi}\tan^{-1}\left(\frac{\left(\cos^2(\theta/2)-(1-2t_m)^2\right)^{1/2}}{\sin(\theta/2)}\right), & \theta < \theta_{t_m}, \\ 1/2 - t_m, & \theta \geq \theta_{t_m}, \end{cases}$$

$$\psi_3^{\mathrm{R}_t}(\theta) = \begin{cases} \frac{1}{2} + t_m - \frac{\theta+\theta_{t_m}}{2\pi} + \frac{1}{\pi}\left\{\frac{\sin(\theta_{t_m})}{2} + \tan\left(\frac{\theta}{2}\right)\cos^2\left(\frac{\theta_{t_m}}{2}\right)\right\}, & \theta < \theta_{t_m}, \\ 1/2 - t_m, & \theta \geq \theta_{t_m}, \end{cases}$$

*where* $\theta_{t_m} := 2\cos^{-1}\left(F_{d-1}^{-1}(1-t_m)\right)$*, and, if* $d \geq 3$*,*

$$\psi_{d-1}^{\mathrm{R}_t}(\theta) = \begin{cases} t_m - \frac{\theta}{2\pi} + 2\int_0^{\cos(\theta_{t_m}/2)} F_{d-2}\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_{d-1}(u), & \theta < \theta_{t_m}, \\ 1/2 - t_m, & \theta \geq \theta_{t_m}. \end{cases}$$

*Proof.* The second addend in (5.18) is

$$\int_0^1 u(1-nu)\,\mathrm{d}W_t(u) = \frac{1}{2}(1-n) + nt(1-t). \tag{5.29}$$

Denote $\bar{\theta} := \theta/(2\pi)$. If $d = 2$, the first addend in (5.18) is

$$\begin{aligned} \psi_1^{\mathrm{R}_t}(\theta) =& \frac{1}{2} - \bar{\theta} + \int_0^{\bar{\theta}} \left(1_{\{t_m \leq u\}} + 1_{\{1-t_m \leq u\}}\right)\mathrm{d}u \\ =& \frac{1}{2} - \bar{\theta} + \left(\bar{\theta} - t_m\right)_+ + \left(\bar{\theta} - (1-t_m)\right)_+ \\ =& \frac{1}{2} - \bar{\theta} + \left(\bar{\theta} - t_m\right)_+. \end{aligned} \tag{5.30}$$

The proof for $d = 2$ readily follows from (5.29) and (5.30). For $d \geq 3$, by Proposition 5.1.5,

$$\begin{aligned}\psi_{d-1}^{\mathrm{R}_t}(\theta) &= -\frac{1}{2} + \bar{\theta} + 2\int_0^{1/2} W_t(u)\,\mathrm{d}u\\ &\quad + 4\int_0^{\cos(\theta/2)} W_t(F_{d-1}(u))\left(1 - F_{d-2}\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\right)\mathrm{d}F_{d-1}(u)\\ &= -t_m + \frac{1}{2} + 2\left(F_{d-1}(\cos(\theta/2)) - 1 + t_m\right)_+ - 2F_{d-1}(\cos(\theta/2)) - \bar{\theta} + \frac{3}{2}\\ &\quad + 2\int_0^{\cos(\theta/2)\wedge F_{d-1}^{-1}(1-t_m)} F_{d-2}\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_{d-1}(u), \end{aligned} \tag{5.31}$$

where the last equality follows from Lemma 5.1.2. If $\cos(\theta/2) < F_{d-1}^{-1}(1 - t_m)$, from (5.31) and using Lemma 5.1.2, we have

$$\psi_{d-1}^{\mathrm{R}_t}(\theta) = -t_m + \frac{1}{2} - 2F_{d-1}(\cos(\theta/2)) - \bar{\theta} + \frac{3}{2} + 2F_{d-1}(\cos(\theta/2)) + \bar{\theta} - \frac{3}{2} = \frac{1}{2} - t_m.$$

If $\cos(\theta/2) > F_{d-1}^{-1}(1 - t_m)$, from (5.31), we have

$$\begin{aligned}\psi_{d-1}^{\mathrm{R}_t}(\theta) &= t_m + 2 + 2F_{d-1}\left(\cos(\theta/2)\right) - 2 - 2F_{d-1}\left(\cos(\theta/2)\right) - \bar{\theta}\\ &\quad + 2\int_0^{F_{d-1}^{-1}(1-t_m)} F_{d-2}\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_{d-1}(u)\\ &= t_m - \frac{3}{2} + 2F_{d-1}\left(\cos(\theta/2)\right) - 2\int_{F_{d-1}^{-1}(1-t_m)}^{\cos(\theta/2)} F_{d-2}\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_{d-1}(u),\end{aligned}$$

where the last equality follows from Lemma 5.1.2.

Take now $d = 3$. Then $F_2^{-1}(1 - t_m) = 1 - 2t_m$ and, in addition,

$$\begin{aligned}2\int_0^{F_2^{-1}(1-t_m)} F_1\left(\tfrac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_2(u) &= 1 - 2t_m - \cos^{-1}\left(\tfrac{(1-2t_m)\tan(\theta/2)}{2(t_m(1-t_m))^{1/2}}\right)\tfrac{(1-2t_m)}{\pi}\\ &\quad + \frac{1}{\pi}\tan^{-1}\left(\tfrac{\left(\cos^2\left(\frac{\theta}{2}\right)-(1-2t_m)^2\right)^{1/2}}{\sin\left(\frac{\theta}{2}\right)}\right) + \tfrac{\theta-\pi}{2\pi}.\end{aligned}$$

Hence, if $\cos(\theta/2) > F_2^{-1}(1 - t_m)$, from (5.31), we have

$$\begin{aligned}\psi_2^{\mathrm{R}_t}(\theta) &= t_m + \pi + \left(1 + \frac{1}{2\pi}\right)\theta + 2(1 - 2t_m)\cos^{-1}\left(\frac{(1/2 - t_m)\tan\left(\frac{\theta}{2}\right)}{(t_m(1 - t_m))^{1/2}}\right)(1 - 2t_m)\\ &\quad - 2\tan^{-1}\left(\frac{\left(\cos^2\left(\frac{\theta}{2}\right) - (1 - 2t_m)^2\right)^{1/2}}{\sin\left(\frac{\theta}{2}\right)}\right).\end{aligned}$$

If $d = 4$, note that

$$
\begin{aligned}
2\int_0^{F_3^{-1}(1-t_m)} F_2\left(\frac{u\tan(\theta/2)}{(1-u^2)^{1/2}}\right)\mathrm{d}F_3(u) &= \frac{2}{\pi}\int_0^{F_3^{-1}(1-t_m)}\left(\tan\left(\tfrac{\theta}{2}\right)u+(1-u^2)^{1/2}\right)\mathrm{d}u\\
&= \frac{\tan\left(\frac{\theta}{2}\right)}{\pi}\left(F_3^{-1}(1-t_m)\right)^2+\frac{1}{\pi}\sin^{-1}\left(F_3^{-1}(1-t_m)\right)\\
&\quad+\frac{1}{\pi}F_3^{-1}(1-t_m)\left(1-\left(F_3^{-1}(1-t_m)\right)^2\right)^{1/2}.
\end{aligned}
$$

Then, if $\cos(\theta/2) > F_3^{-1}(1-t_m)$, from (5.31), we have

$$
\begin{aligned}
\psi_3^{\mathrm{R}_t}(\theta) = t_m - \frac{\theta}{2\pi} + \frac{\tan(\theta/2)}{\pi}\left(F_3^{-1}(1-t_m)\right)^2+\frac{1}{\pi}\sin^{-1}\left(F_3^{-1}(1-t_m)\right)\\
+\frac{1}{\pi}F_3^{-1}(1-t_m)\left(1-\left(F_3^{-1}(1-t_m)\right)^2\right)^{1/2}.
\end{aligned}
$$

The proof ends by taking into account that $F_3^{-1}(1-t_m) = \cos(\theta_{t_m}/2)$ by (2.5) and the definition of $\theta_{t_m}$. □

**Remark 5.1.13.** *For $d=2$, $P_{n,1}^{\mathrm{R}_t}$ equals the Rothman test statistic, $P_{n,1}^{\mathrm{R}_t} = R_{n,t}$. Therefore, $P_{n,d-1}^{\mathrm{R}_t}$ can be regarded as an extension of the Rothman statistic to $\Omega^d$, $d\geq 2$. Furthermore, the extension is coherent with Prentice [114]'s extension of Ajne [4]'s $A_n$ to $\Omega^d$ for $d \geq 2$, given by the test statistic $A_{n,d-1} := n/4 - 1/(n\pi)\sum_{i<j}\theta_{ij}$. Indeed, considering $t = 1/2$, by Lemma 5.1.2, we obtain that $P_{n,d-1}^{\mathrm{R}_{1/2}}$ has the dimension-independent kernel $\psi_{d-1}(\theta) = 1/2-\theta/(2\pi)$, $d\geq 2$, and therefore the tests based on $P_{n,d-1}^{\mathrm{R}_{1/2}}$ and $A_{n,d-1}$ are equivalent.*

**Remark 5.1.14.** *The statistic $P_{n,d-1}^{\mathrm{R}_t}$ only depends on $t_m$, hence $P_{n,d-1}^{\mathrm{R}_t} = P_{n,d-1}^{\mathrm{R}_{1-t}}$, $t\in(0,1)$. Also, since $F_{n,\gamma}(F^{-1}(1-t_m)) = F_{n,\gamma}(-F^{-1}(t_m)) = 1-F_{n,-\gamma}(F^{-1}(t_m)^-)$, we have that*

$$
\begin{aligned}
P_{n,d-1}^{\mathrm{R}_t} &= n\int_{\Omega^{d-1}}\left(F_{n,\gamma}(F_{d-1}^{-1}(t_m))-t_m\right)^2\nu_{d-1}(\mathrm{d}\gamma)\\
&= \frac{1}{n}\int_{\Omega^{d-1}}\left(N_{d-1}\left(\gamma,\cos^{-1}(F_{d-1}^{-1}(1-t_m))\right)-nt_m\right)^2\nu_{d-1}(\mathrm{d}\gamma),
\end{aligned}
$$

*where $N_{d-1}(\gamma,\theta) := \#\{\mathbf{X}_1,\ldots,\mathbf{X}_n : \mathbf{X}_i\in\mathcal{C}_{d-1}(\gamma,\theta)\}$ and $\mathcal{C}_{d-1}(\gamma,\theta)$ is the hyperspherical cap centred at $\gamma$ and with solid angle $\theta\in[0,\pi]$. Therefore, geometrically $P_{n,d-1}^{\mathrm{R}_t}$ is comparing the number of observed and expected points in $\mathcal{C}_{d-1}\left(\gamma,\cos^{-1}(F_{d-1}^{-1}(1-t_m))\right)$ under $\mathbf{H}_0$, for all $\gamma\in\Omega^{d-1}$, and hence its connection with (2.18) is evident.*

### 5.1.5 An Anderson–Darling-like test

We introduce now a new test based on the Anderson–Darling weight. Contrary to the Kolmogorov–Smirnov and CvM tests, up to our knowledge, this is the first adaptation of the celebrated Anderson and Darling [8] test to deal with directional data, despite the test being one of the three most well-known ecdf-based goodness-of-fit tests.

**Proposition 5.1.15** (An Anderson–Darling test). *If* $w(u) = 1/(u(1-u))$. *Then,*

$$P_{n,d-1}^{\mathrm{AD}} = \frac{2}{n}\sum_{i<j}\psi_{d-1}^{\mathrm{AD}}(\theta_{ij}) + n,$$

*where, for* $\theta \in (0, \pi]$,

$$\psi_{d-1}^{\mathrm{AD}}(\theta) = \begin{cases} -2\log(2\pi) + \frac{1}{\pi}\left\{\theta\log(\theta) + (2\pi-\theta)\log(2\pi-\theta)\right\}, & d=2,\\ -\log(4) + \frac{2}{\pi}\int_0^{\cos(\theta/2)}\log\left(\frac{1+t}{1-t}\right)\cos^{-1}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}t, & d=3,\\ -2\log(2\pi) + \frac{1}{\pi}\left\{s(\theta)\log(s(\theta)) + (2\pi - s(\theta))\log(2\pi - s(\theta))\right\} & \\ \quad -\frac{4}{\pi}\tan\left(\frac{\theta}{2}\right)\int_0^{\cos(\theta/2)} t\log\left(\frac{\pi}{\cos^{-1}(t) - t(1-t^2)^{1/2}} - 1\right)\mathrm{d}t, & d=4, \end{cases}$$

*with* $s(\theta) := \theta - \sin(\theta)$*, and, if* $d \geq 3$,

$$\psi_{d-1}^{\mathrm{AD}}(\theta) = -\log(4) + 4\int_0^{\cos(\theta/2)}\log\left(\frac{F_{d-1}(t)}{1-F_{d-1}(t)}\right)\left(1 - F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\right)\mathrm{d}F_{d-1}(t). \tag{5.32}$$

*Proof.* If $d = 2$, then by (5.7)

$$\begin{aligned} P_{n,1}^{\mathrm{AD}} &= \frac{1}{\pi}\int_0^{\pi}\left\{\frac{1}{n}\sum_{i\neq j}A(\theta_{ij},\cos(\alpha)) + F_1(\cos(\alpha))(1 - nF_1(\cos(\alpha)))\right\}w(F_1(\cos(\alpha)))\,\mathrm{d}\alpha \\ &= \frac{2}{n}\sum_{i<j}\frac{1}{\pi}\int_0^{\pi}\left(\frac{A(\theta_{ij},\cos(\alpha))}{F_1(\cos(\alpha))} + \frac{1}{n-1}(1 - nF_1(\cos(\alpha)))\right)\frac{1}{1-F_1(\cos(\alpha))}\,\mathrm{d}\alpha. \end{aligned}$$

From $F_1(\cos(\alpha)) = 1 - \alpha/\pi$, we have that

$$\begin{aligned} P_{n,1}^{\mathrm{AD}} &= \frac{2}{n}\sum_{i<j}\int_0^{\pi}\left(\frac{\pi}{\alpha(\pi-\alpha)}A(\theta_{ij},\cos(\alpha)) + \frac{1-n}{(n-1)\alpha} + \frac{n}{(n-1)\pi}\right)\mathrm{d}\alpha \\ &= n + \frac{2}{n}\sum_{i<j}\int_0^{\pi}\left(\frac{\pi}{\alpha(\pi-\alpha)}A(\theta_{ij},\cos(\alpha)) - \frac{1}{\alpha}\right)\mathrm{d}\alpha. \end{aligned} \tag{5.33}$$

Since $\alpha \mapsto 1/(\alpha(\pi-\alpha))$ and $\alpha \mapsto 1/\alpha$ are not integrable on $[0,\pi]$, we first compute the sum of the integrand. By Lemma 5.1.3, we have to consider three cases depending on the value of $\alpha$:

$$\frac{\pi}{\alpha(\pi-\alpha)}A(\theta_{ij},\cos(\alpha)) - \frac{1}{\alpha} = \begin{cases} -\frac{1}{\pi-\alpha}, & 0 \leq \alpha \leq \frac{\theta_{ij}}{2},\\ -\frac{\theta_{ij}}{2\alpha(\pi-\alpha)}, & \frac{\theta_{ij}}{2} < \alpha < \pi - \frac{\theta_{ij}}{2},\\ -\frac{1}{\alpha}, & \pi - \frac{\theta_{ij}}{2} \leq \alpha \leq \pi. \end{cases}$$

Consequently, from (5.33) we have that

$$\begin{aligned}P_{n,1}^{\mathrm{AD}} &= n + \frac{2}{n}\sum_{i<j}\left\{\log\left(\pi - \frac{\theta_{ij}}{2}\right) - \log(\pi) - \frac{\theta_{ij}}{2\pi}\int_{\theta_{ij}/2}^{\pi-\theta_{ij}/2}\left(\frac{1}{\alpha} + \frac{1}{\pi-\alpha}\right)\mathrm{d}\alpha\right.\\ &\quad\left. - \log(\pi) + \log\left(\pi - \frac{\theta_{ij}}{2}\right)\right\}\\ &= n + \frac{2}{n}\sum_{i<j}\left\{-2\log(\pi) + \frac{1}{\pi}\left[\theta_{ij}\log(\theta_{ij}) + (2\pi-\theta_{ij})\log(2\pi-\theta_{ij})\right]\right\},\end{aligned}$$

and thus the expression for $\psi_1^{\mathrm{AD}}$ follows. For $d \geq 3$, write the statistic as

$$P_{n,d-1}^{\mathrm{AD}} = \lim_{\varepsilon\to 0} P_{n,d-1}^{\mathrm{AD},\varepsilon},$$

where, for $\varepsilon > 0$,

$$\begin{aligned}P_{n,d-1}^{\mathrm{AD},\varepsilon} &:= \frac{1}{n}\sum_{i\neq j} W_{ij}^{\varepsilon} + \int_{-1+\varepsilon}^{1-\varepsilon} F_{d-1}(x)(1-nF_{d-1}(x))w(F_{d-1}(x))\,\mathrm{d}F_{d-1}(x), \qquad (5.34)\\ W_{ij}^{\varepsilon} &:= \int_{-1+\varepsilon}^{1-\varepsilon} A(\theta_{ij},x)w(F_{d-1}(x))\,\mathrm{d}F_{d-1}(x).\end{aligned}$$

For the second term of (5.34) we have that

$$\begin{aligned}\int_{-1+\varepsilon}^{1-\varepsilon}\frac{F_{d-1}(x)(1-nF_{d-1}(x))}{F_{d-1}(x)(1-F_{d-1}(x))}\,\mathrm{d}F_{d-1}(x) &= n\left(F_{d-1}(1-\varepsilon) - F_{d-1}(-1+\varepsilon)\right)\\ &\quad + (n-1)\log\left(\frac{1-F_{d-1}(1-\varepsilon)}{1-F_{d-1}(-1+\varepsilon)}\right). \qquad (5.35)\end{aligned}$$

By the same arguments as in (5.22) and (5.23), the first term of (5.34) is

$$W_{ij}^{\varepsilon} = 2\int_{-1}^{1} H_{d-1}^{\varepsilon}(t)\left\{\int_{-1}^{1} 1_{\{t\cos(\theta_{ij})+u(1-t^2)^{1/2}\sin(\theta_{ij})\leq t\}}\,\mathrm{d}F_{d-2}(u)\right\}\mathrm{d}F_{d-1}(t) =: \psi_{d-1}^{\varepsilon}(\theta_{ij}),$$

where, for $x \in (-1+\varepsilon, 1-\varepsilon)$,

$$H_{d-1}^{\varepsilon}(x) := \int_{x}^{1-\varepsilon} w(F_{d-1}(t))\,\mathrm{d}F_{d-1}(t) = \log\left(\frac{F_{d-1}(1-\varepsilon)}{1-F_{d-1}(1-\varepsilon)}\right) + \log\left(\frac{1-F_{d-1}(x)}{F_{d-1}(x)}\right).$$

Analogously to (5.23), we have that

$$\begin{aligned}\psi_{d-1}^{\mathrm{AD},\varepsilon}(\theta) &= 2\int_{\cos(\theta/2)}^{1} H_{d-1}^{\varepsilon}(t)\,\mathrm{d}F_{d-1}(t) + 2\int_{-\cos(\theta/2)}^{\cos(\theta/2)} H_{d-1}^{\varepsilon}(t)\,F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t)\\ &=: \psi_{d,1}^{\mathrm{AD},\varepsilon}(\theta) + \psi_{d,2}^{\mathrm{AD},\varepsilon}(\theta).\end{aligned}$$

Each term here can be computed separately:

$$
\begin{aligned}
\psi_{d,1}^{\mathrm{AD},\varepsilon}(\theta) ={}& -2F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right)\log\left(\frac{F_{d-1}(1-\varepsilon)}{1-F_{d-1}(1-\varepsilon)}\right) - 2\int_0^{\cos(\theta/2)}\log\left(\frac{1-F_{d-1}(x)}{F_{d-1}(x)}\right)\mathrm{d}F_{d-1}(t)\\
&-\log(4) - 2\log\left(1-F_{d-1}(1-\varepsilon)\right), \qquad (5.36)\\
\psi_{d,2}^{\mathrm{AD},\varepsilon}(\theta) ={}& \left(2F_{d-1}(\cos\left(\tfrac{\theta}{2}\right)) - 1\right)\log\left(\frac{F_{d-1}(1-\varepsilon)}{1-F_{d-1}(1-\varepsilon)}\right)\\
&+2\int_0^{\cos(\theta/2)}\log\left(\frac{1-F_{d-1}(t)}{F_{d-1}(t)}\right)\left(2F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)-1\right)\mathrm{d}F_{d-1}(t), \qquad (5.37)
\end{aligned}
$$

where in the first term of (5.37) it is employed that

$$
\int_{-\cos(\theta/2)}^{\cos(\theta/2)} F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t) = F_{d-1}\left(\cos\left(\tfrac{\theta}{2}\right)\right) - \frac{1}{2}
$$

and in the second that

$$
\begin{aligned}
&\int_{-\cos(\theta/2)}^{\cos(\theta/2)}\log\left(\frac{1-F_{d-1}(t)}{F_{d-1}(t)}\right)F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\mathrm{d}F_{d-1}(t)\\
&\quad= \int_0^{\cos(\theta/2)}\log\left(\frac{1-F_{d-1}(t)}{F_{d-1}(t)}\right)\left(2F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)-1\right)\mathrm{d}F_{d-1}(t).
\end{aligned}
$$

From (5.32), it results

$$
\psi_{d-1}^{\mathrm{AD},\varepsilon}(\theta) = \psi_{d-1}^{\mathrm{AD}}(\theta) - \log\left(1-F_{d-1}(1-\varepsilon)\right) - \log\left(F_{d-1}(1-\varepsilon)\right). \tag{5.38}
$$

Consequently, by (5.34), (5.35), (5.36), and (5.37):

$$
\begin{aligned}
P_{n,d-1}^{\mathrm{AD}} ={}& \lim_{\varepsilon\to 0}\left\{\frac{2}{n}\sum_{i<j}\left[\psi_{d-1}^{\mathrm{AD}}(\theta_{ij}) - \log\left(1-F_{d-1}(1-\varepsilon)\right) - \log\left(F_{d-1}(1-\varepsilon)\right)\right]\right.\\
&\left.+\,n\left(F_{d-1}(1-\varepsilon) - F_{d-1}(-1+\varepsilon)\right) + (n-1)\log\left(\frac{1-F_{d-1}(1-\varepsilon)}{1-F_{d-1}(-1+\varepsilon)}\right)\right\}\\
={}& n + \frac{2}{n}\sum_{i<j}\psi_{d-1}^{\mathrm{AD}}(\theta_{ij}).
\end{aligned}
$$

The particular expression for $d = 3$ follows trivially from (5.32). For $d = 4$, from (5.32) and taking into account that $F_3(\cos(\theta/2)) = 1 + (\sin(\theta) - \theta)/(2\pi)$, it follows

$$
\begin{aligned}
\psi_3^{\mathrm{AD}}(\theta) ={}& -\log(4) - 2\log(\pi) + 2\log(2\pi + \sin(\theta) - \theta) + \frac{1}{\pi}(\sin(\theta)-\theta)\log\left(\frac{2\pi + \sin(\theta) - \theta}{\theta - \sin(\theta)}\right)\\
&-\frac{4\tan\left(\frac{\theta}{2}\right)}{\pi}\int_0^{\cos(\theta/2)} t\log\left(\frac{\pi}{\cos^{-1}(t) - t(1-t^2)^{1/2}} - 1\right)\mathrm{d}t.
\end{aligned}
$$

□

**Remark 5.1.16.** *Because* $\lim_{\theta \to 0^+} \psi_{d-1}^{\mathrm{AD}}(\theta) = 0$*, by continuity,* $\psi_{d-1}^{\mathrm{AD}}(0) := 0$.

## 5.1.6 Summary

We conclude this section by summarizing in Table 5.1 the new test statistics proposals and by showing in Figure 5.1 the associated kernel functions of $P_{n,d-1}^{\mathrm{CvM}}$, $P_{n,d-1}^{\mathrm{R}_t}$, and $P_{n,d-1}^{\mathrm{AD}}$.

In Figure 5.1, the formulation in (5.17) is considered to achieve a standardization of the three kernel functions. Recall the difference in vertical scales, indicating the larger variability of $P_{n,d-1}^{\mathrm{R}_t}$ and $P_{n,d-1}^{\mathrm{AD}}$ with respect to $P_{n,d-1}^{\mathrm{CvM}}$. The kernels $\tilde{\psi}_{d-1}^{\mathrm{R}_{1-t}}$ and $\tilde{\psi}_{d-1}^{\mathrm{R}_t}$ coincide (Remark 5.1.14).

**Table 5.1.:** Extensions and connections given by the new family of projected tests.

| Test | $\Omega^1$- or $\Omega^2$-specific | $\Omega^{d-1}, d \geq 2$ |
|---|---|---|
| Watson | $U_n^2$ for $\Omega^1$ (Watson [135]) | $P_{n,d-1}^{\mathrm{CvM}}$ |
| Ajne | $A_n$ for $\Omega^1$ (Ajne [4]) | $P_{n,d-1}^{\mathrm{R}_{1/2}}$ [1] |
| Rothman | $R_{n,t}$ for $\Omega^1$ (Rothman [117]) | $P_{n,d-1}^{\mathrm{R}_t}$ |
| Chordal-based | $N_{n,2}$ for $\Omega^2$ (Bakshaev [11]) | $P_{n,d-1}^{\mathrm{CvM}}$ |
| Anderson–Darling | — | $P_{n,d-1}^{\mathrm{AD}}$ |

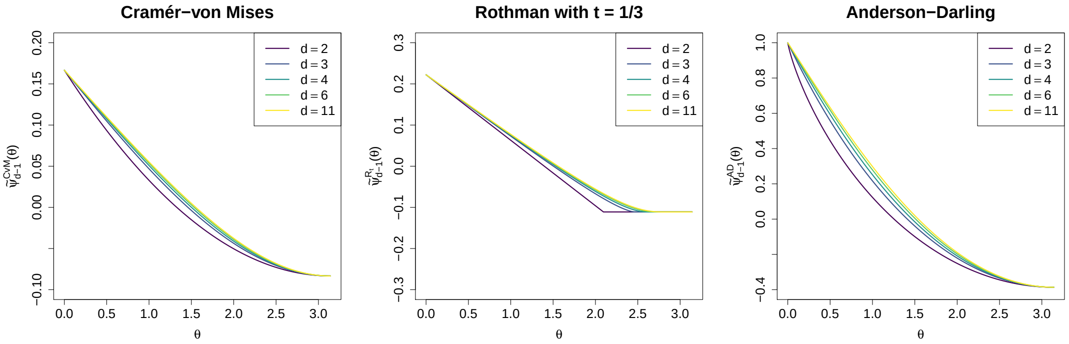


**Figure 5.1.:** From left to right, depiction of $\tilde{\psi}_{d-1}^{\mathrm{CvM}}(\theta) = \psi_{d-1}^{\mathrm{CvM}}(\theta) - \frac{1}{3}$, $\tilde{\psi}_{d-1}^{\mathrm{R}_{1/3}}(\theta) = \psi_{d-1}^{\mathrm{R}_{1/3}}(\theta) - \frac{5}{18}$, and $\tilde{\psi}_{d-1}^{\mathrm{AD}}(\theta) = \psi_{d-1}^{\mathrm{AD}}(\theta) + 1$, for $d = 2, 3, 4, 6, 11$.

[1] $P_{n,d-1}^{\mathrm{R}_{1/2}}$ coincides with Prentice [114]'s extension of $A_n$ to $\Omega^{d-1}$.

## 5.2 Connection with the Sobolev class

This section analyses the relation of the new class with the Sobolev class and gives its asymptotics. The notation about the class of Sobolev tests given in Subsection 2.6.1 will be useful.

### 5.2.1 Relation between Sobolev and projected classes

Despite being rooted on a different motivation, the class of projected statistics is indeed related with the Sobolev class, as explained in the Introduction. Thus, an important issue to address is the existence of a bijection between both classes:

Q1) For any projected statistic with measure $W$, is there a $f^W \in \mathcal{F}_{d-1}$ such that the statistic is expressible as a Sobolev one with $g_{f^W}(z) = \tilde{\psi}_{d-1}^W(\cos^{-1}(z))$ in (2.14)?

Q2) For any Sobolev statistic with $f \in \mathcal{F}_{d-1}$, is there a measure $W_f$ such that the statistic is expressible as a projected one with $\tilde{\psi}_{d-1}^{W_f}(\theta) = g_f(\cos(\theta))$ in (5.17)?

To elucidate these queries, we introduce next several classes of *projected-ecdf* test statistics with varying generality. We exclude the weights with $W(\{0\}) > 0$ due to its lack of statistical interest and related technical difficulties.

**Definition 5.2.1** (Projected-ecdf classes)**.** *The* projected-ecdf classes *of statistics (a) $\mathcal{P}_+$, (b) $\mathcal{P}_{\sigma+}$, and (c) $\mathcal{P}_\pm$ are defined as the collection of statistics $P_{n,d-1}^W$ indexed by $W$, a measure on $[0,1]$, with $W(\{0\}) = 0$, such that, for each class, (a) $W$ is a probability, (b) $W$ is positive $\sigma$-finite, (c) $W$ is signed, finite, and absolutely continuous with respect to the Lebesgue measure.*

Obviously, from the definition $\mathcal{P}_+ \subset \mathcal{P}_{\sigma+}$. Also, $P_{n,d-1}^{\mathrm{CvM}}, P_{n,d-1}^{\mathrm{R}_t} \in \mathcal{P}_+$ and $P_{n,d-1}^{\mathrm{AD}} \in \mathcal{P}_{\sigma+}$.The next theorem answers the questions raised in Q1) and Q2) above in light of the different projected-ecdf classes. Its first statement concludes that "sensible" projected-ecdf statistics (associated to probabilities $W$), as well as the Anderson–Darling statistic, are indeed Sobolev statistics. The second statement shows that finite Sobolev tests are projected-ecdf statistics with absolutely continuous, potentially signed, measures $W$.

**Theorem 5.2.2** (Projected-ecdf and Sobolev classes relations)**.** *It happens that:*

*i.* $\mathcal{P}_+ \subset \mathcal{S}$ *and* $P_{n,d-1}^{\mathrm{AD}} \in \mathcal{S}$.

*ii.* $\mathcal{S}_\ell \subset \mathcal{P}_\pm$*, for all* $\ell \geq 1$.

The proof of Theorem 5.2.2 is split for each statement. We begin with some lemmas required to prove *i*.

**Lemma 5.2.3.** *If $a > b$ and $\gamma \in \mathbb{R}$, then the order of $\Gamma(\gamma k + a)/\Gamma(\gamma k + b)$ as $k \to \infty$ is $(\gamma k + a)^{a-b}$.*

*Proof.* Stirling's equivalence gives

$$\begin{aligned}\frac{\Gamma(\gamma k + a)}{\Gamma(\gamma k + b)} &\sim \frac{e^{-(\gamma k+a)}}{e^{-(\gamma k+b)}} \frac{(\gamma k + a)^{\gamma k+a}}{(\gamma k + b)^{\gamma k+b}} \frac{(\gamma k + b)^{1/2}}{(\gamma k + a)^{1/2}} \\ &\sim e^{b-a} \left(\frac{\gamma k + a}{\gamma k + b}\right)^{\gamma k+b} (\gamma k + a)^{a-b} \\ &\sim e^{b-a} e^{a-b} (\gamma k + a)^{a-b}.\end{aligned}$$

□

**Lemma 5.2.4.** *Let $k \geq 1$ and $d \geq 3$. For $x \in [-1, 1]$, we have*

$$\left|C_k^{(d-3)/2}(x)\right| \leq \frac{\Gamma(k + (d-2)/2)}{\Gamma(d-2)\Gamma(k+1)}.$$

*Proof.* The result follows from equation (11) in Lohöfer [101] and equation 18.14.4 in [46]. □

**Lemma 5.2.5.** *Let $x \in [-1, 1]$ and consider the function $f_{d-1}^x(z) := (F_{d-1}(x))^{-1} 1_{\{z \leq x\}}, z \in [-1, 1]$. For $k \geq 1$, denote by $(1 + 2k/(d-2))v_{k,d-1}^x$ and by $2v_{k,d-1}^x$ to the Gegenbauer ($d \geq 3$) and Chebyshev coefficients ($d = 2$) of $f_{d-1}^x$, respectively. Then,*

$$|v_{k,d-1}^x| \leq \begin{cases} (k\pi F_1(x))^{-1}, & d = 2, \\ 2^{d-3/2} \left(\Gamma\left(\frac{d}{2}\right)\right)^2 (\pi F_{d-1}(x))^{-1} \mathcal{O}\left((k + (d-1)/2)^{-d/2}\right), & d \geq 3. \end{cases}$$

*Proof.* For $d = 2$, we have that

$$|v_{k,1}^x| = \frac{1}{2c_{k,1} F_1(x)} \left| \int_{-1}^{x} T_k(z)(1 - z^2)^{-1/2} \, \mathrm{d}z \right| = \frac{1}{k\pi F_1(x)} |\sin(k \cos^{-1}(x))| \leq \frac{1}{k\pi F_1(x)}.$$

For $d \geq 3$, equation 18.17.1 in [46] gives

$$\int_0^x C_k^{(d-2)/2}(z)(1 - z^2)^{(d-3)/2} \, \mathrm{d}z = \frac{d-2}{k(k+d-2)} \left(C_{k-1}^{d/2}(0) - (1 - x^2)^{(d-1)/2} C_{k-1}^{d/2}(x)\right). \tag{5.39}$$

The parity of $C_k^{(d-2)/2}$ jointly with the definition of the Gegenbauer coefficients and (5.39) give that

$$\begin{aligned}
v^x_{k,d-1} &= \frac{1}{c_{k,d-1}\left(1+\frac{2k}{d-2}\right)F_{d-1}(x)}\left(\int_{-1}^{0} C_k^{(d-2)/2}(z)(1-z^2)^{(d-3)/2}\,\mathrm{d}z\right.\\
&\qquad\left.+\int_0^x C_k^{(d-2)/2}(z)(1-z^2)^{(d-3)/2}\,\mathrm{d}z\right)\\
&= \frac{2^{d-2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi(d-2)F_{d-1}(x)}\,\frac{(1-q)(1-x^2)^{(d-1)/2}C^{d/2}_{k-1}(x)}{k(k+d-2)}\\
&= \frac{-2^{d-2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\,\frac{(k-1)!}{\Gamma(k+d-1)}(1-x^2)^{(d-1)/2}C^{d/2}_{k-1}(x). \qquad (5.40)
\end{aligned}$$

Then, from (5.40), we have

$$\begin{aligned}
|v^x_{k,d-1}| &\le \frac{2^{d-3/2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\,\frac{(k-1)!}{\Gamma(k+d-1)}C^{d/2}_{k-1}(x)\\
&\le \frac{2^{d-3/2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\,\frac{(k-1)!}{\Gamma(k+d-1)}\sup_{x\in[-1,1]}\left|C^{d/2}_{k-1}(x)\right|\\
&\le \frac{2^{d-3/2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\,\frac{(k-1)!}{\Gamma(k+d-1)}\,\frac{\Gamma(k+(d-2)/2)}{\Gamma(k)} \qquad (5.41)\\
&= \frac{2^{d-3/2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\mathcal{O}\left((k+(d-1)/2)^{-d/2}\right), \qquad (5.42)
\end{aligned}$$

where the inequality (5.41) comes from Lemma 5.2.4 and equality (5.42) from Lemma 5.2.3. □

In the proof of the previous lemma, we have obtained relation (5.40) that we state separately for further reference.

**Corollary 5.2.6.** *Let $x \in [-1,1]$. For $k \ge 1$, consider the Gegenbauer ($d \ge 3$) and Chebyshev coefficients ($d = 2$) of $f^x_{d-1}$ defined in Lemma 5.2.5 $(1 + 2k/(d-2))v^x_{k,d-1}$ and $2v^x_{k,d-1}$, respectively. Then,*

$$v^x_{k,d-1} = \frac{-2^{d-2}\left(\Gamma\left(\frac{d}{2}\right)\right)^2}{\pi F_{d-1}(x)}\,\frac{(k-1)!}{\Gamma(k+d-1)}(1-x^2)^{(d-1)/2}C^{d/2}_{k-1}(x).$$

**Lemma 5.2.7.** *It happens that $b^{\mathrm{AD}}_{k,1} = \mathcal{O}(\log(k)/k^2)$ and, consequently, $\sum_{k=1}^{\infty} b^{\mathrm{AD}}_{k,1} < \infty$.*

*Proof.* It happens that

$$b^{\mathrm{AD}}_{k,1} \le \frac{2}{\pi k^2}\left(\int_0^{\pi/2}\frac{1-\cos(2k\theta)}{\theta}\,\mathrm{d}\theta + \int_{\pi/2}^{\pi}\frac{1-\cos(2k\theta)}{\pi-\theta}\,\mathrm{d}\theta\right). \qquad (5.43)$$

We will only consider the first integral in (5.43) because the other one is handled similarly. Obviously, if $k \geq 1$, then

$$\int_0^{\pi/2} \frac{1-\cos(2k\theta)}{\theta}\,\mathrm{d}\theta = \int_0^{k\pi} \frac{1-\cos(u)}{u}\,\mathrm{d}u \leq I_1 + \sum_{h=1}^{k-1} \frac{1}{h\pi} \int_{h\pi}^{(h+1)\pi} (1-\cos(u))\,\mathrm{d}u,$$

where $I_1 = \int_0^{\pi} \frac{1-\cos(u)}{u}\,\mathrm{d}u < \infty$. Therefore,

$$\int_0^{\pi/2} \frac{1-\cos(2k\theta)}{\theta}\,\mathrm{d}\theta \leq I_1 + \sum_{h=1}^{k-1} \frac{1}{h} = \mathcal{O}(\log(k))$$

and (5.43) gives that $b_{k,1}^{\mathrm{AD}} = \mathcal{O}(\log(k)/k^2)$, which proves the lemma. Note that $\sum_{k=1}^{\infty} \frac{\log(k)}{k^2} = -\frac{1}{6}\pi^2(-12\log(A) + \gamma + \log(2) + \log(\pi)) < \infty$, where $A$ and $\gamma$ are the Glaisher–Kinkelin and Euler–Mascheroni constants, respectively. □

**Remark 5.2.8.** *Lemma 5.2.7 also shows that* $\psi_1^{\mathrm{AD}} \in L_{d-1}^2[-1,1]$.

*Proof of $\mathcal{P}_+ \subset \mathcal{S}$ in i) in Theorem 5.2.2.* The proof goes along two steps.

*Step 1.* Assume in this step that $W$ is the Dirac's delta on the point $F_{d-1}(x) \in (0,1]$, $\delta_{F_{d-1}(x)}$, given by $x \in (-1,1]$. Clearly, $W(F_{d-1}(A)) = \delta_x(A)$ for $A \subset [-1,1]$, and therefore from (5.3) we have

$$P_{n,d-1}^{\delta_{F_{d-1}(x)}} = n\mathrm{E}_\gamma\left(\{F_{n,\gamma}(x) - F_{d-1}(x)\}^2\right) = \frac{F_{d-1}(x)^2}{n}\mathrm{E}_\gamma\left(\left\{\sum_{i=1}^{n} \frac{1_{\{\gamma'\mathbf{X}_i \leq x\}}}{F_{d-1}(x)} - 1\right\}^2\right). \tag{5.44}$$

Given $x \in [-1,1]$, the function $f_{d-1}^x(z)$ defined in Lemma 5.2.5 is bounded, thus $f_{d-1}^x \in L_{d-1}^2[-1,1]$. Also, its first Gegenbauer coefficient equals one since: if $d = 2$, we have that

$$b_{0,1} = \frac{1}{c_{0,1}} \int_{-1}^{1} f_1^x(z)(1-z^2)^{-1/2}\,\mathrm{d}z = \frac{1}{\pi F_1(x)}\mathrm{B}\left(\tfrac{1}{2}, \tfrac{1}{2}\right) F_1(x) = 1,$$

and, if $d \geq 3$, the Legendre duplication formula gives

$$\begin{aligned} b_{0,d-1} &= \frac{1}{c_{0,d-1}} \int_{-1}^{1} F_{d-1}^x(z)(1-z^2)^{(d-3)/2}\,\mathrm{d}z \\ &= \frac{2^{d-2}\Gamma(d/2)^2}{\pi(d-2)\Gamma(d-2)} \frac{1}{F_{d-1}(x)}\mathrm{B}\left(\tfrac{1}{2}, \tfrac{d-1}{2}\right) F_{d-1}(x) = 1. \end{aligned}$$

Based on (2.11), denote by $(1+2k/(d-2))v_{k,d-1}^x$ to the remaining Gegenbauer coefficients of $F_{d-1}^x$ for $d \geq 2$ (employing extension (2.9) if $d = 2$). By showing $\sum_{k=1}^{\infty}(v_{k,d-1}^x)^2\ell_{k,d-1} < \infty$, (5.44) and (2.10) would give the equality $P_{n,d-1}^{\delta_{F_{d-1}(x)}} = F_{d-1}(x)^2 S_{n,d-1}(f_{d-1}^x)$ and the proof will be completed for the case $W = \delta_{F_{d-1}(x)}$.

For $d = 2$, $d_{k,1} = 2$ for $k \geq 1$ so, by Lemma 5.2.5, $\sum_{k=1}^{\infty} 2(v_{k,d-1}^x)^2 < \infty$ as desired. For $d \geq 3$, according to (2.6), $\ell_{k,d-1} \sim k^{d-2}$, so by Lemma 5.2.5 we have that $(v_{k,d-1}^x)^2 \ell_{k,d-1} \sim k^{-2}$ and therefore $\sum_{k=1}^{\infty} (v_{k,d-1}^x)^2 \ell_{k,d-1} < \infty$ as sought.

*Step 2.* Assume now that $W$ is a probability measure defined on the Borel sets on $[0,1]$ with $W(\{0\}) = 0$. The first step is to make explicit the relation between the functions $\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}$ and $g_{f_{d-1}^x}$ (defined in (5.17) and (2.13), respectively). Since $P_{n,d-1}^{\delta_{F_{d-1}(x)}} = F_{d-1}(x)^2 S_{n,d-1}(f_{d-1}^x)$ holds for every $x \in [-1,1]$ and $n \geq 1$, if we apply this relation with $n = 1$, from (5.17) and (2.13), we obtain that $\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\theta_{ii}) = F_{d-1}(x)^2 g_{f_{d-1}^x}(\mathbf{X}_i'\mathbf{X}_i)$ for every $i = 1, \ldots, n$. From here, and (5.17) and (2.13) with $n = 2$, we obtain that $\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\theta_{12}) = F_{d-1}(x)^2 g_{f_{d-1}^x}(\mathbf{X}_1'\mathbf{X}_2)$ for every selection of $\mathbf{X}_1, \mathbf{X}_2 \in \Omega^{d-1}$, therefore implying that

$$\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\theta) = F_{d-1}(x)^2 g_{f_{d-1}^x}(\cos\theta), \text{ for all } \theta \in [0,\pi]. \tag{5.45}$$

Additionally, notice that $\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\theta) = A(\theta, x) - F_{d-1}(x)^2$ due to (5.17) and since the Gegenbauer coefficients of $g_{F_{d-1}^x}$ are $0$ and $(1 + 2k/(d-2))(v_{k,d-1}^x)^2$ for $k \geq 1$ by (2.12).

Now, consider the projected ecdf statistic $P_{n,d-1}^W$ with kernel $\tilde{\psi}_{d-1}^W$. We have that

$$\begin{aligned} P_{n,d-1}^W &= \int_{-1}^{1} \left[ n\mathrm{E}_\gamma \left( \{F_{n,\gamma}(x) - F_{d-1}(x)\}^2 \right) \right] \mathrm{d}W(F_{d-1}(x)) \\ &= \int_{-1}^{1} P_{n,d-1}^{\delta_{F_{d-1}(x)}} \,\mathrm{d}W(F_{d-1}(x)) \\ &= \frac{1}{n} \sum_{i,j=1}^{n} \int_{-1}^{1} F_{d-1}(x)^2 g_{f_{d-1}^x}(\mathbf{X}_i'\mathbf{X}_j) \,\mathrm{d}W(F_{d-1}(x)), \end{aligned}$$

where the last equality follows from (5.15) and (5.45). Let us consider the function

$$g^W(z) := \int_{-1}^{1} F_{d-1}(x)^2 g_{f_{d-1}^x}(z) \,\mathrm{d}W(F_{d-1}(x)), \quad z \in [-1,1]. \tag{5.46}$$

We have that

$$|F_{d-1}(x)^2 g_{f_{d-1}^x}(z)| = \left|\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\cos^{-1}(z))\right| = |A(\cos^{-1}(z), x) - F_{d-1}(x)^2| \leq 1,$$

thus $g^W$ is bounded and $g^W \in L_{d-1}^2[-1,1]$. Denote by $(1 + 2k/(d-2))u_{k,d-1}^W$, $k \geq 1$, to its Gegenbauer coefficients. If we prove that $u_{0,d-1}^W = 0$, $u_{k,d-1}^W \geq 0$, $k \geq 1$, and that $\sum_{k=1}^{\infty} u_{k,d-1}^W \ell_{k,d-1} < \infty$, from (2.13), we would have that $P_{n,d-1}^W = S_{n,d-1}(\{v_{k,d-1}^W\})$ for $v_{k,d-1}^W := (u_{k,d-1}^W)^{1/2}$ and $\mathcal{P}_+ \subset \mathcal{S}$ would be proved. Since the map $(x,z) \mapsto F_{d-1}(x)^2 g_{f_{d-1}^x}(z)$ is bounded, Fubini's theorem gives for $q \geq 1$ and $k \geq 1$ that

$$u_{k,d-1}^W = \frac{1}{(1+\frac{2k}{d-2})c_{k,d-1}} \int_{-1}^{1} g^W(z) C_k^{(d-2)/2}(z)(1-z^2)^{(d-3)/2} \,\mathrm{d}z$$

$$= \frac{1}{(1+\frac{2k}{d-2})c_{k,d-1}} \int_{-1}^{1} F_{d-1}(x)^2 \left[ \int_{-1}^{1} g_{f^x_{d-1}}(z) C_k^{\frac{d-2}{2}}(z)(1-z^2)^{\frac{d-3}{2}} \,\mathrm{d}z \right] \mathrm{d}W(F_{d-1}(x))$$
$$= \int_{-1}^{1} \left( F_{d-1}(x) v^x_{k,d-1} \right)^2 \mathrm{d}W(F_{d-1}(x)) \geq 0.$$

This reasoning also shows that $u^W_{0,d-1} = 0$. Moreover, from here and Lemma 5.2.5 we have

$$u^W_{k,d-1} \leq \int_{-1}^{1} \mathrm{d}W(F_{d-1}(x)) \frac{2^{2d-3}\left(\Gamma\left(\frac{d}{2}\right)\right)^4}{\pi^2 F_{d-1}(x)^2} \mathcal{O}\left( (k + (d-1)/2)^{-d} \right)$$

for $d \geq 3$, which implies that $\sum_{k=1}^{\infty} u^W_{k,d-1} \ell_{k,d-1} < \infty$ for $d \geq 2$. □

*Proof of $P^{\mathrm{AD}}_{n,d-1} \in \mathcal{S}$ in i) in Theorem 5.2.2.* For $d \geq 3$, relation (5.38) shows that, for every $\varepsilon > 0$, there exists a finite and positive measure $W^\varepsilon$ such that

$$\psi^{\mathrm{AD}}_d(\theta) = \psi^{W^\varepsilon}_d(\theta) + h(\varepsilon),$$

where $h$ is a real function depending on $\varepsilon$ but not on $\theta$. Therefore, $\psi^{\mathrm{AD}}_{d-1}$ equals to a constant plus the $L^2_{d-1}[-1,1]$ map $z \mapsto \psi^{W^\varepsilon}_{d-1}(\cos(z))$.

Consequently, $\psi^{\mathrm{AD}}_{d-1} \in L^2_{d-1}[-1,1]$ and all its Gegenbauer coefficients for $k \geq 1$, denoted by $(1 + 2k/(d-2))u^{\mathrm{AD}}_{k,d-1}$, coincide with those of $\psi^{W^\varepsilon}_{d-1}$, which are positive and satisfy that $\sum_{k=1}^{\infty} u^{\mathrm{AD}}_{k,d-1} \ell_{k,d-1} < \infty$ because, trivially, the proof when $W$ is a probability can be extended to cover $W^\varepsilon$ for fixed $\varepsilon > 0$.

With respect to $d = 2$, in Proposition 5.2.22 we obtain the Gegenbauer coefficients of $\psi^{\mathrm{AD}}_1$, $\{b^{\mathrm{AD}}_{k,1}\}$, which are obviously non-negative. Moreover, Lemma 5.2.7 shows that $b^{\mathrm{AD}}_{k,1} = O(\log(k)/k^2)$, so there exists a function $f \in L^2_{d-1}[-1,1]$ whose Gegenbauer coefficients are 1 and $(b^{\mathrm{AD}}_{k,1})^{1/2}$, $k \geq 1$. This shows that $P^{\mathrm{AD}}_{n,1} = S_{n,1}(f)$ since $d_{k,1} = 2$ for $k \geq 1$. □

**Remark 5.2.9.** *The contention $\mathcal{P}_+ \subset \mathcal{S}$ is strict since $P^{\mathrm{AD}}_{n,d-1} \in \mathcal{S}$ while $P^{\mathrm{AD}}_{n,d-1} \notin \mathcal{P}_+$.*

Notice that in the proof of the previous theorem we have proved relation (5.45) that we state separately for further reference.

**Corollary 5.2.10.** *Let $x \in [-1,1]$. Consider the functions $f^x_{d-1}$, $\tilde{\psi}^{\delta_{F_{d-1}(x)}}_{d-1}$ and $g_{f^x_{d-1}}$ defined in Lemma 5.2.5, in (5.17) and in (2.13), respectively. Then for $d \geq 2$ and any $\theta \in [0,\pi]$,*

$$\tilde{\psi}^{\delta_{F_{d-1}(x)}}_{d-1}(\theta) = F_{d-1}(x)^2 g_{f^x_{d-1}}(\cos\theta).$$

**Remark 5.2.11.** *In Step 1 of the proof of $\mathcal{P}_+ \subset \mathcal{S}$, we set the function $f^x_{d-1}(z) = (F_{d-1}(x))^{-1}$ $1_{\{z \leq x\}}$ and obtain that $P^{\delta_{F_{d-1}(x)}}_{n,d-1} = F_{d-1}(x)^2 S_{n,d-1}(F^x_{d-1})$. Step 2, (5.46) with $W = \delta_{F_{d-1}(x)}$*

*gives a different function, $g^{\delta_{F_{d-1}(x)}}$, such that its Gegenbauer coefficients $\left(v_{k,d-1}^{\delta_{F_{d-1}(x)}}\right)^2$, satisfy $P_{n,d-1}^{\delta_{F_{d-1}(x)}} = S_{n,d-1}\left(\left\{v_{k,d-1}^{\delta_{F_{d-1}(x)}}\right\}\right)$. Taking $f^{\delta_{F_{d-1}(x)}}(z) = 1_{\{z\leq x\}} - F_{d-1}(x) + 1$, it is simple to check that $P_{n,d-1}^{\delta_{F_{d-1}(x)}} = S_{n,d-1}(f^{\delta_{F_{d-1}(x)}})$.*

The proof of *ii*) in Theorem 5.2.2 requires some results on integral equations that are given in Section 2.7 and Theorem 5.2.12. In such theorem we show how to apply Proposition 2.7.3 to the kernel $(s,t) \mapsto A(\cos^{-1}(s),t)$ with $A$ defined as in Proposition 5.1.4. Note that this is trivially an $L^2$-kernel due to its boundedness.

**Theorem 5.2.12.** *Let be the $L^2$-kernel $K(s,t) = A(\cos^{-1}(s),t)$, $s,t \in [-1,1]$, and let $\{([u_n,v_n];\mu_n)\}_{n=1}^{\infty}$ be a singular system of $K$. Let $y \in L^2[-1,1]$ satisfying a) in Proposition 2.7.3. Then, there exists $g \in L^2[-1,1]$ such that*

$$y(s) = \int_{-1}^{1} K(s,t)g(t)\,\mathrm{d}t \ \textit{ for almost every } s \in [-1,1]$$

*Proof.* Since condition a) in Proposition 2.7.3 holds by assumption, only condition b) remains to be proved. For this, since $K^*(s,t) = A(\cos^{-1}(t),s)$, we show that if $\int_{-1}^{1} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s = 0$ for almost every $t \in [-1,1]$ and $u \in L^2[-1,1]$, then $u(s) = 0$ for almost every $s \in [-1,1]$.

Assume $d \geq 3$. By (5.13) and since $\cos(2\cos^{-1}(s)) = 2s^2 - 1$,

$$A(\cos^{-1}(s),t) = \begin{cases} 2F_{d-1}(t) - 1, & -1 \leq s \leq 2t^2 - 1, \\ \frac{1}{2} - \frac{\cos^{-1}(s)}{2\pi} + 2\int_0^t F_{d-2}\left(\frac{z\tan(\cos^{-1}(s)/2)}{(1-z^2)^{1/2}}\right)\mathrm{d}F_{d-1}(z), & 2t^2 - 1 \leq s < 1. \end{cases} \tag{5.47}$$

Also, by assumption, for any $t \in [-1,1]$, we have

$$\frac{\partial}{\partial t}\int_{-1}^{1} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s = 0. \tag{5.48}$$

Therefore,

$$\begin{aligned} 0 &= \frac{\partial}{\partial t}\left(\int_{-1}^{2t^2-1} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s + \int_{2t^2-1}^{1} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s\right) \\ &= 4t(2F_{d-1}(t) - 1)u(2t^2-1) + \int_{-1}^{2t^2-1} \frac{\partial}{\partial t}A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\ &\quad - 4tu(2t^2-1)\left(\frac{1}{2} - \frac{\cos^{-1}(2t^2-1)}{2\pi}\right. \\ &\quad \left. +2\int_0^t F_{d-2}\left(\frac{z}{\sqrt{1-z^2}}\tan\left(\frac{\cos^{-1}(2t^2-1)}{2}\right)\right)\mathrm{d}F_{d-1}(z)\right) \end{aligned}$$

$$
\begin{aligned}
&+\int_{2t^2-1}^{1} \frac{\partial}{\partial t} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\
&= 4tu(2t^2-1)g_{d-1}(t) + \int_{-1}^{1} \frac{\partial}{\partial t} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\
&= 4tu(2t^2-1)g_{d-1}(t),
\end{aligned}
$$

with $g_{d-1}(t) := 2F_{d-1}(t) - \frac{3}{2} + \frac{\cos^{-1}(2t^2-1)}{2\pi} + 2\int_0^t F_{d-2}\left(\frac{z}{\sqrt{1-z^2}}\tan\left(\frac{\cos^{-1}(2t^2-1)}{2}\right)\right)\mathrm{d}F_{d-1}(z)$ and where the last equality follows from (5.48) by exchanging integral and differential. Hence, since $F_{d-1}(t) > F_{d-1}(0) = 1/2$, the result follows because $g_{d-1}(t) \geq 1 - 3/2 + \int_0^t \mathrm{d}F_{d-1}(y) > -1/2 + F_{d-1}(t) > 0$.

Showing that $\partial A(\cos^{-1}(s),t)/\partial t$ exists and is bounded is enough to guarantee that switching differentiation and integration in (5.48) is possible. From (5.47), we have

$$
\frac{\partial}{\partial t} A(\cos^{-1}(s),t) = \begin{cases} 2f_{d-1}(t), & -1 \leq s \leq 2t^2-1, \\ 2F_{d-2}\left(\frac{t\tan(\cos^{-1}(s)/2)}{(1-t^2)^{1/2}}\right)F_{d-1}(t), & 2t^2-1 \leq s < 1, \end{cases}
$$

which is obviously bounded by the definitions of the functions $f_{d-1}$ and $F_{d-2}$, and since $d \geq 3$.

Assume now that $d = 2$. Obviously, it suffices to show that $\int_{-1}^{1} A(\cos^{-1}(s),t)u(s)\,\mathrm{d}s = 0$ for almost every $t \in (0,1)$ implies $u \equiv 0$ almost everywhere. Then, by assumption, $\int_{-1}^{1} \tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s = 0$ for almost every $t \in (0,1)$, with

$$
\begin{aligned}
\tilde{A}(\cos^{-1}(s),t) &:= (1-t^2)A(\cos^{-1}(s),t) \\
&= \begin{cases} (1-t^2)\left(1-\frac{2}{\pi}\cos^{-1}(t)\right), & -1 \leq s \leq 2t^2-1, \\ (1-t^2)\left(1-\frac{1}{\pi}\cos^{-1}(t)-\frac{1}{2\pi}\cos^{-1}(s)\right), & 2t^2-1 \leq s < 1, \end{cases}
\end{aligned}
$$

due to Proposition 5.1.4. Therefore,

$$
\begin{aligned}
0 &= \frac{\partial}{\partial t}\int_{-1}^{1} \tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\
&= \frac{\partial}{\partial t}\left(\int_{-1}^{2t^2-1} \tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s + \int_{2t^2-1}^{1} \tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s\right) \qquad (5.49) \\
&= 4tu(2t^2-1)\left(-2t\left(1-\frac{2}{\pi}\cos^{-1}(t)\right) + \frac{2(1-t^2)^{1/2}}{\pi}\right) + \int_{-1}^{2t^2-1} \frac{\partial}{\partial t}\tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\
&\quad - 4tu(2t^2-1)\left(-2t\left(1-\frac{1}{\pi}\cos^{-1}(t)-\frac{1}{2\pi}\cos^{-1}(2t^2-1)\right) + \frac{(1-t^2)^{1/2}}{\pi}\right) \\
&\quad + \int_{2t^2-1}^{1} \frac{\partial}{\partial t}\tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s \\
&= 4tu(2t^2-1)g_1(t) + \int_{-1}^{1} \frac{\partial}{\partial t}\tilde{A}(\cos^{-1}(s),t)u(s)\,\mathrm{d}s
\end{aligned}
$$

$$= 4tu(2t^2 - 1)g_1(t),$$

with $g_1(t) := \left(1 - t^2\right)^{1/2}/\pi$, because $\cos^{-1}(2t^2 - 1) = 2\cos^{-1}(t)$ since $t \in (0,1)$. Then, the result follows because $g_1(t) > 0$ for $t \in (0,1)$.

Notice that the exchange of integral and differential in (5.49) is possible because by Proposition 5.1.4 it is easy to see that

$$\frac{\partial}{\partial t}\tilde{A}(\cos^{-1}(s), t) = \begin{cases} -2t\left(1 - \frac{2}{\pi}\cos^{-1}(t)\right) + \frac{2}{\pi}\left(1 - t^2\right)^{1/2}, & -1 \leq s \leq 2t^2 - 1, \\ -2t\left(1 - \frac{1}{\pi}\cos^{-1}(t) - \frac{1}{2\pi}\cos^{-1}(s)\right) + \frac{1}{\pi}\left(1 - t^2\right)^{1/2}, & 2t^2 - 1 \leq s < 1 \end{cases}$$

is well-defined and bounded. □

*Proof of ii) in Theorem 5.2.2.* Under the conditions of Theorem 5.2.12 in the Appendix, if we additionally consider that the first Gegenbauer coefficient of $f$ is one, then there exists an absolutely continuous finite, potentially signed, Borel measure on $[-1,1]$ such that $S_{n,d-1}(f) = P^W_{n,d-1}$. This result follows directly from Theorem 5.2.12 in the Appendix, because this measure is the one whose density with respect to the Lebesgue measure is $f$, which acts as $g$ in the statement of such theorem. □

**Remark 5.2.13.** *The contention $\mathcal{P}_+ \subset \mathcal{S}$ is strict since $P^{\mathrm{AD}}_{n,d-1} \in \mathcal{S}$ while $P^{\mathrm{AD}}_{n,d-1} \notin \mathcal{P}_+$.*

**Remark 5.2.14.** *The proof of $\mathcal{P}_+ \subset \mathcal{S}$ does not guarantee the non-negativeness of $\tilde{f} \in \mathcal{F}_{d-1}$ such that $P^W_{n,d-1} = S_{n,d-1}(\tilde{f})$, unless $W$ is a Dirac's delta. This is allowed, as $\mathcal{F}_{d-1}$ contains non-positive functions, but obscures the local optimality view of $S_{n,d-1}(\tilde{f})$. Of course, when $\kappa \to 0$, (2.15) is always well-defined as a pdf, even if $f$ is non-positive.*

**Remark 5.2.15.** *The next counterexample shows that $\mathcal{P}_\pm \not\subset \mathcal{S}$. If $d = 2$ and $W(x) = \cos(4\pi x)$, then, by Proposition 5.1.5, $\psi_1(\theta) = 1 - (\cos(\theta)\sin(\theta))/(2\pi)$ and its Gegenbauer coefficients are $b_{k,1} = ((-1)^k - 1)/(\pi^2(4 - k^2)) \ngeq 0$, hence it can not be written as (2.13).*

### 5.2.2 Asymptotic null distributions and local optimality

In virtue of Theorem 5.2.2, all the projected-ecdf statistics within the class $\mathcal{P}_+$ belong to the Sobolev class. Therefore, the asymptotic distribution and local optimality results stated in Theorem 2.6.3 readily apply to the class of statistics $\mathcal{P}_+ \cup \{P^{\mathrm{AD}}_{n,d-1}\}$. This is collected in the next corollary, which follows directly from Theorems 2.6.3 and *i* by noting that the Gegenbauer coefficients of $\psi^W_{d-1}$ equal to $(1 + 2k/(d-2))v_{k,d-1}$ by (2.12). Remember that the coefficients $\ell_{k,d-1}$ were defined in (2.6).

**Corollary 5.2.16** (Asymptotic null distribution and local optimality)**.** *Let $W$ be a probability on $[0,1]$ or the Anderson–Darling measure. For $d \geq 2$, the Gegenbauer and Chebyshev coefficients of $\psi_{d-1}^W$, defined as*

$$b_{k,d-1}^W := \begin{cases} \dfrac{1}{c_{k,1}} \displaystyle\int_0^\pi \psi_{d-1}^W(\theta) T_k(\cos\theta)\,\mathrm{d}\theta, & d = 2, \\ \dfrac{1}{c_{k,d-1}} \displaystyle\int_0^\pi \psi_{d-1}^W(\theta) C_k^{(d-2)/2}(\cos\theta)\sin^{d-2}(\theta)\,\mathrm{d}\theta, & d \geq 3 \end{cases} \tag{5.50}$$

*for $k \geq 0$, are non-negative sequences satisfying $\sum_{k=1}^\infty b_{k,d-1}^W \ell_{k,d-1} < \infty$. Under $\mathbf{H}_0$,*

$$P_{n,d-1}^W \overset{d}{\rightsquigarrow} \begin{cases} \displaystyle\sum_{k=1}^\infty 2^{-1} b_{k,1}^W Y_k, & d = 2, \\ \displaystyle\sum_{k=1}^\infty \left(1 + \frac{2k}{d-2}\right)^{-1} b_{k,d-1}^W Y_k, & d \geq 3, \end{cases}$$

*where $Y_k \overset{d}{=} \chi^2_{\ell_{k,d-1}}$, $k \geq 1$, are independent rv's. In addition, the test that rejects for large values of $P_{n,d-1}^W$ is asymptotically and locally (in $\kappa \to 0$) most powerful rotation-invariant (except $O(\kappa^3)$ terms) against any pdf (2.15) based on $|\upsilon_{k,d-1}| = (b_{k,d-1}^W)^{1/2}$. Furthermore, if $b_{k,d-1}^W > 0$, for all $k \geq 1$, the test is consistent against all non-uniform alternatives with pdf in $L^2(\Omega^{d-1}, \nu_{d-1})$.*

The asymptotic distribution and local optimality of the $(\mathcal{P}_+ \cup \{P_{n,d-1}^{\mathrm{AD}}\})$-based tests are governed by (5.50). The following results are aimed to facilitate these coefficients.

**Theorem 5.2.17.** *For $x \in [-1,1]$ and $\theta \in [0,\pi]$, consider $A(\theta,x)$ defined as in (5.8). Then, for $d \geq 2$,*

$$A(\theta,x) = \sum_{k=0}^\infty a_{k,d-1}^x C_k^{(d-2)/2}(\cos\theta),$$

*where $a_{0,d-1}^x := F_{d-1}(x)^2$ and, for $k \geq 1$,*

$$a_{k,d-1}^x := \begin{cases} \dfrac{1}{k^2\pi^2}\left(1 - T_{2k}(x)\right), & d = 2, \\ \left(1 + \dfrac{2k}{d-2}\right)\left(\dfrac{2^{d-2}\Gamma\left(d/2\right)^2 \Gamma(k)}{\pi\Gamma(k+d-1)}\right)^2 (1-x^2)^{(}d-1)\left(C_{k-1}^{d/2}(x)\right)^2, & d \geq 3. \end{cases}$$

*Proof.* Suppose first that $d = 2$ and $k \geq 1$. By Proposition 5.1.4 the Gegenbauer coefficients of $A(\theta,x)$ are

$$\begin{aligned} b_{k,1}^x &= \frac{1}{c_{k,1}} \int_0^\pi A(\theta,x) T_k(\cos(\theta)) \sin^{d-2}(\theta)\,\mathrm{d}\theta \\ &= \frac{2}{\pi} \int_0^\pi A(\theta,x)\cos(k\theta)\,\mathrm{d}\theta \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\pi}\left(\int_0^{2\cos^{-1}(x)}\left(1-\frac{\cos^{-1}(x)}{\pi}-\frac{\theta}{2\pi}\right)\cos(k\theta)\,\mathrm{d}\theta\right.\\
&\qquad\left.+\int_{2\cos^{-1}(x)}^{\pi}\left(1-\frac{2\cos^{-1}(x)}{\pi}\right)\cos(k\theta)\,\mathrm{d}\theta\right)\\
&= \frac{2}{\pi}\int_0^{2\cos^{-1}(x)}\left(\frac{\cos^{-1}(x)}{\pi}-\frac{\theta}{2}\right)\cos(k\theta)\,\mathrm{d}\theta, \qquad (5.51)
\end{aligned}
$$

where (5.51) follows from the orthogonality of the Gegenbauer polynomials. Hence

$$
\begin{aligned}
a_{k,1}^x &= \frac{2}{\pi}\left(\frac{1}{k\pi}\cos^{-1}(x)\,[\sin(k\theta)]_0^{2\cos^{-1}(x)} - \left[\frac{1}{2k^2\pi}\cos(k\theta)+\frac{\theta}{2k\pi}\sin(k\theta)\right]_0^{2\cos^{-1}(x)}\right)\\
&= \frac{1}{k^2\pi^2}\left(1-\cos(2k\cos^{-1}(x))\right).
\end{aligned}
$$

For $k=0$, Proposition 5.1.4 gives that

$$
\begin{aligned}
a_{0,1}^x &= \frac{1}{\pi}\int_0^{\pi} A(\theta,x)\,\mathrm{d}\theta\\
&= \frac{1}{\pi}\int_0^{\pi}\left(2F_1(x)-1+\frac{1}{\pi}\left(\cos^{-1}(x)-\frac{\theta}{2}\right)_+\right)\mathrm{d}\theta\\
&= 2F_1(x)-1+\frac{1}{\pi^2}\int_0^{2\cos^{-1}(x)}\left(\cos^{-1}(x)-\frac{\theta}{2}\right)\mathrm{d}\theta\\
&= 2F_1(x)-1+\frac{1}{\pi^2}\left(\cos^{-1}(x)\right)^2\\
&= \left(F_1(x)\right)^2,
\end{aligned}
$$

where the last equality follows from (2.4).

Assume $d\geq 3$. From Corollary 5.2.10 and the identity $\tilde{\psi}_{d-1}^{\delta_{F_{d-1}(x)}}(\theta) = A(\theta,x)-F_{d-1}(x)^2$, we have that

$$
\begin{aligned}
A(\theta,x) &= F_{d-1}(x)^2(1+g_{f_{d-1}^x}(\cos\theta))\\
&= F_{d-1}(x)^2+\sum_{k=1}^{\infty}\left(v_{k,d-1}^x F_{d-1}(x)\right)^2\left(1+\frac{2k}{d-2}\right)C_k^{(d-2)/2}(\cos\theta),
\end{aligned}
$$

and the Gegenbauer coefficients of $A(\theta,x)$ are obtained from Corollary 5.2.6 for $k\geq 1$, and from the equality $C_k^{(d-2)/2}(x)=1$ when $k=0$. □

**Corollary 5.2.18** (Gegenbauer coefficients of $\psi_{d-1}^W$). *Let $P_{n,d-1}^W\in\mathcal{P}_{\sigma+}\cup\mathcal{P}_{\pm}$. Then the Gegenbauer coefficients of $\psi_{d-1}^W$ are $b_{k,d-1}^W=\int_{-1}^1 a_{k,d-1}^x\,\mathrm{d}W(F_{d-1}(x))$, for $d\geq 2$ and $k\geq 0$.*

*Proof.* The coefficients of $\psi_{d-1}^W$ are trivially deduced from (5.16) and Theorem 5.2.17. □

The Rayleigh [115] and Bingham [19] statistics belong to $\mathcal{S}_1$, the former with $v_{1,d-1} = \delta_{k1}$ and the latter with $v_{2,d-1} = \delta_{k2}$ (the rest of $v_{k,d-1}$'s being null). The next result, consequence of Theorem 5.2.2 and Corollary 5.2.18, identifies, in the circular case, the statistic in $\mathcal{P}_\pm$ equating them and any other statistic in $\mathcal{S}_1$. It also highlights that any $\mathcal{P}_\pm$-statistic is decomposable into a weighted difference of two $\mathcal{P}_+$-statistics.

**Corollary 5.2.19.** *For any circular statistic $S_{n,1}(\{v_{k,1}\}) \in \mathcal{S}_\ell$, with $\ell \geq 1$, the weights*

$$w(\{v_{k,1}\})(x) := \sum_{k=1}^{\ell} v_{k,1}^2 w_k(x) \quad \textit{and} \quad w_k(x) := -2\pi^2 k^2 \cos(2kx\pi), \quad \textit{with } k \geq 1$$

*generates $P_{n,1}^{W(\{v_{k,1}\})} \in \mathcal{P}_\pm$ such that $S_{n,1}(\{v_{k,1}\}) = P_{n,1}^{W(\{v_{k,1}\})}$. Also, there exist $P_{n,1}^{W^+(\{v_{k,1}\})}$, $P_{n,1}^{W^-(\{v_{k,1}\})} \in \mathcal{P}_+$ and $a^+, a^- \geq 0$ such that $P_{n,1}^{W(\{v_{k,1}\})} = a^+ P_{n,1}^{W^+(\{v_{k,1}\})} - a^- P_{n,1}^{W^-(\{v_{k,1}\})}$.*

*Proof.* Since $F_1^{-1}(x) = -\cos(\pi x)$, then

$$w_k(x) = -2\pi^2 k^2 \cos(2kx\pi) = -2\pi^2 k^2 \cos(2k\cos^{-1}(F_1^{-1}(x))) = -2k^2\pi^2 T_{2k}(F_1^{-1}(x)).$$

Consider the signed measure $W_k$ associated to $w_k$. From Corollary 5.2.18, we have

$$\begin{aligned} b_{j,1}^{W_k} &= \int_{-1}^{1} a_{j,1}^x w_k(F_1(x))\, \mathrm{d}F_1(x) \\ &= -2k^2\pi^2 \int_{-1}^{1} \frac{1}{j^2\pi^2}(1 - T_{2j}(x))\, T_{2k}(x)\, \mathrm{d}F_1(x) \\ &= \frac{2k^2}{j^2\pi} \int_{-1}^{1} T_{2j}(x) T_{2k}(x)(1-x^2)^{-1/2}\, \mathrm{d}x \\ &= \delta_{jk}. \end{aligned}$$

Therefore, $w(\{v_{k,1}\})(x) = \sum_{k=1}^{\ell} v_{k,1}^2 w_k(x)$ is such that $b_{j,1}^{W(\{v_{k,1}\})} = v_{k,1}^2$, $k \geq 1$, and $S_{n,1}(\{v_{k,1}\}) = P_{n,1}^{W(\{v_{k,1}\})}$. Defining $\tilde{w}^\pm(\{v_{k,1}\})(x) := \max(\pm w(\{v_{k,1}\})(x), 0)$, then $w(\{v_{k,1}\}) = a^+ w^+(\{v_{k,1}\}) - a^- w^-(\{v_{k,1}\})$, where $(a^\pm)^{-1} := \int_{-1}^{1} \tilde{w}^\pm(\{v_{k,1}\})(x)\, \mathrm{d}x \geq 0$ and $w^\pm(\{v_{k,1}\}) := (a^\pm)^{-1}\tilde{w}(\{v_{k,1}\})$ with associated measure $W^\pm(\{v_{k,1}\})$. Then, $P_{n,1}^{W(\{v_{k,1}\})} = a^+ P_{n,1}^{W^+(\{v_{k,1}\})} - a^- P_{n,1}^{W^-(\{v_{k,1}\})}$. □

The following results provide relatively explicit control on the Gegenbauer coefficients of $\psi_{d-1}^{\mathrm{R}_t}$, $\psi_{d-1}^{\mathrm{CvM}}$, and $\psi_{d-1}^{\mathrm{AD}}$. The first is a direct consequence of Theorem 5.2.17, whereas the latter involve direct computations on the kernel functions. General closed-form expressions for $b_{k,d-1}^{\mathrm{AD}}$ are highly challenging to obtain, except for $d = 2, 3$.

**Corollary 5.2.20** (Gegenbauer coefficients for $\psi_{d-1}^{\mathrm{R}_t}$)**.** *Let $t_m$ defined as in Proposition 5.1.10 and $a_{k,d-1}^x$ as in Theorem 5.2.17. The Gegenbauer coefficients of $\psi_{d-1}^{\mathrm{R}_t}$ for $d \geq 2$ are*

$$b_{k,d-1}^{\mathrm{R}_t} = \begin{cases} \dfrac{2}{k^2\pi^2}\sin^2\left(k\pi t_m\right), & d = 2, \\ a_{k,d-1}^{F_{d-1}^{-1}(t_m)}, & d \geq 3, \end{cases} \qquad b_{0,d-1}^{\mathrm{R}_t} = \frac{1}{2} - t_m(1-t_m), \quad d \geq 2.$$

*Proof.* Assume $d \geq 3$ and $k \geq 1$. Due to the symmetry of $x \mapsto a_{k,d-1}^x$, by Corollary 5.2.18,

$$b_{k,d-1}^{\mathrm{R}_t} = \int_{-1}^{1} a_{k,d-1}^x \,\mathrm{d}W_t(F_{d-1}(x)) = \frac{1}{2}a_{k,d-1}^{F_{d-1}^{-1}(1-t_m)} + \frac{1}{2}a_{k,d-1}^{F_{d-1}^{-1}(t_m)} = a_{k,d-1}^{F_{d-1}^{-1}(t_m)}.$$

The particular case $d = 2$ is obtained from $F_1^{-1}(x) = \cos(\pi(1-x))$ and $\sin^2(x) = (1-\cos(2x))/2$. For $k = 0$, by the definition of Gegenbauer coefficients and Proposition 5.1.12, we have

$$\begin{aligned} b_{0,d-1}^{\mathrm{R}_t} &= \int_{-1}^{1} F_{d-1}(x)^2 \,\mathrm{d}W_t(F_{d-1}(x)) = \tfrac{1}{2}\left((F_{d-1}(F_{d-1}^{-1}(1-t_m)))^2 + (F_{d-1}(F_{d-1}^{-1}(t_m)))^2\right) \\ &= \tfrac{1}{2}\left((1-t_m)^2 + t_m^2\right). \end{aligned}$$ □

**Proposition 5.2.21** (Gegenbauer coefficients for $\psi_{d-1}^{\mathrm{CvM}}$)**.** *The Gegenbauer coefficients of $\psi_{d-1}^{\mathrm{CvM}}$ for $d = 2, 3, 4$ are*

$$b_{k,d-1}^{\mathrm{CvM}} = \begin{cases} \dfrac{1}{\pi^2k^2}, & d = 2, \\ \dfrac{1}{2(2k+3)(2k-1)}, & d = 3, \\ \dfrac{35}{72\pi^2}1_{\{k=1\}} + \dfrac{1}{2\pi^2}\dfrac{3k^2+6k+4}{k^2(k+1)(k+2)^2}1_{\{k>1\}}, & d = 4, \end{cases}$$

*and if $d \geq 3$,*

$$b_{k,d-1}^{\mathrm{CvM}} = \frac{(d-2)^2(2k+d-2)\Gamma\left(\frac{d-2}{2}\right)^3\Gamma\left(\frac{3(d-1)}{2}\right)}{8\pi(d-1)^2\Gamma\left(\frac{d-1}{2}\right)^3\Gamma\left(\frac{3d-2}{2}\right)}{}_4F_3\left(1-k, d-1+k, \tfrac{d}{2}, \tfrac{3(d-1)}{2}; d, \tfrac{d-1}{2}+1, \tfrac{3d-2}{2}; 1\right)$$

*Additionally, $b_{0,d-1} = 1/3$ for $d \geq 2$. Therefore, $b_{k,d-1}^{\mathrm{CvM}} > 0$ for $k \geq 0$ and $d = 2, 3, 4$.*

*Proof.* If $d = 2$, by Corollary 5.2.18, the Gegenbauer coefficients of $\psi_1^{\mathrm{CvM}}$ are, for $k \geq 1$,

$$\begin{aligned} b_{k,1}^{\mathrm{CvM}} &= \int_{-1}^{1} a_{k,1}^x \,\mathrm{d}F_1(x) \\ &= \frac{1}{k^2\pi^3}\int_{-1}^{1}\left(1 - T_{2k}(x)\right)\left(1-x^2\right)^{-1/2}\mathrm{d}x \\ &= \frac{1}{k^2\pi^3}\int_{-1}^{1}\left(1 - \cos\left(2k\cos^{-1}(x)\right)\right)\left(1-x^2\right)^{-1/2}\mathrm{d}x \end{aligned}$$

$$= \frac{1}{\pi^2 k^2}, \tag{5.52}$$

where (5.52) follows from the orthogonality of the Gegenbauer polynomials.

If $d \geq 3$, by Corollary 5.2.18, the Gegenbauer coefficients of $\psi_{d-1}^{\mathrm{CvM}}$ are, for $k \leq 1$,

$$\begin{aligned}
b_{k,d-1}^{\mathrm{CvM}} &= \int_{-1}^{1} a_{k,d-1}^{x} \, \mathrm{d}F_{d-1}(x) \\
&= \left(1 + \frac{2k}{d-2}\right) \left( \frac{2^{d-2}\Gamma\left(d/2\right)^2 \Gamma(k)}{\pi\Gamma(k+d-1)} \right)^2 \frac{1}{\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} \int_{-1}^{1} (1-x^2)^{(3d-5)/2} \left( C_{k-1}^{d/2}(x) \right)^2 \mathrm{d}x \\
&= \left(1 + \frac{2k}{d-2}\right) \left( \frac{2^{d-2}\Gamma\left(d/2\right)^2 \Gamma(k)}{\pi\Gamma(k+d-1)} \right)^2 \frac{1}{\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} \frac{\mathrm{B}\left(\frac{3(d-1)}{2}, \frac{1}{2}\right)}{\left((k-1)\mathrm{B}(d, k-1)\right)^2} \qquad (5.53) \\
&\quad {}_4F_3\left(1-k, d-1+k, \frac{d}{2}, \frac{3(d-1)}{2}; d, \frac{d-1}{2}+1, \frac{3d-2}{2}; 1\right) \\
&= \frac{(d-2)^2(2k+d-2)\Gamma\left(\frac{d-2}{2}\right)^3 \Gamma\left(\frac{3(d-1)}{2}\right)}{32\pi\Gamma\left(\frac{d-1}{2}+1\right)^2 \Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{3d-2}{2}\right)} {}_4F_3\left(1-k, d-1+k, \tfrac{d}{2}, \frac{3(d-1)}{2}; d, \tfrac{d-1}{2}+1, \tfrac{3d-2}{2}; 1\right) \\
&= \frac{(d-2)^2(2k+d-2)\Gamma\left(\frac{d-2}{2}\right)^3 \Gamma\left(\frac{3(d-1)}{2}\right)}{8\pi(d-1)^2\Gamma\left(\frac{d-1}{2}\right)^3 \Gamma\left(\frac{3d-2}{2}\right)} {}_4F_3\left(1-k, d-1+k, \frac{d}{2}, \frac{3(d-1)}{2}; d, \frac{d-1}{2}+1, \frac{3d-2}{2}; 1\right)
\end{aligned} \tag{5.54}$$

where the last equality follows from the Legendre duplication formula and (5.53) follows from equations (16) and (18) in Laursen and Mita [96].

For $d = 3$, by equation (21) in Laursen and Mita [96], we have a special expression of (5.53):

$$\begin{aligned}
b_{k,2}^{\mathrm{CvM}} &= \frac{(1+2k)}{2} \left( \frac{\Gamma(k)}{2\Gamma(k+2)} \right)^2 \frac{4\Gamma(k+2)(2+(k-1)(k+2))}{(k-1)!(2k-1)(2k+1)(2k+3)} \\
&= \frac{1}{2} \frac{\Gamma(k)}{\Gamma(k+2)} \frac{(2+(k-1)(k+2))}{(2k-1)(2k+3)} \\
&= \frac{1}{2(2k-1)(2k+3)}.
\end{aligned}$$

If $d = 4$, (5.54) becomes in

$$b_{k,3}^{\mathrm{CvM}} := \frac{35\,(1+k)}{144\pi^2} {}_4F_3\left(1-k, 3+k, 2, \frac{9}{2}; 4, \frac{5}{2}, 5; 1\right),$$

however this expression is not easily tractable. For that reason, we directly work with the definition of the Gegenbauer coefficients, i.e. using $\psi_3^{\mathrm{CvM}} = \psi_1^{\mathrm{CvM}} + \sin(\theta)(\pi - \theta - \sin(\theta))/((4\pi^2)(1+\cos(\theta)))$. We employ that the Gegenbauer polynomials of order 1 coincide with the Chebyshev polynomials of the second type (equation 18.5.2 in [46]) to obtain the Gegenbauer coefficients of $\psi_3^{\mathrm{CvM}}$ for $k > 1$:

$$
\begin{aligned}
b_{k,3}^{\mathrm{CvM}} &:= \frac{2}{\pi} \int_0^{\pi} \psi_3(\theta) \sin((k+1)\theta) \sin(\theta)\, \mathrm{d}\theta \\
&= \frac{1}{\pi} \int_0^{\pi} \left( 1 + \tfrac{1}{2\pi^2} \left\{ \theta(\theta - 2\pi) + \frac{\sin(\theta)(\pi - \theta - \sin(\theta))}{(1 + \cos(\theta))} \right\} \right) \sin((k+1)\theta) \sin(\theta)\, \mathrm{d}\theta \\
&= \frac{1}{\pi} \bigg( \tfrac{1}{2\pi^2} \int_0^{\pi} (1 - \cos(\theta))(\pi - \theta - \sin(\theta)) \sin((k+1)\theta)\, \mathrm{d}\theta \\
&\qquad + \int_0^{\pi} \sin((k+1)\theta) \sin(\theta)\, \mathrm{d}\theta + \tfrac{1}{2\pi^2} \int_0^{\pi} \theta(\theta - 2\pi) \sin((k+1)\theta) \sin(\theta)\, \mathrm{d}\theta \bigg) \\
&= \frac{1}{\pi} \bigg( \tfrac{1}{8\pi^2} \bigg[ \tfrac{2}{k^2} \sin(k\theta) - \tfrac{2}{k} \left( \tfrac{k^2 \sin(\theta)}{k^2-1} + \theta - \pi \right) \cos(k\theta) + \tfrac{2\cos(\theta)\sin(k\theta)}{k^2-1} - \tfrac{2}{k} \sin(k\theta) \\
&\qquad\qquad + \tfrac{1}{k+1} \sin((k+1)\theta) - \tfrac{4}{(k+1)^2} \sin((k+1)\theta) + \tfrac{2}{k+2} \sin((k+2)\theta) \\
&\qquad\qquad + \tfrac{2}{(k+2)^2} \sin((k+2)\theta) - \tfrac{1}{k+3} \sin((k+3)\theta) - \tfrac{4(\pi-\theta)}{k+1} \cos((k+1)\theta) \\
&\qquad\qquad - \tfrac{2\theta}{k+2} \cos((k+2)\theta) + \tfrac{2\pi}{k+2} \cos((k+2)\theta) \bigg]_0^{\pi} + \tfrac{1}{2} \left[ \tfrac{\sin(k\theta)}{k} - \tfrac{\sin((k+2)\theta)}{k+2} \right]_0^{\pi} \\
&\qquad\qquad + \tfrac{1}{2\pi^2} \bigg[ -\tfrac{\pi}{k^2} \cos(k\theta) - \tfrac{\pi}{k} \theta \sin(k\theta) + \tfrac{\theta \cos(k\theta)}{k^2} + \tfrac{1}{2k^3} (k^2\theta^2 - 2) \sin(k\theta) \\
&\qquad\qquad\qquad - \frac{((k+2)^2\theta^2 - 2) \sin((k+2)\theta) + 2(k+2)\theta \cos((k+2)\theta)}{2(k+2)^3} \\
&\qquad\qquad\qquad + \frac{\pi}{k+2} \theta \sin((k+2)\theta) + \frac{\pi}{(k+2)^2} \cos((k+2)\theta) \bigg]_0^{\pi} \bigg) \\
&= \frac{1}{2\pi^2} \frac{3k^2 + 6k + 4}{k^2(k+1)(k+2)^2}.
\end{aligned}
$$

For $k = 1$, we have

$$
\begin{aligned}
b_{1,3}^{\mathrm{CvM}} &:= \frac{4}{\pi} \int_0^{\pi} \psi_3^{\mathrm{CvM}}(\theta) \sin^2(\theta) \cos(\theta)\, \mathrm{d}\theta \\
&= \frac{1}{\pi^3} \bigg\{ \int_0^{\pi} \theta(\theta - 2\pi) \sin^2(\theta) \cos(\theta)\, \mathrm{d}\theta \\
&\qquad + \int_0^{\pi} (\pi - \theta - \sin(\theta))(1 - \cos(\theta)) \sin(\theta) \cos(\theta))\, \mathrm{d}\theta \bigg\} \\
&= \frac{35}{72\pi^2}.
\end{aligned}
$$

If $k = 0$ and $d \geq 2$, by Corollary 5.2.18, we have that

$$
b_{0,d-1}^{\mathrm{CvM}} = \int_{-1}^{1} F_{d-1}(x)^2 f_{d-1}(x)\, \mathrm{d}x = \frac{1}{3} \left[ F_{d-1}(x)^3 \right]_{-1}^{1} = \frac{1}{3}. \qquad \square
$$

**Proposition 5.2.22** (Gegenbauer coefficients for $\psi_d^{\mathrm{AD}}$)**.** *The Gegenbauer coefficients of* $\psi_1^{\mathrm{AD}}$ *and* $\psi_2^{\mathrm{AD}}$ *are*

$$b_{k,d-1}^{\mathrm{AD}} = \begin{cases} \frac{1}{\pi k^2} \int_0^\pi \frac{1-\cos(2k\theta)}{(\pi-\theta)\theta} \,\mathrm{d}\theta, & d = 2, \\ \frac{1}{k(k+1)}, & d = 3, \end{cases} \qquad b_{0,d-1}^{\mathrm{AD}} = -1, \quad d \geq 2.$$

*Therefore,* $b_{k,d-1}^{\mathrm{AD}} > 0$ *for* $k \geq 1$ *and* $d = 2, 3$.

*Proof.* If $d = 2$ and $k \geq 1$, by Corollary 5.2.18 we have that

$$\begin{aligned} b_{k,1}^{\mathrm{AD}} &= \int_{-1}^1 a_{k,2}^x \frac{f_1(x)}{F_1(x)(1 - F_1(x))} \,\mathrm{d}x \\ &= \frac{1}{k^2\pi} \int_0^\pi (1 - T_{2k}(x)) \frac{(1 - x^2)^{-1/2}}{(\pi - \cos^{-1}(x))(\cos^{-1}(x))^2} \,\mathrm{d}x \\ &= \frac{1}{\pi k^2} \int_0^\pi \frac{1 - \cos(2k\theta)}{(\pi - \theta)\theta} \,\mathrm{d}\theta. \end{aligned}$$

If $d = 3$ and $k \geq 1$, by Corollary 5.2.18 we have that

$$\begin{aligned} b_{k,2}^{\mathrm{AD}} &= \int_{-1}^1 a_{k,2}^x \frac{f_2(x)}{F_2(x)(1 - F_2(x))} \,\mathrm{d}x \\ &= 2\,(1 + 2k) \left( \frac{2\Gamma\,(3/2)^2\,\Gamma(k)}{\pi\Gamma(k + 2)} \right)^2 \int_{-1}^1 (1 - x^2) \left( C_{k-1}^{3/2}(x) \right)^2 \,\mathrm{d}x \\ &= \frac{4}{\pi} \frac{(\Gamma(3/2))^2\,\Gamma(k)}{\Gamma(k + 2)} \qquad (5.55) \\ &= \frac{1}{k(k + 1)} \end{aligned}$$

where (5.55) follows from equation ET II 281(8) in Gradshteyn and Ryzhik [74] with $\nu = 3/2$.

If $k = 0$ and for any $d \geq 2$, from (5.38) we have that

$$\begin{aligned} \psi_{d-1}^{\mathrm{AD}}(\theta) &= \lim_{\varepsilon \to 0} \left( \psi_{d-1}^{\mathrm{AD},\varepsilon}(\theta) + \log\,(1 - F_{d-1}(1 - \varepsilon)) + \log\,(F_{d-1}(1 - \varepsilon)) \right) \\ &= \lim_{\varepsilon \to 0} \left( \int_{-1+\varepsilon}^{1-\varepsilon} \frac{A(\theta,x)}{F_{d-1}(x)(1-F_{d-1}(x))} \mathrm{d}F_{d-1}(x) + \log\,(1-F_{d-1}(1-\varepsilon)) + \log\,(F_{d-1}(1-\varepsilon)) \right), \end{aligned}$$

and consequently by Corollary 5.2.18 we have that

$$\begin{aligned} b_{0,d-1}^{\mathrm{AD}} &= \frac{1}{c_{0,d-1}} \int_0^\pi \psi_{d-1}^{\mathrm{AD}}(\theta) \sin^{d-2}(\theta) \,\mathrm{d}\theta \\ &= \frac{1}{c_{0,d-1}} \int_0^\pi \lim_{\varepsilon \to 0} \left( \int_{-1+\varepsilon}^{1-\varepsilon} \frac{A(\theta, x)}{F_{d-1}(x)(1 - F_{d-1}(x))} \,\mathrm{d}F_{d-1}(x) + \log\,(F_{d-1}(-1 + \varepsilon)) \right) \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad + \log\left(F_{d-1}(1-\varepsilon)\right)\Bigg)\sin^{d-2}(\theta)\,\mathrm{d}\theta \\
&= \lim_{\varepsilon\to 0}\left\{\frac{1}{c_{0,d-1}}\int_0^{\pi}\left(\int_{-1+\varepsilon}^{1-\varepsilon}\frac{A(\theta,x)}{F_{d-1}(x)(1-F_{d-1}(x))}\,\mathrm{d}F_{d-1}(x)\right)\sin^{d-2}(\theta)\,\mathrm{d}\theta\right. \\
&\qquad \left. + \log\left(F_{d-1}(-1+\varepsilon)\right) + \log\left(F_{d-1}(1-\varepsilon)\right)\right\} \qquad (5.56) \\
&= \lim_{\varepsilon\to 0}\left\{\int_{-1+\varepsilon}^{1-\varepsilon}\frac{F_{d-1}(x)^2\mathrm{d}F_{d-1}(x)}{F_{d-1}(x)(1-F_{d-1}(x))} + \log\left(F_{d-1}(-1+\varepsilon)\right) + \log\left(F_{d-1}(1-\varepsilon)\right)\right\} \\
&\qquad\qquad (5.57) \\
&= \lim_{\varepsilon\to 0}\left\{-\left[F_{d-1}(x)\right]_{-1+\varepsilon}^{1-\varepsilon} + 2\log\left(F_{d-1}(1-\varepsilon)\right)\right\} = -1,
\end{aligned}
$$

where (5.57) trivially follows from Fubini's Theorem and (5.56) is obtained from the dominated convergence theorem, which can be applied since $|\psi_{d-1}^{\mathrm{AD}}(\theta)|$ is bounded because by Proposition 5.1.15, we have

$$
\begin{aligned}
|\psi_{d-1}^{\mathrm{AD}}(\theta)| &= \left|-\log(4) + 4\int_0^{\cos(\theta/2)}\log\left(\frac{F_{d-1}(t)}{1-F_{d-1}(t)}\right)\left(1 - F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^2)^{1/2}}\right)\right)\mathrm{d}F_{d-1}(t)\right| \\
&\leq \log(4) + 4\int_0^{\cos(\theta/2)}\left|\log\left(\frac{F_{d-1}(t)}{1-F_{d-1}(t)}\right)\right|\mathrm{d}F_{d-1}(t) \\
&= \log(4) + 4\int_{1/2}^{F_{d-1}(\cos(\theta/2))}\log\left(\frac{u}{1-u}\right)\mathrm{d}u \\
&\leq \log(4) + 4\int_{1/2}^{1}\log\left(\frac{u}{1-u}\right)\mathrm{d}u \\
&= 6\log(2).
\end{aligned}
$$

□

The previous results do not guarantee the omnibussness for $d \geq 2$ of the tests based on $P_{n,d-1}^{\mathrm{CvM}}$ and $P_{n,d-1}^{\mathrm{AD}}$. The next corollary gives a simple sufficient condition, satisfied by these tests, to guarantee omnibussness. It also shows that non-omnibus tests, related with very specific discrete measures $W$, are somehow a rarity in the projected-ecdf class.

**Corollary 5.2.23** (Omnibusness of projected-ecdf tests). *The statistic $P_{n,d-1}^{W} \in \mathcal{P}_{\sigma+}$ generates an omnibus test if and only if $W$ does not concentrate its measure entirely in $F_{d-1}(Z_{d-1})$, $Z_{d-1} := [-1,1] \cap \left(\cup_{k\geq 1} Z_{k,d-1}\right)$, where*

$$
Z_{k,d-1} := \begin{cases} \{\cos(m/k\pi) : m = 0,1,\ldots,k\}, & d = 2, \\ \{-1,1\} \cup \left\{x \in (-1,1) : C_{k-1}^{d/2}(x) = 0\right\}, & d \geq 3. \end{cases}
$$

*In particular, any $W$ that assigns a positive measure to a fixed set of $[0,1]$ with non-zero Lebesgue measure, generates an omnibus projected-ecdf test.*

*Proof.* By Theorem 5.2.17, the $k$-th Gegenbauer coefficient of $\theta \mapsto A(x,\theta)$ is $a_{k,d-1}^x$. Due to the properties of orthogonal polynomials, the function $x \mapsto a_{k,d-1}^x$ has exactly $k-1$ different real zeros in $(-1,1)$ for $d \geq 3$, hence $k+1$ different zeros in $[-1,1]$. For $d=2$, the $k+1$ different roots for $T_{2k}(x)=1$ with $x \in [-1,1]$ are $x_m = \cos(m/k\pi)$, $m = 0,1,\ldots,k$. Then, the sets $Z_{k,d-1}$ have cardinality $k-1$ and $Z_{d-1} \subset [-1,1]$ is an at most denumerable set.

From the previous statements, for any $x^* \in [-1,1] \backslash Z_{d-1}$, $a_{k,d-1}^{x^*} > 0$, for all $k \geq 1$. As a consequence, for any $x^* \in Z_{d-1}$ there is an associated $W_{x^*} := \delta_{F_{d-1}(x^*)}$ whose kernel $\psi_{d-1}^{W_{x^*}}$ has positive coefficients $b_{k,d-1}^{W_{x^*}} = a_{k,d-1}^{x^*}$, therefore generating an omnibus test by Theorem 2.6.3. In addition, any $\sigma$-finite measure $W$ that assigns a positive measure to $[0,1] \backslash F_{d-1}(Z_{d-1})$ generates a test with coefficients $b_{k,d-1}^W = \int_{-1}^1 a_{k,d-1}^x \, \mathrm{d}W(F_{d-1}(x)) > 0$, hence omnibus.

Finally, since $F_{d-1}(Z_{d-1})$ has null Lebesgue measure, trivially every $L \subset [0,1]$ with non-null Lebesgue measure satisfying $L \cap ([0,1] \backslash F_{d-1}(Z_{d-1})) \neq \varnothing$, thus a measure $W$ such that $W(L) > 0$ gives an omnibus test. □

**Remark 5.2.24.** *Since $F_1(Z_1) = \mathbb{Q} \cap [0,1]$, any measure $W$ concentrated on $\mathbb{Q} \cap [0,1]$ generates a non-omnibus projected-ecdf test. This is the case of the $P_{n,1}^{\mathrm{R}_t}$-based test with $t \in \mathbb{Q} \cap [0,1]$ (if $t \in \mathbb{I} \cap [0,1]$, the test is omnibus). For $d \geq 3$, the explicit characterization of $F_{d-1}(Z_{d-1})$ is more cumbersome, since it depends on the zeros of the Gegenbauer polynomials (see, e.g., Theorem 2 in Dimitrov and Nikolov [45] for their lower and upper bounds). Nevertheless, it is easy to see that $\{0, 1/2, 1\} \subset F_{d-1}(Z_{d-1})$, thus, as already known, the $P_{n,d-1}^{\mathrm{R}_{1/2}}$-based test is not omnibus.*

## 5.2.3 Computation of asymptotic tail probabilities

Computing tail probabilities in the asymptotic distribution $\sum_{k=1}^{\infty} w_k Y_k$ in (2.14) is not trivial. [114] provided approximations, for specific choices of $\{w_k\}$, based on inverting the characteristic function [16]. A more general approach is to compute, for a sufficiently large $K$, the tail probability of $\sum_{k=1}^{K} w_k Y_k$ by the fast Hall–Buckley–Eagleson (HBE) approximation (three-moment match to a Gamma distribution, see [23]) or by [82]'s exact method; see [21] for a review of approaches for evaluating the cdf of $\sum_{k=1}^{K} w_k Y_k$.

Our proposal is to use Algorithm 5 for computation of asymptotic tail probabilities.

According to the desired precision, Step *2)* in Algorithm 5 can be omitted taking a reduced $K_{\max}$ since in our empirical investigations we have seen that, for dimensions $2 \leq d \leq 11$ and the previous statistics, such step can be effectively omitted. Indeed, values of $K_{\max} = 10^3, 10^4$ and $5 \times 10^4$ omitting in Step *2)* give a uniform tail probability

**Algorithm 5:** Asymptotic $p$-value computation for a test based on $P^W_{n,d-1} \in \mathcal{P}_{\sigma+}, \mathcal{P}_{\pm}$

```
1) Compute the sequence $\{b^W_{k,d-1}\}$ for $k = 1, \ldots, K_{\max}$ using the most adequate expression in Corollaries 5.2.18 and 5.2.20 or Propositions 5.2.21 and 5.2.22.
2) Reduce $K_{\max}$ to $K_{\delta,x} = \min\{K \geq 1 : |p^{\mathrm{HBE}}_{K_{\max},x} - p^{\mathrm{HBE}}_{K,x}| \leq \delta\}$ for $\delta \in [0,1]$, where $p^{\mathrm{HBE}}_{K,x}$ represents the HBE-approximated tail probability $\mathbb{P}[\sum_{k=1}^{K} (v^W_{k,d-1})^2 Y_k > x]$, $\{v^W_{k,d-1}\}$ stem from $\{b^W_{k,d-1}\}$ using (2.12), and $x$ equals the observed statistic $P^W_{n,d-1}$.
3) Use Imhof [82]'s method to compute $\mathbf{P}\big[\sum_{k=1}^{K_{\delta,x}} \big(v^W_{k,d-1}\big)^2 Y_k > x\big]$.
```

accuracy (in $x \in [0,1]$) of two, three, and four digits, respectively. Due to this, we set $K_{\max} = 5 \times 10^4$ for the application of Algorithm 5 in the next sections.

However, Step *2)* could be useful if we consider $K_{\max} \geq 10^5$ and $\delta = K_{\max}^{-1}$ or if $d$ is larger because the accuracy deteriorates as $d$ increases, e.g., with $d = 51$ a one digit accuracy is lost.

## 5.3 Simulation study

We compare through simulations the empirical performance of three new projected-ecdf tests: CvM, Anderson–Darling (AD), and Rothman with $t = 1/3$ (Rt). We do so by: *i*) evaluating the accuracy of their obtained asymptotic distributions in Section 5.3.1; *ii*) comparing their empirical powers with well-known uniformity tests for $\Omega^{d-1}$, $d \geq 3$, in Section 5.3.2.

We consider $M = 10^6$ Monte Carlo replicates, dimensions $d = 2, 3, 4, 11$, and sample sizes $n = 50, 100, 200$. The implementations of the projected-ecdf test statistics rely on the explicit expressions given in Propositions 5.1.6, 5.1.12, and 5.1.15. Gauss–Legendre quadrature with $160$ nodes is used for approximating the integrals in the kernel functions.

### 5.3.1 Asymptotic null distribution accuracy

Table 5.2 reveals that the exact-$n$ critical values (i.e. the critical values are computed with the Monte Carlo method using a sample of size $n$), approximated by $M$ Monte Carlo replicates, quickly converges to the asymptotic critical values, irrespectively of the investigated dimensions, significance levels or projected-ecdf tests. Table 5.3

Table 5.2.: Exact-$n$ and asymptotic critical values for the significance levels $\alpha = 0.10$, $0.05$, $0.01$ of the CvM, AD, and Rt uniformity tests on $\Omega^{d-1}$, for $d = 2, 3, 4, 11$. The exact-$n$ critical values are approximated by $M$ Monte Carlo replicates, whereas the asymptotic critical values are computed using Algorithm 5.

| Test | $d$ | $\alpha = 0.10$ | | | | $\alpha = 0.05$ | | | | $\alpha = 0.01$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $n = 50$ | $n = 100$ | $n = 200$ | $n = \infty$ | $n = 50$ | $n = 100$ | $n = 200$ | $n = \infty$ | $n = 50$ | $n = 100$ | $n = 200$ | $n = \infty$ |
| CvM | 2 | 0.3025 | 0.3029 | 0.3031 | 0.3035 | 0.3713 | 0.3723 | 0.3731 | 0.3738 | 0.5303 | 0.5336 | 0.5356 | 0.5368 |
| | 3 | 0.2759 | 0.2764 | 0.2769 | 0.2769 | 0.3270 | 0.3277 | 0.3289 | 0.3291 | 0.4412 | 0.4442 | 0.4465 | 0.4469 |
| | 4 | 0.2598 | 0.2605 | 0.2607 | 0.2608 | 0.3010 | 0.3020 | 0.3031 | 0.3029 | 0.3929 | 0.3941 | 0.3960 | 0.3963 |
| | 11 | 0.2202 | 0.2206 | 0.2206 | 0.2208 | 0.2406 | 0.2411 | 0.2412 | 0.2414 | 0.2831 | 0.2844 | 0.2848 | 0.2849 |
| AD | 2 | 1.6824 | 1.6832 | 1.6855 | 1.6875 | 2.0170 | 2.0236 | 2.0272 | 2.0304 | 2.7982 | 2.8122 | 2.8194 | 2.8252 |
| | 3 | 1.5555 | 1.5577 | 1.5613 | 1.5612 | 1.8112 | 1.8156 | 1.8216 | 1.8227 | 2.3856 | 2.3993 | 2.4111 | 2.4122 |
| | 4 | 1.4773 | 1.4805 | 1.4818 | 1.4824 | 1.6869 | 1.6911 | 1.6972 | 1.6961 | 2.1531 | 2.1585 | 2.1690 | 2.1695 |
| | 11 | 1.2780 | 1.2797 | 1.2798 | 1.2810 | 1.3836 | 1.3863 | 1.3863 | 1.3880 | 1.6044 | 1.6100 | 1.6124 | 1.6130 |
| Rt | 2 | 0.4248 | 0.4254 | 0.4259 | 0.4264 | 0.5282 | 0.5297 | 0.5304 | 0.5318 | 0.7666 | 0.7716 | 0.7744 | 0.7764 |
| | 3 | 0.3830 | 0.3837 | 0.3844 | 0.3844 | 0.4585 | 0.4594 | 0.4614 | 0.4617 | 0.6280 | 0.6322 | 0.6355 | 0.6361 |
| | 4 | 0.3584 | 0.3593 | 0.3597 | 0.3598 | 0.4191 | 0.4204 | 0.4220 | 0.4217 | 0.5534 | 0.5556 | 0.5579 | 0.5589 |
| | 11 | 0.2997 | 0.3002 | 0.3002 | 0.3005 | 0.3292 | 0.3299 | 0.3299 | 0.3304 | 0.3907 | 0.3924 | 0.3930 | 0.3933 |

corroborates that the exact-$n$ rejection frequencies remain within the normal 99% confidence interval for $n = 200$ when the asymptotic critical values are employed in the test decision. For $n = 50, 100$, the rejection frequencies are either inside the confidence interval or quite close to the significance level on the conservative side.

Table 5.3.: Rejection frequencies using asymptotic critical values for the significance levels $\alpha = 0.10, 0.05, 0.01$ of the CvM, AD, and Rt uniformity tests on $\Omega^{d-1}$, for $d = 2, 3, 4, 11$. The rejection frequencies are approximated by $M$ Monte Carlo replicates, whereas the asymptotic critical values are computed with Algorithm 5. Boldfaces denote that the rejection rate is within the normal 99% confidence interval for $\alpha$.

| Test | $d$ | $\alpha = 0.10$ | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n = 50$ | $n = 100$ | $n = 200$ | $n = 50$ | $n = 100$ | $n = 200$ | $n = 50$ | $n = 100$ | $n = 200$ |
| CvM | 2 | 0.0990 | **0.0993** | **0.0996** | 0.0488 | 0.0493 | **0.0497** | 0.0094 | 0.0097 | **0.0099** |
| | 3 | 0.0987 | **0.0993** | **0.1000** | 0.0485 | 0.0490 | **0.0499** | 0.0092 | 0.0096 | **0.0099** |
| | 4 | 0.0984 | **0.0995** | **0.0999** | 0.0484 | 0.0492 | **0.0502** | 0.0094 | 0.0096 | **0.0099** |
| | 11 | 0.0982 | **0.0994** | **0.0996** | 0.0485 | **0.0495** | **0.0496** | 0.0093 | **0.0098** | **0.0100** |
| AD | 2 | 0.0989 | 0.0991 | **0.0996** | 0.0486 | 0.0493 | **0.0497** | 0.0095 | **0.0097** | 0.0099 |
| | 3 | 0.0984 | 0.0991 | **0.1000** | 0.0484 | 0.0490 | **0.0498** | 0.0093 | 0.0097 | **0.0100** |
| | 4 | 0.0983 | **0.0994** | **0.0998** | 0.0484 | 0.0492 | **0.0502** | 0.0094 | 0.0096 | **0.0100** |
| | 11 | 0.0980 | **0.0992** | **0.0992** | 0.0486 | 0.0494 | 0.0494 | 0.0094 | **0.0098** | **0.0100** |
| Rt | 2 | 0.0989 | **0.0994** | **0.0996** | 0.0488 | 0.0493 | **0.0495** | 0.0094 | 0.0097 | **0.0099** |
| | 3 | 0.0987 | **0.0993** | **0.1000** | 0.0486 | 0.0490 | **0.0499** | 0.0092 | 0.0096 | **0.0099** |
| | 4 | 0.0985 | **0.0994** | **0.0999** | 0.0485 | 0.0492 | **0.0502** | 0.0093 | 0.0096 | **0.0099** |
| | 11 | 0.0982 | **0.0993** | **0.0993** | 0.0485 | 0.0494 | **0.0495** | 0.0093 | 0.0097 | **0.0099** |

## 5.3.2 Empirical power investigation

We compare the CvM, AD, and Rt projected-ecdf tests with the following classic uniformity tests: Rayleigh (Rayleigh [115]), Bingham (Bingham [19]), Ajne (Ajne [4] and

Prentice [114]), Giné's $G_n$ and $F_n$ (Giné [71] and Prentice [114]), and Cuesta-Albertos *et al.* [37], henceforth abbreviated CCF09. Except the latter (already reviewed in Section 5.1.1), all of them belong to the Sobolev class. The main properties of the competing tests are summarized next (see García-Portugués and Verdebout [64] for a more detailed review):

- All of them are valid for arbitrary dimensions.
- Only Bakshaev and CCF09 are omnibus tests.
- The Rayleigh and Ajne tests are not consistent against axial alternatives (symmetric pdfs with respect to $\mathbf{0}$).
- The Bingham and Giné's $G_n$ tests are designed for axial alternatives, sacrificing power against unimodal alternatives.
- The Rayleigh and Bingham tests are the most powerful rotation invariant tests with respect to von Mises–Fisher and Watson alternatives, respectively.
- The CCF09 test requires simulating $k$ random directions and then performing a Monte Carlo calibration *conditionally* on those random directions. Following the recommendation in CCF09, we considered $k = 50$, then run the simulation study conditionally on a fixed set of $k$ random directions.

We employ six different Data Generating Processes (DGPs). The first three are based on the local alternatives for which a $P^W_{n,d-1}$-based projected-ecdf test is locally asymptotically most powerful rotation-invariant in virtue of Corollary 5.2.16:

$$f^W_{\boldsymbol{\mu},\kappa}(\mathbf{x}) := \frac{1-\kappa}{\omega_{d-1}} + \frac{\kappa f^W(\mathbf{x}'\boldsymbol{\mu})}{\omega_{d-1}}, \quad f^W(z) := 1 + \sum_{k=1}^{\infty} \left(1 + \tfrac{2k}{d-2}\right) (b^W_{k,d-1})^{1/2} C_k^{(d-2)/2}(z). \tag{5.58}$$

Considering $\boldsymbol{\mu} = (\mathbf{0}_{d-1}, 1)$, the first three DGPs use (5.58) with the next coefficients:

**CvM** $\{b^{\mathrm{CvM}}_{k,d-1}\}$ given in Proposition 5.2.21 for $d = 2, 3, 4$, and computed numerically for $d = 11$ using Corollary 5.2.18 and Gauss–Legendre quadrature with $5120$ nodes.

**AD** $\{b^{\mathrm{AD}}_{k,d-1}\}$ given in Proposition 5.2.22 for $d = 2$ and computed numerically for $d \geq 3$ under the previous conditions.

**Rt** $b^{\mathrm{R}_{1/3}}_{k,d-1}$ given in Corollary 5.2.20.

These DGPs can be seen as "unimodal" alternatives: CvM and AD concentrate probability mass about $\boldsymbol{\mu}$, while Rt does so in a constant cap about $\boldsymbol{\mu}$.

The remaining DGPs include the optimally-detected alternatives for the Rayleigh and Bingham tests, and an alternative without an optimal test among the inspected:

**vMF** Von Mises–Fisher pdf $\mathbf{x} \mapsto c^{\mathrm{vMF}}_{d-1,\eta} \exp(\eta \mathbf{x}'\boldsymbol{\mu})$, with $\eta \geq 0$ and $\boldsymbol{\mu} \in \Omega^{d-1}$. We set $\eta = \kappa$ and $\boldsymbol{\mu} = (\mathbf{0}_{d-1}, 1)$.

**Wat** Watson pdf $\mathbf{x} \mapsto c^{\mathrm{W}}_{d-1,\eta} \exp(\eta(\mathbf{x}'\boldsymbol{\mu})^2)$, with $\eta \in \mathbb{R}$ and $\boldsymbol{\mu} \in \Omega^{d-1}$. We set $\eta = 2.5\kappa$ and $\boldsymbol{\mu} = (\mathbf{0}_{d-1}, 1)$.

**SC** Small-circle pdf $\mathbf{x} \mapsto c^{\mathrm{SC}}_{d-1,\tau,\eta} \exp(\eta(\mathbf{x}'\boldsymbol{\mu} - \tau)^2)$, with $\eta \in \mathbb{R}$, $\tau \in [-1, 1]$, and $\boldsymbol{\mu} \in \Omega^{d-1}$. We set $\eta = -1.5\kappa$, $\tau = 0.50$, and $\boldsymbol{\mu} = (\mathbf{0}_{d-1}, 1)$.

The deviation from uniformity is controlled by $\kappa \geq 0$ ($\kappa \leq 1$ necessarily for (5.58)).

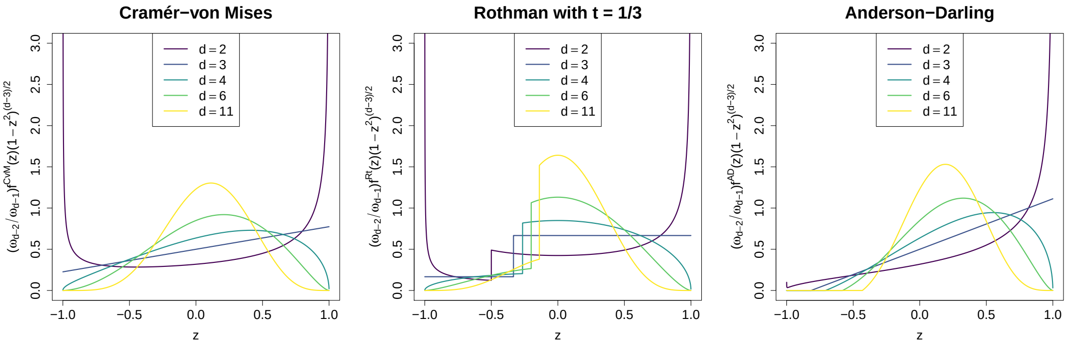


**Figure 5.2.:** From left to right, depiction of the projected densities $z \mapsto \frac{\omega_{d-2}}{\omega_{d-1}} f^W(z)(1-z^2)^{(d-3)/2}$ on $[-1, 1]$ generated by $f^{\mathrm{CvM}}$, $f^{\mathrm{R}_{1/3}}$, and $f^{\mathrm{AD}}$, for $d = 2, 3, 4, 6, 11$. The series in (5.58) is truncated such that 99.9% of the norm in $L^2_{d-1}[-1, 1]$ is retained.

The simulation of all the alternatives was done through the tangent-normal decomposition implemented in the `rotasym` package García-Portugués *et al.* [69] and the (numerical) inversion method to simulate from the pdfs of the projections along $\boldsymbol{\mu}$, which for $f^W$ are $z \mapsto (\omega_{d-2}/\omega_{d-1}) f^W(z)(1-z^2)^{(d-3)/2}$ (see Figure 5.2). The series in $f^W$ is truncated to its first $K_r$ terms explaining $r = 99.95\%$ of the $L^2_{d-1}[-1, 1]$-norm of the series computed with $K_{\max} = 5 \times 10^4$ terms. The simulation from Rt is exact thanks to the closed-form expression $f^{\mathrm{Rt}}(z) = 1_{\{z \geq F^{-1}_{d-1}(t)\}} + t$ and its projected quantile function $F^{-1}_{\mathrm{Rt}}(u) = F^{-1}_{d-1}((u+t)/(1+t)) 1_{\{u > t^2\}} + F^{-1}_{d-1}(u/t) 1_{\{u \leq t^2\}}$. Sampling from CvM for $d = 2$ is simplified due to the closed-form expression $f^{\mathrm{CvM}}(z) = 1 - \sqrt{2} \log(2(1-z))/(2\pi)$.[2]

For the sake of equity, all the tests are calibrated with exact-$n$ critical values approximated by $M$ Monte Carlo replicates at the $\alpha = 0.05$ significance level. Table 5.4 collects the empirical powers for $d = 2, 3, 4, 11$, $n = 100$, and $\kappa = 0.50$. The remaining combinations for $n = 50, 100, 200$ and $\kappa = 0.25, 0.50, 0.75$ are relegated to the Appendix. The following conclusions are extracted from all the tables:

[2] Differs from $f(\theta) = \theta^2/(2\pi^2)$ in Mardia and Jupp [102, page 114] for the Watson test, which is not a circular pdf nor generates the Watson statistic from the book's equation (6.3.70).

(*i*). Overall, the CvM test improves over CCF09. It does so with an average (absolute) power gain equal to $0.0315$ for all DGPs but Wat. The "$[5\%, 95\%]$ Interquantile Range" (IR) of these power gains (taken over all variations of $n$ and $\kappa$, henceforth implicit) is $[0, 0.0930]$. In Wat, the only axial alternative, the IR is $[-0.0703, 0.0085]$. The power gap between CvM and CCF09 stretches with $d$ due to the increasing difficulty of capturing the most $\mathbf{H}_0$-separating directions on $\Omega^{d-1}$ by random sampling.

(*ii*). The AD test performs just slightly better ($0.0005$) than CvM on all DGPs except Wat, where AD notably improves CvM. In Wat, the gains for AD have average $0.0553$ and IR $[0.0002, 0.2331]$. AD test an edge against axial alternatives, dominating the CCF09 test for all the DGPs considered (IR: $[0, 0.1085]$).

(*iii*). The Rt test performs very similarly to CvM in all alternatives except Wat, where it is clearly outperformed by the latter (IR: $[-0.144, 0.002]$).

(*iv*). The Bakshaev test is slightly ($-0.0026$) outperformed by CvM in all alternatives except vMF, where it behaves similarly to the latter ($0.0002$).

(*v*). Unimodal alternatives are well-detected by projected-ecdf tests: (*a*) the CvM, AD, and Rt tests have very similar performance in their corresponding DGPs; (*b*) in the vMF alternative, the optimal Rayleigh test is barely superior to the projected-ecdf tests; e.g., its average power gain with respect to CvM is just $0.002$.

(*vi*). The non-unimodal and non-axial alternative SC is well-detected by the AD test. It outperforms the rest of tests in the majority of situations (especially $n = 100, 200$).

(*vii*). The axial alternative Wat is much harder to detect by projected-ecdf tests. Their performance is consistently below the Bingham test, whose average power gain with respect to AD is $0.253$, a sharp contrast with the situation in the vMF alternative.

(*viii*). Local asymptotic optimalities have very small effect sizes, and many are actually undetected for the settings considered in the study. Indeed, the AD test is among the most powerful tests in most of the CvM and Rt alternatives. This result can be explained by several factors: (*a*) the very small effect sizes of local optimalities; (*b*) the numerical inaccuracy on sampling exactly (5.58); (*c*) the limitation of the explored values for $\kappa$ and $n$; (*d*) the Monte Carlo noise.

(*ix*). The Rayleigh and Ajne tests perform really similarly. So do the Bingham and Giné's $G_n$ tests. Even in the vMF and Wat alternatives, where Rayleigh and Bingham are respectively the optimal tests, the power difference is minimal.

(*x*). Overall, the qualitative behaviour of all the tests except Bingham and Giné's $G_n$ is highly similar, a reflection of (*v*) and (*vi*).

(*xi*). Though all alternatives are harder to detect when $d$ increases, SC and especially Wat have the larger power dropouts in their optimal tests.

Based on the previous conclusions, we regard the AD test as a reference test of uniformity on $\Omega^{d-1}$ due to its omnibussness, great performance against unimodal alternatives, and relative robustness against non-unimodal alternatives.

We conclude by pointing that (*viii*) and (*ix*) may seem surprising, yet they had been partially reported in the literature. Related with (*viii*), Stephens [126] studied the powers of Ajne and Watson tests under the Rt alternative with $t = 1/2$ and different values of $\kappa$, finding that Ajne was only barely more powerful (see his Table 3). With respect to (*ix*), Figueiredo and Gomes [60] compared the Bingham and Giné's $G_n$ tests under different dimensions, sample sizes, and concentrations, finding no remarkable differences between them (see their Table 3). Figueiredo [59] conducted a similar analysis for the Rayleigh and Ajne tests with identical conclusions (see her Tables 2–4). A simulation experiment in the Appendix gives insights about (*viii*).

Table 5.4.: Empirical powers for the investigated uniformity tests on $\Omega_1$ for $n = 100$ and $\kappa = 0.50$. Boldfaces indicate the tests whose empirical powers are *not* significantly smaller than the largest empirical power for each row, according to a McNemar's exact one-sided test Fay [55] performed at $5\%$ significance level.

| DGP | $d$ | Rayleigh | Bingham | Ajne | Giné | CCF09 | Bakshaev | CvM | AD | Rt |
|---|---|---|---|---|---|---|---|---|---|---|
| CvM | 2 | 0.2773 | 0.1004 | 0.2793 | 0.1021 | 0.2653 | 0.2879 | 0.2897 | **0.2918** | 0.2879 |
| | 3 | 0.2367 | 0.0794 | 0.2377 | 0.0801 | 0.2083 | 0.2424 | 0.2424 | **0.2434** | 0.2414 |
| | 4 | 0.2064 | 0.0692 | 0.2068 | 0.0696 | 0.1839 | 0.2098 | 0.2095 | **0.2104** | 0.2087 |
| | 11 | 0.1326 | 0.0554 | **0.1327** | 0.0554 | 0.1049 | **0.1328** | **0.1328** | **0.1328** | **0.1328** |
| AD | 2 | 0.9002 | 0.4463 | 0.9071 | 0.4776 | 0.9019 | 0.9234 | 0.9271 | **0.9319** | 0.9225 |
| | 3 | 0.8507 | 0.3110 | 0.8542 | 0.3213 | 0.8037 | 0.8670 | 0.8670 | **0.8710** | 0.8637 |
| | 4 | 0.8092 | 0.2378 | 0.8113 | 0.2426 | 0.7497 | 0.8218 | 0.8209 | **0.8241** | 0.8180 |
| | 11 | 0.6201 | 0.1021 | 0.6204 | 0.1023 | 0.4625 | 0.6241 | 0.6234 | **0.6246** | 0.6224 |
| Rt | 2 | 0.4020 | 0.1282 | 0.4014 | 0.1303 | 0.4124 | 0.4175 | 0.4203 | **0.4225** | 0.4213 |
| | 3 | 0.3362 | 0.0820 | 0.3360 | 0.0833 | 0.3127 | **0.3413** | **0.3413** | **0.3414** | **0.3413** |
| | 4 | 0.2902 | 0.0686 | 0.2903 | 0.0694 | 0.2756 | **0.2924** | **0.2924** | **0.2923** | **0.2922** |
| | 11 | **0.1744** | 0.0543 | **0.1745** | 0.0544 | 0.1410 | 0.1742 | 0.1742 | 0.1740 | **0.1744** |
| vMF | 2 | **0.8867** | 0.0634 | 0.8859 | 0.0634 | 0.8451 | 0.8837 | 0.8816 | 0.8769 | 0.8823 |
| | 3 | **0.6648** | 0.0558 | 0.6638 | 0.0556 | 0.5895 | 0.6606 | 0.6606 | 0.6560 | 0.6628 |
| | 4 | **0.4806** | 0.0530 | 0.4798 | 0.0532 | 0.4247 | 0.4774 | 0.4779 | 0.4749 | 0.4791 |
| | 11 | **0.1265** | 0.0503 | **0.1264** | 0.0503 | 0.1031 | 0.1261 | 0.1262 | 0.1259 | **0.1264** |
| SC | 2 | 0.9843 | 0.3243 | 0.9841 | 0.3212 | 0.9891 | 0.9910 | 0.9918 | **0.9922** | 0.9919 |
| | 3 | 0.8716 | 0.1489 | 0.8711 | 0.1480 | 0.8545 | 0.8858 | 0.8858 | **0.8887** | 0.8836 |
| | 4 | 0.7111 | 0.0933 | 0.7104 | 0.0933 | 0.6767 | 0.7192 | 0.7187 | **0.7202** | 0.7170 |
| | 11 | **0.2092** | 0.0526 | **0.2092** | 0.0525 | 0.1612 | 0.2086 | 0.2088 | 0.2083 | **0.2090** |
| W | 2 | 0.0538 | **0.9785** | 0.0547 | 0.9773 | 0.5358 | 0.3560 | 0.4916 | 0.6396 | 0.4946 |
| | 3 | 0.0536 | **0.8850** | 0.0543 | 0.8823 | 0.2031 | 0.1643 | 0.1643 | 0.2570 | 0.1265 |
| | 4 | 0.0526 | **0.6728** | 0.0529 | 0.6698 | 0.1065 | 0.0980 | 0.0916 | 0.1216 | 0.0769 |
| | 11 | 0.0504 | **0.0832** | 0.0505 | **0.0831** | 0.0515 | 0.0523 | 0.0518 | 0.0526 | 0.0512 |

## 5.4 Real data applications

We illustrate the practical relevance of the proposed tests with three real data applications in astronomy. The first two build on previous applications in $\Omega^1$ and $\Omega^2$, while the third is a novel case study. The end-to-end reproduction of the three applications is possible trough the `sphunif` package García-Portugués and Verdebout [65]. The asymptotic $p$-values were computed using Algorithm 5.

### 5.4.1 Sunspots

Sunspots are darker regions of the Sun generated by local concentrations of the solar magnetic field. They appear in a rotationally symmetric fashion emerging due to the wrapping of the field by the Sun's differential rotation Babcock [10]. As this wrapping advances, sunspots progressively span at lower latitudes until approximately 11 years, when the field reverses its polarity and wrapping is restarted, constituting a *solar cycle*. Non-rotationally symmetric patterns may be triggered by "preferred zones of occurrence" where sunspots had originated previously Babcock [10, pages 574 and 581].

The significance of non-rotationally symmetric patterns was investigated in García-Portugués *et al.* [68] using processed data from the Debrecen photoheliographic sunspot catalogue Baranyi *et al.* [12] and Győri *et al.* [75]. Their analysis considered tests for rotational symmetry that inspect the circular uniformity of the longitudes of sunspots with respect to an axis $\boldsymbol{\theta}$. However, due to the non-omnibusness of the tests employed in their analysis, non-rotationally symmetric deviations for which the tests are not consistent may have been undetected.

To further investigate the rotational symmetry of sunspots, we applied the CvM, AD, and Rt tests to the longitudes about the north pole $\boldsymbol{\theta} = (0, 0, 1)'$ of the $5373$ sunspots observed in the cycle 23 (1996–2008), obtaining the asymptotic $p$-values $0.3595$, $0.8393$, and $0.3285$, respectively. We repeated the analysis for the cycle 22 (1986–1996; $4551$ sunspots), obtaining the asymptotic $p$-values $0.0067$, $0.0139$, and $0.0091$. The outcomes of the analysis are coherent with those in García-Portugués *et al.* [68], where the $p$-values of a non-omnibus test for rotational symmetry about $\boldsymbol{\theta}$ are $0.2710$ and $0.0103$ for the cycles 23 and 22, respectively. Therefore, our analysis shows that these outcomes hold when omnibus tests are used and highlights the varying behaviour of different cycles.

### 5.4.2 Long-period comets

Orbits of celestial bodies, such as planets and comets, have attracted scientists' attention for a long time. [17] already discussed whether the clustering of the planets' orbits about the ecliptic, nowadays explained by their origin in the protoplanetary disk, could have happened "by chance". The study of comet orbits has been more intricate. Long-period comets (with periods larger than $200$ years) are thought to arise from the roughly spherical Oort cloud, containing icy planetesimals that were ejected from protoplanetary disks by giant planets. These icy planetesimals became heliocentric comets when their orbits were affected by random perturbations of passing stars and the galactic tide (see, e.g., Sections 5 and 7.2 in Dones *et al.* [47] and references therein). This conjectured past of the Oort cloud explains the nearly isotropic distribution of long-period comets [evidenced, e.g., in 140], sharply contrasting with the ecliptic-clustered orbits of short-period comets originating in the flattened Kuiper Belt.

As illustrated in [137] and [90], assessing the uniformity of orbits can be formalized as testing the uniformity on the sphere of their directed unit normal vectors. An orbit with *inclination* $i \in [0, \pi]$ and *longitude of the ascending node* $\Omega \in [0, 2\pi)$ (see [90]) has directed normal vector $(\sin(i)\sin(\Omega), -\sin(i)\cos(\Omega), \cos(i))'$ to the orbit's plane. The sign of the vector reflects if the orbit is prograde or retrograde.

We applied the CvM, AD, and Rt tests to revisit [137]'s testing of the uniformity of the planets' orbits with updated measurements on $(i, \Omega)$. Unsurprisingly, uniformity is rejected with null Monte Carlo $p$-values. More interesting is the analysis of long-period comets, for which we:

*i*) considered the $208$ long-period elliptic-type single-apparition comets, as of 7th of December 2007, used in Cuesta-Albertos *et al.* [37].

*ii*) performed the same search in [37], restricted to comets with distinct $(i, \Omega)$ up to the second digit, obtaining $438$ comets as of 7th May 2020. The source of both datasets is the JPL Small-Body Database Search Engine (https://ssd.jpl.nasa.gov/sbdb_query.cgi).

The dynamic nature of the database, with additions of first-ever observed comets and updates on the data for former comets, generated the noticeable differences between (*i*) and (*ii*).

In (*i*), the asymptotic $p$-values for the CvM, AD, and Rt tests are, respectively, $0.1011$, $0.0744$, and $0.1207$. Therefore, uniformity is not rejected at significance level $5\%$ and the outcome is in agreement with the analysis in [37]. In (*ii*), however, the same tests gave asymptotic $p$-values $0.0041$, $0.0023$, and $0.0052$. Therefore, contrarily to the analysis

Table 5.5.: Asymptotic $p$-values of the CvM, AD, and Rt tests when applied to the crater locations of the planets and moons with more than 30 IUA-named craters.

| Class | Name | Craters | CvM | AD | Rt |
|---|---|---|---|---|---|
| **Planets** | Mars | 1127 | 0 | 0 | $1 \cdot 10^{-8}$ |
| | Venus | 881 | 0.2726 | 0.2749 | 0.2806 |
| | Mercury | 409 | 0 | 0 | 0 |
| *Dwarf* | Ceres | 115 | 0.0133 | 0.0127 | 0.0150 |
| **Moons** | Moon | 1578 | 0 | $1 \cdot 10^{-8}$ | $1 \cdot 10^{-8}$ |
| | Callisto | 141 | 0 | 0 | $3 \cdot 10^{-8}$ |
| | Ganymede | 129 | 0.0132 | 0.0087 | 0.0184 |
| | Europa | 41 | 0.0010 | 0.0009 | 0.0010 |
| *Saturn's moons* | Rhea | 128 | 0.2793 | 0.2954 | 0.2705 |
| | Dione | 73 | 0.5195 | 0.4989 | 0.5418 |
| | Iapetus | 58 | 0.0034 | 0.0037 | 0.0032 |
| | Enceladus | 53 | $1 \cdot 10^{-7}$ | $2 \cdot 10^{-8}$ | $5 \cdot 10^{-7}$ |
| | Tethys | 50 | 0.7910 | 0.8425 | 0.7199 |
| | Mimas | 35 | 0.1701 | 0.1704 | 0.1754 |

in ($i$), significant non-uniformity is detected in the orbits of long-period comets with updated records. The observational bias of long-period comets, as described in [90], may explain the leading rejection cause.

### 5.4.3 Craters on Rhea

Craters are roughly circular depressions resulting from impact or volcanic activity. Impact craters give valuable insights on the planetary subsurface structure, past geologic processes, resurfacing history, and relative surfaces ages Barlow [13]. Indeed, crater counting is the primary method for determining remotely the relative age of a planetary surface; see Fassett [54] for a review on crater statistics and their applications.

Short-period comets, especially dominant of the cratering process in the outer Solar System, are among the main generators of non-isotropic impact cratering (see Zahnle *et al.* [141] and references therein). To evaluate the rareness of uniform crater distributions in the Solar System, we analysed the *named* craters from the Gazetteer of Planetary Nomenclature database (`https://planetarynames.wr.usgs.gov/AdvancedSearch`) of the International Astronomical Union (IUA). As of May 31st 2020, the database contained $5235$ craters for $44$ bodies. Filtering for non-asteroid bodies with at least $30$ craters results in $4818$ observations on $\Omega^2$ containing the planetocentric coordinates of the craters' centers. Table 5.5 reveals that, for this dataset, crater uniformity is rejected at significance level $5\%$ in all bodies except Venus and four Saturnian moons. These few non-rejections, however, are suspected to be driven by a uniformity bias in the data: well-separated craters that cover the body are likely more probable to be named than those that cluster (see Figure 5.3). Bypassing this source limitation requires from

Table 5.6.: Asymptotic $p$-values of the CvM, AD, and Rt tests when applied to [78]'s Rhea crater database.

| Diameter | Craters | CvM | AD | Rt |
|---|---|---|---|---|
| $15\text{km} < D < 20\text{km}$ | 867 | 0.1176 | 0.0721 | 0.1856 |
| $D > 20\text{km}$ | 1373 | $2 \cdot 10^{-9}$ | 0 | $3 \cdot 10^{-9}$ |
| $D > 15\text{km}$ | 2240 | $2 \cdot 10^{-8}$ | 0 | $3 \cdot 10^{-8}$ |

detailed crater databases, available only for certain bodies such as Venus (see, e.g., the analysis in García-Portugués *et al.* [70]) and Rhea.

We investigate in detail the crater distribution of Rhea, the second most cratered body in Table 5.5 with a uniform-like distribution. Rhea orbits Saturn synchronously, thus it has a *leading hemisphere* that always faces forward into the orbit motion and a *trailing hemisphere* that faces backward (see Figure 5.3). Preferred cratering on the leading hemisphere is expected from heliocentric impactors, whereas planetocentric impactors weakly favour the centers of the leading and trailing hemispheres, referred to as *apex* and *antapex*, respectively (see Hirata [78] and references therein). Both populations of impactors may therefore induce a non-uniform crater distribution. Hirata [78] found apex-antapex asymmetry for large craters (diameter $D$ larger than 20km) and no apparent apex-antapex asymmetry in small craters ($15\text{km} < D < 20\text{km}$). We assess the significance of these findings, for the stronger hypothesis of uniformity, from his database of 2440 craters with $D > 15\text{km}$. (The full database contains 3596 craters, but [78]'s analysis only considers those with $D > 15\text{km}$ as the detection of almost all craters above this diameter threshold is guaranteed from the available imagery of Rhea.)

The tests in Table 5.6 reveal that uniformity:

*i*) is not rejected for small craters ($15\text{km} < D < 20\text{km}$) at significance level 5%.

*ii*) is emphatically rejected for large craters ($D > 20\text{km}$); (*iii*) is emphatically rejected for all reliable-detected craters ($D > 15\text{km}$).

The non-rejection in (*i*) may be attributed to Rhea's "crater saturation" Squyres *et al.* [124] or to the dominance of planetocentric impactors Hirata [78], as the largest craters generated by the debris ejected from large crater impacts is $D \approx 20\text{km}$ Alvarellos *et al.* [5]. In turn, the rejections in (*ii*) and (*iii*) may be explained by the predominantly heliocentric origins of the impactors associated to large craters Hirata [78].

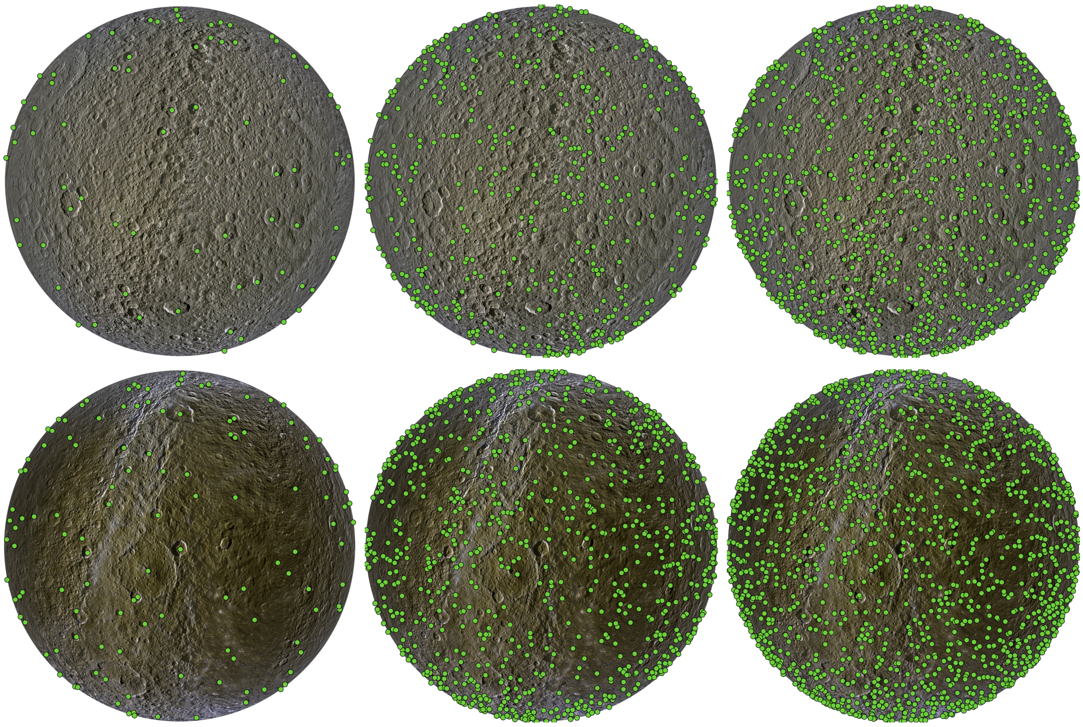


Figure 5.3.: Craters on Rhea. The upper (lower) row shows the leading (trailing) hemisphere. The north and south poles on each hemisphere correspond to the usual top and bottom positions. From left to right, the columns represent the locations of craters for the IUA-named database, and the [78] database for $15\text{km} < D < 20\text{km}$ and $D > 20$km, respectively. The locations are superimposed over the PIA 18438 map produced by NASA/JPL-Caltech/Space Science Institute/Lunar and Planetary Institute using data from the Cassini spacecraft.

# 6 Future work

This research leaves several open problems, some of them were posed from its beginning and other have been arisen through its development. We summarise now some of these interesting issues that remain open for future research. We divide them according to the topic that they tackle:

- Open problems linked to outlier detection:

  a) In this work, we focus on high-dimensional outlier analysis (see Chapters 3 and 4), but it still remains the extension to the infinite-dimensional scenario which seems completely feasible.

  b) Extension to other families of distributions (such as generalized elliptical families) and even to the non-parametric case. In those cases the unidimensional procedure could be, for instance, based on *bootstrap* or on *kernel density functions*.

- Other applications of sequential analysis:

  a) As we exposed in the Introduction, the sequential analysis can be applied to other type of problems with the possibility of developing a general theory which can include different tests (even in the functional case). These problems should have some type of linear invariance such as those proposed in [32] to [39] (excluding Cuesta-Albertos *et al.* [35]).

- Open problems related to uniformity tests on the hypersphere:

  a) A very clear alternative research direction is to proceed *à la* Escanciano [50] and replace $\nu_{d-1}$ by $F_{n,\gamma}$ in (5.2). This approach replaces the analytically challenging integration on $\Omega^{d-1}$ by a sum of $n$ addends, thought, it would have less explicit connection with dimension-specific tests. Moreover if we suppose that the calculations are made with the exact expressions of the statistics (as in (5.15)) and fix $d$, we would replace a test with complexity $O(n^2)$ by one with $O(n^3)$. However, if we admit that the expressions of the statistics can be approximated, the complexities will depend on how these approximations are made.

b) Another alternative is to replace the CvM norm in (5.3) by the “3-point CvM statistic” of Feltz and Goldin [58].

c) As we mentioned in Subsection 5.1.2, Proposition 5.1.5 allows to construct new uniformity tests based on determined choices of $W$. For instance, the consideration of $W_{a,b}(x) := \mathrm{I}_x(a, b)$, where $a, b$ are strictly positive constants, generates a flexible and fairly general two-parameter family of uniformity tests on $\Omega^{d-1}$, although with somehow challenging forms for $\psi_{d-1}^W$.

   This is just one among the many $W$-specific instances of $P_{n,d-1}^W$ that could be constructed. Moreover, the extension to higher dimensions of other well-known tests could be studied.

d) Goodness-of-fit testing of non-uniform distributions on $\Omega^{d-1}$ is possible, yet challenging, by determining the proper substitute for $F_{d-1}$ in (5.2).

e) Bakshaev [11] proposed a uniformity test for which he only obtained the asymptotic distribution in dimensions $2$ and $3$ (and not explicitly). The fact that this test belongs to the family introduced here (see Remark 5.1.8) makes us consider the possibility of obtaining an explicit expression for this distribution in any dimension.

# Bibliography

[1]D. Achlioptas, “Database-friendly random projections: Johnson-Lindenstrauss with binary coins”, *J. Comput. System Sci.*, vol. 66, no. 4, pp. 671–687, 2003 (cit. on p. 12).

[2]J. Adrover and V. J. Yohai, “Projection estimates of multivariate location”, *Ann. Statist.*, vol. 30, no. 6, pp. 1760–1781, 2002 (cit. on p. 17).

[3]C. C. Aggarwal, *Outlier Analysis*. New York: Springer, 2017 (cit. on p. 16).

[4]B. Ajne, “A simple test for uniformity of a circular distribution”, *Biometrika*, vol. 55, no. 2, pp. 343–354, 1968 (cit. on pp. xxv, 7, 26, 105, 109, 113, 133, 163).

[5]J. L. Alvarellos, K. J. Zahnle, A. R. Dobrovolskis, and P. Hamill, “Fates of satellite ejecta in the Saturn system”, *Icarus*, vol. 178, no. 1, pp. 104–123, 2005 (cit. on p. 141).

[6]T. W. Anderson, “The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities”, *Proc. Amer. Math. Soc.*, vol. 6, no. 2, pp. 170–176, 1955 (cit. on p. 70).

[7]T. W. Anderson and D. A. Darling, “Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes”, *Ann. Math. Statistics*, vol. 23, no. 2, pp. 193–212, 1952 (cit. on pp. xxiv, 95).

[8]——, “A test of goodness of fit”, *J. Amer. Statist. Assoc.*, vol. 49, no. 268, pp. 765–769, 1954 (cit. on p. 110).

[9]R. I. Arriaga and S. Vempala, “An algorithmic theory of learning: Robust concepts and random projection”, *Mach. Learn.*, vol. 63, pp. 161–182, 2006 (cit. on pp. xvii, 1).

[10]H. W. Babcock, “The topology of the Sun’s magnetic field and the 22-year cycle”, *Astrophys. J.*, vol. 133, no. 2, pp. 572–587, 1961 (cit. on p. 138).

[11]A. Bakshaev, “$N$-distance tests of uniformity on the hypersphere”, *Nonlinear Anal. Model. Control*, vol. 15, no. 1, pp. 15–28, 2010 (cit. on pp. xxvii, 26, 103, 105, 113, 144).

[12]T. Baranyi, L. Győri, and A. Ludmány, “On-line tools for solar data compiled at the Debrecen observatory and their extensions with the Greenwich sunspot data”, *Sol. Phys.*, vol. 291, no. 9, pp. 3081–3102, 2016 (cit. on p. 138).

[13]N. G. Barlow, “Constraining geologic properties and processes through the use of impact craters”, *Geomorphology*, vol. 240, pp. 18–33, 2015 (cit. on p. 140).

[14]V. Barnett and T. Lewis, *Outliers in Statistical Data*, ser. Wiley Series in Probability and Statistics. Chichester: Wiley, 1994 (cit. on pp. xx, 4, 16).

[15] C. Becker and U. Gather, "The masking breakdown point of multivariate outlier identification rules", *J. Amer. Statist. Assoc.*, vol. 94, no. 447, pp. 947–955, 1999 (cit. on pp. 14, 16, 72).

[16] R. J. Beran, "Testing for uniformity on a compact homogeneous space", *J. Appl. Probab*, vol. 5, no. 1, pp. 177–195, 1968 (cit. on pp. 22, 23, 131).

[17] D. Bernoulli, "Quelle est la cause physique de l'inclinaison des plans des orbites des planètes par rapport au plan de l'équateur de la révolution du soleil autour de son axe; et d'où vient que les inclinaisons de ces orbites sont différentes en elles", in *Recueil des pièces qui ont remporté le prix de l'Académie Royale des Sciences*, A. R. des Sciences, Ed., vol. 3, Paris: Académie Royale des Sciences, 1735, pp. 93–122 (cit. on p. 139).

[18] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices", *Ann. Statist.*, vol. 36, no. 1, pp. 199–227, 2008 (cit. on p. 18).

[19] C. Bingham, "An antipodally symmetric distribution on the sphere", *Ann. Statist.*, vol. 2, no. 6, pp. 1201–1225, 1974 (cit. on pp. 26, 125, 133).

[20] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data", in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: Association for Computing Machinery, 2001, pp. 245–250 (cit. on pp. xvii, 1).

[21] D. A. Bodenham and N. M. Adams, "A comparison of efficient approximations for a weighted sum of chi-squared random variables", *Stat. Comput.*, vol. 26, no. 4, pp. 917–928, 2016 (cit. on p. 131).

[22] G. E. Box, A. Luceño, and M. C. Paniagua-Quinones, *Statistical Control by Monitoring and Adjustment*. Hoboken, New Jersey: John Wiley & Sons, 2009 (cit. on p. 28).

[23] M. Buckley and G. Eagleson, "An approximation to the distribution of quadratic forms in normal random variables", *Aust. N. Z. J. Stat.*, vol. 30A, no. 1, pp. 150–159, 1988 (cit. on p. 131).

[24] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation", *J. Amer. Statist. Assoc.*, vol. 106, no. 494, pp. 672–684, 2011 (cit. on p. 18).

[25] T. Cai, J. Fan, and T. Jiang, "Distributions of angles in random packing on spheres", *J. Mach. Learn. Res.*, vol. 14, pp. 1837–1864, 2013 (cit. on pp. xxiii, xxiv, 7).

[26] H. Cardot, A. Mas, and P. Sarda, "CLT in functional linear regression models", *Probab. Theory and Related Fields*, vol. 138, no. 3-4, pp. 325–361, 2007 (cit. on p. 19).

[27] A. Cerioli, "Multivariate outlier detection with high-breakdown estimators", *J. Amer. Statist. Assoc.*, vol. 105, no. 489, pp. 147–156, 2010 (cit. on pp. 14, 17).

[28] A. Cerioli, M. Riani, and A. C. Atkinson, "Controlling the size of multivariate outlier tests with the mcd estimator of scatter", *Stat. Comput.*, vol. 19, no. 3, pp. 341–353, 2009 (cit. on pp. 14, 72).

[29] S. Chang, P. C. Cosman, and L. B. Milstein, "Chernoff-type bounds for the gaussian error function", *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 2939–2944, 2011 (cit. on p. 48).

[30]H. Cramér and H. Wold, "Some theorems on distribution functions", *J. Lond. Math. Soc.*, vol. 11, no. 4, pp. 290–294, 1936 (cit. on pp. xvii, 1).

[31]M. Csörgő, *Quantile Processes with Statistical Applications*. Philadelphia, Pennsylvania: SIAM, 1983, vol. 42 (cit. on p. 74).

[32]J. A. Cuesta-Albertos and M. Febrero-Bande, "A simple multiway anova for functional data", *TEST*, vol. 19, no. 3, pp. 537–557, 2010 (cit. on pp. xix, xxvi, 2, 143).

[33]J. A. Cuesta-Albertos and A. Nieto-Reyes, "The random Tukey depth", *Comput. Statist. Data Anal.*, vol. 52, no. 11, pp. 4979–4988, 2008 (cit. on pp. xix, 2).

[34]J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford, "Random projections and goodness-of-fit tests in infinite-dimensional spaces", *Bull. Braz. Math. Soc.*, vol. 37, no. 4, pp. 477–501, 2006 (cit. on pp. xix, 2).

[35]——, "A sharp form of the Cramér–Wold theorem", *J. Theor. Probab.*, vol. 20, no. 2, pp. 201–209, 2007 (cit. on pp. xvii, xxvi, 1, 11, 12, 143).

[36]J. A. Cuesta-Albertos, E. del Barrio, R. Fraiman, and C. Matrán, "The random projection method in goodness of fit for functional data", *Comput. Statist. Data Anal.*, vol. 51, no. 10, pp. 4814–4831, 2007 (cit. on pp. xviii, xix, 1, 2).

[37]J. A. Cuesta-Albertos, A. Cuevas, and R. Fraiman, "On projection-based tests for directional and compositional data", *Stat. Comput.*, vol. 19, no. 4, pp. 367–380, 2009 (cit. on pp. xxv, 3, 27, 95, 96, 134, 139).

[38]J. A. Cuesta-Albertos, F. Gamboa, and A. Nieto-Reyes, "A random-projection based procedure to test if a stationary process is gaussian", *Comput. Statist. Data Anal.*, vol. 75, pp. 124–141, 2014 (cit. on pp. xix, 2).

[39]J. A. Cuesta-Albertos, E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga, "Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes", *Ann. Stat.*, vol. 47, no. 1, pp. 439–467, 2019 (cit. on pp. xix, xxvi, 2, 3, 143).

[40]R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-fit techniques*, ser. Statistics: Textbooks and Monographs. New York: Marcel Dekker, Inc., 1986, vol. 68 (cit. on p. 96).

[41]F. Dai and Y. Xu, *Approximation theory and harmonic analysis on spheres and balls*, ser. Springer Monographs in Mathematics. New York: Springer, 2013 (cit. on p. 22).

[42]S. Dasgupta and A. Gupta, "An elementary proof of the Johnson-Lindenstrauss lemma", *International Computer Science Institute, Technical Report*, vol. 22, no. 1, pp. 1–5, 1999 (cit. on pp. 11, 12).

[43]L. Davies, "The asymptotics of Rousseeuw's minimum volume ellipsoid estimator", *Ann. Statist.*, vol. 20, no. 4, pp. 1828–1843, 1992 (cit. on p. 16).

[44]L. Davies and U. Gather, "The identification of multiple outliers", *J. Amer. Statist. Assoc.*, vol. 88, no. 423, pp. 782–792, 1993 (cit. on p. 14).

[45] D. K. Dimitrov and G. P. Nikolov, “Sharp bounds for the extreme zeros of classical orthogonal polynomials”, *J. Approx. Theory*, vol. 162, no. 10, pp. 1793–1804, 2010 (cit. on p. 131).

[46] DLMF, *NIST Digital Library of Mathematical Functions*, http://dlmf.nist.gov/, Release 1.0.27 of 2020-06-15, F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds., 2020 (cit. on pp. 23, 115, 128).

[47] L. Dones, R. Brasser, N. Kaib, and H. Rickman, “Origin and evolution of the cometary reservoirs”, *Space Sci. Rev*, vol. 197, no. 1, pp. 191–269, 2015 (cit. on p. 139).

[48] D. L. Donoho, “Breakdown properties of multivariate location estimators”, *Ph.D. qualifying paper*, 1982 (cit. on p. 19).

[49] K. Esbensen, D. Guyot, F. Westad, and L. Houmoller, *Multivariate Data Analysis in Practice: an Introduction to Multivariate Data Analysis and Experimental Design*. Oslo: CAMO, 2002 (cit. on p. 93).

[50] J. C. Escanciano, “A consistent diagnostic test for regression models using projections”, *Econ. Theory*, vol. 22, no. 6, pp. 1030–1051, 2006 (cit. on pp. xix, xxvi, 3, 143).

[51] K.-T. Fang and Y. Wang, *Number Theoretic Methods in Statistics*. London: Chapman and Hall, 1994, vol. 51 (cit. on p. 20).

[52] K.-T. Fang and Y. Zhang, *Generalized Multivariate Analysis*. New York: Springer-Verlag, 1990 (cit. on pp. 20, 21).

[53] J. C. Farman, B. G. Gardiner, and J. D. Shanklin, “Large losses of total ozone in Antarctica reveal seasonal $\mathrm{ClO}_x$ / $\mathrm{NO}_x$ interaction”, *Nature*, vol. 315, no. 6016, pp. 207–210, 1985 (cit. on p. 4).

[54] C. I. Fassett, “Analysis of impact crater populations and the geochronology of planetary surfaces in the inner solar system”, *J. Geophys. Res.*, vol. 121, no. 10, pp. 1900–1926, 2016 (cit. on p. 140).

[55] M. P. Fay, “Two-sided exact tests and matching confidence intervals for discrete data”, *R J.*, vol. 2, no. 1, pp. 53–58, 2010 (cit. on p. 137).

[56] M. Febrero, P. Galeano, and W. González-Manteiga, “A functional analysis of $\mathrm{NO}_x$ levels: Location and scale estimation and outlier detection”, *Comput. Statist.*, vol. 22, no. 3, pp. 411–427, 2007 (cit. on p. 13).

[57] ——, “Outlier detection in functional data by depth measures, with application to identify abnormal nox levels”, *Environmetrics*, vol. 19, no. 4, pp. 331–345, 2008 (cit. on p. 13).

[58] C. J. Feltz and G. A. Goldin, “Partition-based goodness-of-fit tests on the line and the circle”, *Aust. N. Z. J. Stat.*, vol. 43, no. 2, pp. 207–220, 2001 (cit. on pp. xxvi, 144).

[59] A. Figueiredo, “Comparison of tests of uniformity defined on the hypersphere”, *Statist. Probab. Lett.*, vol. 77, no. 3, pp. 329–334, 2007 (cit. on p. 137).

[60] A. Figueiredo and P. Gomes, “Power of tests of uniformity defined on the hypersphere”, *Comm. Statist. Simulation Comput.*, vol. 32, no. 1, pp. 87–94, 2003 (cit. on p. 137).

[61] P. Filzmoser, R. Maronna, and M. Werner, “Outlier identification in high dimensions”, *Comput. Statist. Data Anal.*, vol. 52, no. 3, pp. 1694–1711, 2008 (cit. on pp. xxii, 19, 86, 90).

[62] J. H. Friedman and J. W. Tukey, “A projection pursuit algorithm for exploratory data analysis”, *IEEE Trans. Comput.*, vol. C-23, no. 9, pp. 881–890, 1974 (cit. on p. 18).

[63] V. Fritsch, G. Varoquaux, B. Thyreau, J. B. Poline, and B. Thirion, “Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators”, *Med. Image Anal.*, vol. 16, no. 7, pp. 1359 –1370, 2012 (cit. on p. 17).

[64] E. García-Portugués and T. Verdebout, “An overview of uniformity tests on the hypersphere”, *arXiv:1804.00286*, 2019 (cit. on pp. xxiv, 7, 25, 134).

[65] ——, *sphunif: Uniformity tests on the circle, sphere, and hypersphere*, https://github.com/egarpor/sphunif, 2020 (cit. on pp. xxv, 8, 138).

[66] E. García-Portugués, W. González-Manteiga, and M. Febrero-Bande, “A goodness-of-fit test for the functional linear model with scalar response”, *J. Comput. Graph. Stat.*, vol. 23, no. 3, pp. 761–778, 2014 (cit. on pp. xix, 3).

[67] E. García-Portugués, P. Navarro-Esteban, and J. A. Cuesta-Albertos, “On a projection-based class of uniformity tests on the hypersphere”, *arXiv:2008.09897*, 2020 (cit. on pp. xx, 3, 8).

[68] E. García-Portugués, D. Paindaveine, and T. Verdebout, “On optimal tests for rotational symmetry against new classes of hyperspherical distributions”, *J. Amer. Statist. Assoc.*, vol. to appear, 2020 (cit. on pp. xxiv, 7, 138).

[69] ——, *Rotasym: Tests for rotational symmetry on the hypersphere*, R package version 1.0.7, 2020 (cit. on p. 135).

[70] E. García-Portugués, P. Navarro-Esteban, and J. A. Cuesta-Albertos, “A Cramér–von Mises test of uniformity on the hypersphere”, in *Under review in Cladag 2019 Post Proceedings*, S. Balzano, G. C. Porzio, R. Salvatore, D. Vistocco, and M. Vichi, Eds., ser. Studies in Classification, Data Analysis and Knowledge Organization, Cham: Springer, 2021 (cit. on pp. xx, 3, 141).

[71] E. Giné, “Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms”, *Ann. Statist.*, vol. 3, no. 6, pp. 1243–1266, 1975 (cit. on pp. 22–26, 134).

[72] N. Goel, G. Bebis, and A. Nefian, “Face recognition experiments with random projection”, *Proc. of SPIE*, vol. 5779, pp. 426–437, 2005 (cit. on pp. xvii, 1).

[73] M. Gölz, M. Fauss, and A. Zoubir, “A bootstrapped sequential probability ratio test for signal processing applications”, in *IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec. 2017, pp. 1–5 (cit. on p. 30).

[74] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, Eighth Edition. Boston: Elsevier Science, 2014 (cit. on p. 129).

[75] L. Győri, T. Baranyi, and A. Ludámny, “Comparative analysis of Debrecen sunspot catalogues”, *Mon. Not. R. Astron. Soc.*, vol. 465, no. 2, pp. 1259–1273, Dec. 2016 (cit. on p. 138).

[76]D. Hawkins, *Identification of Outliers*. London: Springer, 1980 (cit. on p. 13).

[77]M. J. R. Healy, “Multivariate normal plotting”, *J. R. Stat. Soc. Ser. C. Appl. Stat.*, vol. 17, no. 2, pp. 157–161, 1968 (cit. on p. 14).

[78]N. Hirata, “Differential impact cratering of Saturn’s satellites by heliocentric impactors”, *J. Geophys. Res. Planets*, vol. 121, no. 2, pp. 111–117, 2016 (cit. on pp. 141, 142).

[79]D. C. Hoaglin, B. Iglewicz, and J. W. Tukey, “Performance of some resistant rules for outlier labeling”, *J. Amer. Statist. Assoc.*, vol. 81, no. 396, pp. 991–999, 1986 (cit. on p. 14).

[80]M. Hubert, P. J. Rousseeuw, and P. Segaert, “Multivariate functional outlier detection”, *Stat. Methods Appl.*, vol. 24, no. 2, pp. 177–202, 2015 (cit. on pp. xxiii, 90, 91, 93).

[81]B. Iglewicz and S. A Banerjee, “A simple univariate outlier identification procedure”, *Proc. Annu. Meeting Amer. Statist. Assoc.*, 2001 (cit. on pp. 15, 91, 92).

[82]J. P. Imhof, “Computing the distribution of quadratic forms in normal variables”, *Biometrika*, vol. 48, no. 3/4, pp. 419–426, 1961 (cit. on pp. 131, 132).

[83]S. R. Jammalamadaka, S. Meintanis, and T. Verdebout, “On Sobolev tests of uniformity on the circle with an extension to the sphere”, *Bernoulli*, vol. 20, no. 3, pp. 2226–2252, 2020 (cit. on p. 24).

[84]J. M. Jobe and M. Pokojovy, “A cluster-based outlier detection scheme for multivariate data”, *J. Amer. Statist. Assoc.*, vol. 110, no. 512, pp. 1543–1551, 2015 (cit. on p. 17).

[85]W. Johnson and J. Lindenstrauss, “Extensions of Lipschitz maps into a Hilbert space”, *Contemporary Mathematics*, vol. 26, pp. 189–206, Jan. 1984 (cit. on p. 11).

[86]I. M. Johnstone and A. Y. Lu, “On consistency and sparsity for principal components analysis in high dimensions”, *J. Amer. Statist. Assoc.*, vol. 104, no. 486, pp. 682–693, 2009 (cit. on p. 19).

[87]I. Jolliffe, *Principal Component Analysis*. New York: John Wiley and Sons, Ltd, 2002 (cit. on p. 18).

[88]P. E. Jupp, “Data-driven Sobolev tests of uniformity on compact Riemannian manifolds”, *Ann. Statist.*, vol. 36, no. 3, pp. 1246–1260, 2008 (cit. on p. 24).

[89]P. E. Jupp and A. Kume, “Measures of goodness of fit obtained by almost-canonical transformations on Riemannian manifolds”, *J. Multivar. Anal.*, vol. 176, p. 104 579, 2020 (cit. on pp. xxiv, 7).

[90]P. E. Jupp, P. T. Kim, J.-Y. Koo, and P. Wiegert, “The intrinsic distribution and selection bias of long-period cometary orbits”, *J. Amer. Statist. Assoc.*, vol. 98, no. 463, pp. 515–521, 2003 (cit. on pp. 139, 140).

[91]J. B. Kruskal, “Toward a practical method which helps uncover the structure of a set of observations by finding the line transformation which optimizes a new “index of condensation””, *Stat. Comput.*, R. C. Milton and J. A. Nelder, Eds., pp. 427–440, 1969 (cit. on p. 18).

[92]N. H. Kuiper, “Tests concerning random points on a circle”, *Proc. K. Ned. Akad. Wet. A*, vol. 63, pp. 38–47, 1960 (cit. on p. 164).

[93]T. L. Lai, “Asymptotic optimality of invariant sequential probability ratio tests”, *Ann. Statist.*, vol. 9, no. 2, pp. 318–333, 1981 (cit. on p. 30).

[94]F. H. Larsen, F. van den Berg, and S. B. Engelsen, “An exploratory chemometric study of 1h nmr spectra of table wines”, *J. Chemom.*, vol. 20, no. 5, pp. 198–208, (cit. on p. 90).

[95]B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection”, *Ann. Math. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000 (cit. on p. 46).

[96]M. L. Laursen and K. Mita, “Some integrals involving associated Legendre functions and Gegenbauer polynomials”, *J. Phys. A: Math. Gen.*, vol. 14, no. 5, p. 1065, 1981 (cit. on p. 127).

[97]Y. Lee and W. C. Kim, “Concise formulas for the surface area of the intersection of two hyperspherical caps”, Korea Advanced Institute of Science and Technology, Tech. Rep., 2014 (cit. on p. 98).

[98]C. Ley and T. Verdebout, *Modern Directional Statistics*, ser. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton: CRC Press, 2017 (cit. on pp. xxiii, 6).

[99]——, *Applied Directional Statistics: Modern Methods and Case Studies*, ser. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton: CRC Press, 2018 (cit. on pp. xxiii, 6).

[100]R. Y. Liu, J. M. Parelius, and K. Singh, “Multivariate analysis by data depth: Descriptive statistics, graphics and inference”, *Ann. Statist.*, vol. 27, no. 3, pp. 783–858, 1999 (cit. on p. 17).

[101]G. Lohöfer, “Inequalities for Legendre functions and Gegenbauer functions”, *J. Approx. Theory*, vol. 64, no. 2, pp. 226–234, 1991 (cit. on p. 115).

[102]K. V. Mardia and P. E. Jupp, *Directional Statistics*, ser. Wiley Series in Probability and Statistics. London: John Wiley & Sons, 2000 (cit. on pp. xxiii, xxiv, 6, 7, 21, 26, 106, 135).

[103]R. A. Maronna and V. J. Yohai, “The behavior of the Stahel-Donoho robust multivariate estimator”, *J. Amer. Statist. Assoc.*, vol. 90, no. 429, pp. 330–341, 1995 (cit. on p. 19).

[104]R. A. Maronna and R. H. Zamar, “Robust estimates of location and dispersion for high-dimensional datasets”, *Technometrics*, vol. 44, no. 4, pp. 307–317, 2002 (cit. on p. 19).

[105]R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. Hoboken: John Wiley & Sons, 2019 (cit. on pp. 16, 31).

[106]T. K. Matthes, “On the optimality of sequential probability ratio tests”, *Ann. Math. Statist.*, vol. 34, no. 1, pp. 18–21, 1963 (cit. on p. 29).

[107]A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton, NJ, USA: Princeton University Press, 2015 (cit. on p. 20).

[108]D. C. Montgomery, *Introduction to Statistical Quality Control*. New York: Wiley, 2008 (cit. on p. 28).

[109]P. Navarro-Esteban and J. A. Cuesta-Albertos, “High-dimensional outlier detection using random projections”, *arXiv:2005.08923*, 2020, Preprint (cit. on pp. xx, 3, 8).

[110]R. Niu and P. K. Varshney, “Sampling schemes for sequential detection with dependent observations”, *IEEE Trans. on Signal Process.*, vol. 58, no. 3, pp. 1469–1481, 2009 (cit. on p. 30).

[111]J. Pan, W. Fung, and K. Fang, “Multiple outlier detection in multivariate data using projection pursuit techniques”, *J. Statist. Plann. Inference*, vol. 83, no. 1, pp. 153–167, 2000 (cit. on p. 20).

[112]D. Peña and F. J. Prieto, “Multivariate outlier detection and robust covariance matrix estimation”, *Technometrics*, vol. 43, no. 3, pp. 286–310, 2001 (cit. on pp. 16, 19).

[113]A. Pewsey and E. García-Portugués, “Recent advances in directional statistics”, *arXiv: 2005.06889*, 2020 (cit. on pp. xxiii, 6).

[114]M. J. Prentice, “On invariant tests of uniformity for directions and orientations”, *Ann. Statist.*, vol. 6, no. 1, pp. 169–176, 1978 (cit. on pp. 23, 24, 26, 105, 109, 113, 131, 133, 134).

[115]L. Rayleigh, “On the problem of random vibrations, and of random flights in one, two, or three dimensions”, *Philos. Mag*, vol. 37, no. 220, pp. 321–347, 1919 (cit. on pp. 25, 125, 133).

[116]K. Ro, C. Zou, Z. Wang, and G. Yin, “Outlier detection for high-dimensional data”, *Biometrika*, vol. 102, no. 3, pp. 589–599, 2015 (cit. on pp. xxii, 17, 86, 90).

[117]E. D. Rothman, “Tests for uniformity of a circular distribution”, *Sankhyā*, vol. 34, no. 1, pp. 23–32, 1972 (cit. on pp. xxv, 7, 26, 105, 106, 113).

[118]P. J. Rousseeuw, “Multivariate estimation with high breakdown point”, *Math. Stat. Appl.*, vol. 8, no. 283–297, p. 37, 1985 (cit. on p. 16).

[119]P. J. Rousseeuw and K. Van Driessen, “A fast algorithm for the minimum covariance determinant estimator”, *Technometrics*, vol. 41, no. 3, pp. 212–23, 1998 (cit. on p. 17).

[120]P. J. Rousseeuw and B. C. Van Zomeren, “Unmasking multivariate outliers and leverage points”, *J. Amer. Statist. Assoc.*, vol. 85, no. 411, pp. 633–639, 1990 (cit. on p. 16).

[121]P. J. Rousseeuw, M. Debruyne, S. Engelen, and M. Hubert, “Robustness and outlier detection in chemometrics”, *Crit. Rev. Anal. Chem.*, vol. 36, no. 3–4, pp. 221–242, 2006 (cit. on p. 93).

[122]R. Serfling and S. Mazumder, “Computationally easy outlier detection via projection pursuit with finitely many directions”, *J. Nonparametr. Stat.*, vol. 25, no. 2, pp. 447–461, 2013 (cit. on p. 20).

[123]F. Smithies, *Integral Equations*, ser. Cambridge Tracts in Mathematics and Mathematical Physics. London: Cambridge University Press, 1958, vol. 49 (cit. on pp. 27, 28).

[124]S. W. Squyres, C. Howell, M. C. Liu, and J. J. Lissauer, “Investigation of crater “saturation” using spatial statistics”, *Icarus*, vol. 125, no. 1, pp. 67–82, 1997 (cit. on p. 141).

[125]W. A. Stahel, “Breakdown of covariance estimators”, *Fachgruppe fur Statistik*, 1981 (cit. on p. 19).

[126]M. A. Stephens, “A goodness-of-fit statistic for the circle, with some comparisons”, *Biometrika*, vol. 56, no. 1, pp. 161–168, 1969 (cit. on pp. 137, 164).

[127]——, “EDF statistics for goodness of fit and some comparisons”, *J. Amer. Statist. Assoc.*, vol. 69, no. 347, pp. 730–737, 1974 (cit. on p. 96).

[128]Y. Sun and M. G. Genton, “Functional boxplots”, *J. Comput. Graphic. Statist*, vol. 20, no. 2, pp. 316–334, 2011 (cit. on p. 17).

[129]——, “Adjusted functional boxplots for spatio-temporal data visualization and outlier detection”, *Environmetrics*, vol. 23, no. 1, pp. 54–64, 2012 (cit. on p. 17).

[130]A. Tartakovski, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. New York: CRC Press, 2014 (cit. on pp. xix, 2, 29).

[131]J. W. Tukey, *Exploratory Data Analysis*. Don Mills: Addison-Wesley, 1977 (cit. on pp. 14, 15).

[132]S. S. Vempala, *The Random Projection Method*. Providence: DIMACS - Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Soc., 2004, vol. 65 (cit. on pp. xvii, 1, 12).

[133]A. Wald, *Sequential Analysis*. New York: Wiley, 1947 (cit. on p. 28).

[134]A. Wald and J. Wolfowitz, “Optimum character of the sequential probability ratio test”, *Ann. Math. Statist.*, vol. 19, pp. 326–339, 1948 (cit. on p. 29).

[135]G. S. Watson, “Goodness-of-fit tests on a circle”, *Biometrika*, vol. 48, no. 1/2, pp. 109–114, 1961 (cit. on pp. xxv, 7, 25, 103, 104, 113).

[136]——, “Another test for the uniformity of a circular distribution”, *Biometrika*, vol. 54, no. 3/4, pp. 675–677, 1967 (cit. on p. 105).

[137]——, “Orientation statistics in the earth sciences”, *Bull. Geol. Inst. Univ. Upps.*, vol. 2, no. 1, pp. 73–89, 1970 (cit. on p. 139).

[138]S. Weisberg, *Applied Linear Regression*, Fourth Edition. Hoboken NJ: Wiley, 2014 (cit. on pp. xx, 4).

[139]G. B. Wetherill and D. W. Brown, *Statistical Process Control: Theory and Practice*. New York: Chapman and Hall, 1991 (cit. on p. 28).

[140]P. Wiegert and S. Tremaine, “The evolution of long-period comets”, *Icarus*, vol. 137, no. 1, pp. 84–121, 1999 (cit. on p. 139).

[141]K. Zahnle, S. P., H. Levison, and L. Dones, “Cratering rates in the outer Solar System”, *Icarus*, vol. 163, no. 2, pp. 263–289, 2003 (cit. on p. 140).

# Appendix A

## A.1 Additional tables of Chapter 3

Table A.1.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_{\Sigma} = C_n^d$, for different values of $n, d$ and $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$ when we use the values of $a$ and $b$ computed with Proposition 3.5.4. We also show the sample mean of $K$. This table is Table 3.6 expanded to $n = 100, 500$ and $d = 5$.

| $d$ | $n$ | $k_I^1$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 50 | 50 | 58 | 0.0534 | 51 | 0.0516 | 73 | 0.0638 | 51 | 0.0482 |
| | | 100 | 119 | 0.0490 | 104 | 0.0524 | 147 | 0.0734 | 101 | 0.0508 |
| | 100 | 50 | 58 | 0.0544 | 52 | 0.0570 | 76 | 0.0698 | 50 | 0.0490 |
| | | 100 | 120 | 0.0536 | 104 | 0.0622 | 147 | 0.0702 | 100 | 0.0528 |
| | 500 | 50 | 58 | 0.0534 | 54 | 0.0578 | 75 | 0.0644 | 50 | 0.0588 |
| | | 100 | 122 | 0.0486 | 106 | 0.0502 | 152 | 0.0734 | 100 | 0.0542 |
| 50 | 50 | 50 | 49 | 0.0544 | 51 | 0.0528 | 190 | 0.0492 | 50 | 0.0448 |
| | | 100 | 101 | 0.0504 | 103 | 0.0534 | 379 | 0.0510 | 100 | 0.0508 |
| | 100 | 50 | 51 | 0.0484 | 49 | 0.0484 | 189 | 0.0466 | 50 | 0.0524 |
| | | 100 | 100 | 0.0502 | 97 | 0.0546 | 374 | 0.0478 | 100 | 0.0492 |
| | 500 | 50 | 50 | 0.0518 | 50 | 0.0572 | 190 | 0.0504 | 52 | 0.0518 |
| | | 100 | 100 | 0.0454 | 100 | 0.0522 | 392 | 0.0504 | 96 | 0.0510 |
| 100 | 50 | 50 | 51 | 0.0500 | 50 | 0.0460 | 262 | 0.0474 | 50 | 0.0470 |
| | | 100 | 101 | 0.0498 | 99 | 0.0478 | 516 | 0.0466 | 101 | 0.0542 |
| | 100 | 50 | 50 | 0.0552 | 49 | 0.0500 | 260 | 0.0474 | 50 | 0.0514 |
| | | 100 | 99 | 0.0482 | 100 | 0.0510 | 546 | 0.0458 | 99 | 0.0456 |
| | 500 | 50 | 50 | 0.0562 | 49 | 0.0496 | 252 | 0.0482 | 50 | 0.0556 |
| | | 100 | 101 | 0.0572 | 100 | 0.0508 | 525 | 0.0472 | 99 | 0.0500 |
| 500 | 50 | 50 | 51 | 0.0476 | 49 | 0.0498 | 548 | 0.0416 | 51 | 0.0498 |
| | | 100 | 99 | 0.0556 | 98 | 0.0514 | 1184 | 0.0484 | 102 | 0.0552 |
| | 100 | 50 | 50 | 0.0514 | 49 | 0.0528 | 589 | 0.0474 | 50 | 0.0492 |
| | | 100 | 99 | 0.0502 | 100 | 0.0578 | 1195 | 0.0470 | 101 | 0.0512 |
| | 500 | 50 | 50 | 0.0488 | 50 | 0.0564 | 578 | 0.0496 | 51 | 0.0490 |
| | | 100 | 100 | 0.0490 | 99 | 0.0536 | 1185 | 0.0480 | 99 | 0.0498 |
| 1000 | 50 | 50 | 50 | 0.0494 | 50 | 0.0520 | 846 | 0.0500 | 50 | 0.0510 |
| | | 100 | 101 | 0.0530 | 101 | 0.0515 | 1610 | 0.0415 | 96 | 0.0525 |
| | 100 | 50 | 50 | 0.0430 | 49 | 0.0510 | 824 | 0.0502 | 51 | 0.0510 |
| | | 100 | 103 | 0.0400 | 101 | 0.0515 | 1744 | 0.0430 | 111 | 0.0470 |
| | 500 | 50 | 49 | 0.0496 | 51 | 0.0468 | 823 | 0.0456 | 50 | 0.0500 |
| | | 100 | 98 | 0.0520 | 101 | 0.0450 | 1696 | 0.0465 | 96 | 0.0540 |

**Table A.2.:** Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$ with $r = 1.2, 2$. We also show the sample mean of $K$. This table expands Table 3.7 to $n = 100, 500$ and $d = 5$ and its description applies.

| $d$ | $n$ | $\|\mathbf{X}\|_\Sigma$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 50 | $1.2C_n^d$ | 50 | 15 | 0.7698 | 28 | 0.5356 | 21 | 0.6414 | 25 | 0.6056 | 15 | 0.7458 |
| | | | 100 | 18 | 0.8472 | 41 | 0.6162 | 30 | 0.7206 | 34 | 0.7010 | 19 | 0.8488 |
| | | $2C_n^d$ | 50 | 3 | 0.9727 | 5 | 0.9336 | 4 | 0.9566 | 4 | 0.9410 | 3 | 0.9714 |
| | | | 100 | 3 | 0.9887 | 6 | 0.9648 | 4 | 0.9742 | 5 | 0.9706 | 3 | 0.9860 |
| | 100 | $1.2C_n^d$ | 50 | 15 | 0.7654 | 28 | 0.5184 | 22 | 0.6330 | 24 | 0.6072 | 15 | 0.7624 |
| | | | 100 | 18 | 0.8588 | 43 | 0.6106 | 30 | 0.7256 | 33 | 0.7062 | 19 | 0.8452 |
| | | $2C_n^d$ | 50 | 3 | 0.9657 | 5 | 0.9334 | 4 | 0.9590 | 4 | 0.9454 | 3 | 0.9722 |
| | | | 100 | 3 | 0.9887 | 5 | 0.9584 | 4 | 0.9738 | 5 | 0.9710 | 3 | 0.9860 |
| | 500 | $1.2C_n^d$ | 50 | 15 | 0.7684 | 28 | 0.5202 | 20 | 0.6452 | 24 | 0.6016 | 16 | 0.7490 |
| | | | 100 | 18 | 0.8500 | 43 | 0.6164 | 31 | 0.7320 | 34 | 0.7024 | 19 | 0.8458 |
| | | $2C_n^d$ | 50 | 3 | 0.9673 | 5 | 0.9324 | 4 | 0.9602 | 4 | 0.9444 | 3 | 0.9688 |
| | | | 100 | 3 | 0.9900 | 5 | 0.9572 | 4 | 0.9758 | 5 | 0.9704 | 3 | 0.9846 |
| 50 | 50 | $1.2C_n^d$ | 50 | 43 | 0.3124 | 44 | 0.2816 | 43 | 0.3020 | 148 | 0.2070 | 44 | 0.3188 |
| | | | 100 | 82 | 0.3638 | 84 | 0.3112 | 82 | 0.3460 | 287 | 0.2122 | 79 | 0.3590 |
| | | $2C_n^d$ | 50 | 8 | 0.9220 | 9 | 0.9146 | 8 | 0.9230 | 34 | 0.7224 | 8 | 0.9226 |
| | | | 100 | 10 | 0.9463 | 11 | 0.9470 | 10 | 0.9526 | 49 | 0.7786 | 10 | 0.9576 |
| | 100 | $1.2C_n^d$ | 50 | 44 | 0.3196 | 45 | 0.2862 | 44 | 0.2974 | 149 | 0.2016 | 44 | 0.3076 |
| | | | 100 | 81 | 0.3672 | 84 | 0.3102 | 84 | 0.3512 | 284 | 0.2192 | 83 | 0.3648 |
| | | $2C_n^d$ | 50 | 8 | 0.9247 | 9 | 0.9120 | 8 | 0.9192 | 34 | 0.7112 | 8 | 0.9286 |
| | | | 100 | 10 | 0.9560 | 11 | 0.9414 | 10 | 0.9562 | 48 | 0.7820 | 10 | 0.9570 |
| | 500 | $1.2C_n^d$ | 50 | 44 | 0.3008 | 44 | 0.2752 | 42 | 0.3090 | 149 | 0.1950 | 43 | 0.2930 |
| | | | 100 | 81 | 0.3556 | 84 | 0.3106 | 80 | 0.3484 | 289 | 0.2170 | 79 | 0.3714 |
| | | $2C_n^d$ | 50 | 8 | 0.9240 | 9 | 0.9114 | 8 | 0.9220 | 33 | 0.7296 | 8 | 0.9190 |
| | | | 100 | 10 | 0.9580 | 12 | 0.9444 | 10 | 0.9508 | 50 | 0.7752 | 10 | 0.9536 |
| 100 | 50 | $1.2C_n^d$ | 50 | 44 | 0.3018 | 46 | 0.2854 | 45 | 0.2890 | 218 | 0.1910 | 45 | 0.2956 |
| | | | 100 | 84 | 0.3354 | 87 | 0.3138 | 83 | 0.3454 | 418 | 0.2034 | 83 | 0.3492 |
| | | $2C_n^d$ | 50 | 9 | 0.9163 | 9 | 0.9152 | 9 | 0.9136 | 49 | 0.6960 | 9 | 0.9156 |
| | | | 100 | 11 | 0.9513 | 12 | 0.9440 | 11 | 0.9452 | 73 | 0.7376 | 11 | 0.9420 |
| | 100 | $1.2C_n^d$ | 50 | 44 | 0.2874 | 45 | 0.2750 | 45 | 0.2874 | 219 | 0.1836 | 44 | 0.2834 |
| | | | 100 | 83 | 0.3448 | 86 | 0.3100 | 84 | 0.3310 | 406 | 0.1948 | 82 | 0.3392 |
| | | $2C_n^d$ | 50 | 8 | 0.9213 | 9 | 0.9114 | 9 | 0.9106 | 47 | 0.6922 | 8 | 0.9116 |
| | | | 100 | 11 | 0.9427 | 12 | 0.9454 | 11 | 0.9494 | 72 | 0.7546 | 11 | 0.9526 |
| | 500 | $1.2C_n^d$ | 50 | 44 | 0.2922 | 46 | 0.2838 | 44 | 0.2932 | 211 | 0.1960 | 45 | 0.2956 |
| | | | 100 | 83 | 0.3312 | 87 | 0.3062 | 83 | 0.3368 | 410 | 0.1996 | 83 | 0.3398 |
| | | $2C_n^d$ | 50 | 8 | 0.9280 | 9 | 0.9156 | 9 | 0.9200 | 47 | 0.7036 | 8 | 0.9210 |
| | | | 100 | 11 | 0.9513 | 12 | 0.9552 | 11 | 0.9512 | 74 | 0.7406 | 10 | 0.9466 |
| 500 | 50 | $1.2C_n^d$ | 50 | 46 | 0.2824 | 45 | 0.2818 | 45 | 0.2790 | 488 | 0.1958 | 45 | 0.2808 |
| | | | 100 | 86 | 0.3324 | 88 | 0.3106 | 87 | 0.3174 | 932 | 0.1912 | 88 | 0.3294 |
| | | $2C_n^d$ | 50 | 9 | 0.9007 | 9 | 0.9124 | 9 | 0.9106 | 106 | 0.6764 | 9 | 0.9142 |
| | | | 100 | 11 | 0.9493 | 11 | 0.9426 | 11 | 0.9520 | 164 | 0.7196 | 11 | 0.9488 |
| | 100 | $1.2C_n^d$ | 50 | 45 | 0.2876 | 47 | 0.2790 | 46 | 0.2802 | 487 | 0.1828 | 44 | 0.2852 |
| | | | 100 | 83 | 0.3270 | 86 | 0.3074 | 84 | 0.3202 | 916 | 0.1826 | 86 | 0.3266 |
| | | $2C_n^d$ | 50 | 9 | 0.9173 | 9 | 0.9118 | 9 | 0.9174 | 106 | 0.6782 | 9 | 0.9128 |
| | | | 100 | 11 | 0.9437 | 12 | 0.9450 | 11 | 0.9446 | 166 | 0.7192 | 11 | 0.9428 |
| | 500 | $1.2C_n^d$ | 50 | 46 | 0.2900 | 44 | 0.2824 | 46 | 0.2728 | 480 | 0.1748 | 45 | 0.2758 |
| | | | 100 | 83 | 0.3366 | 86 | 0.3006 | 85 | 0.3254 | 918 | 0.1856 | 87 | 0.3324 |
| | | $2C_n^d$ | 50 | 9 | 0.9207 | 9 | 0.9168 | 9 | 0.9228 | 105 | 0.6756 | 9 | 0.9200 |
| | | | 100 | 11 | 0.9503 | 11 | 0.9462 | 11 | 0.9476 | 166 | 0.7108 | 11 | 0.9482 |
| 1000 | 50 | $1.2C_n^d$ | 50 | 45 | 0.2753 | 45 | 0.2840 | 45 | 0.2815 | 660 | 0.1810 | 46 | 0.2755 |
| | | | 100 | 86 | 0.3140 | 84 | 0.3095 | 86 | 0.3255 | 1367 | 0.1875 | 83 | 0.3215 |
| | | $2C_n^d$ | 50 | 9 | 0.9127 | 9 | 0.9147 | 9 | 0.9193 | 152 | 0.6867 | 9 | 0.9123 |
| | | | 100 | 11 | 0.9430 | 11 | 0.9447 | 11 | 0.9443 | 237 | 0.7043 | 12 | 0.9463 |
| | 100 | $1.2C_n^d$ | 50 | 46 | 0.2867 | 46 | 0.2682 | 46 | 0.2828 | 687 | 0.1778 | 46 | 0.2688 |
| | | | 100 | 85 | 0.3320 | 86 | 0.3250 | 88 | 0.3100 | 1300 | 0.1875 | 85 | 0.3275 |
| | | $2C_n^d$ | 50 | 9 | 0.9053 | 9 | 0.9137 | 9 | 0.9123 | 155 | 0.6697 | 9 | 0.9093 |
| | | | 100 | 12 | 0.9460 | 11 | 0.9427 | 11 | 0.9420 | 227 | 0.7250 | 11 | 0.9523 |
| | 500 | $1.2C_n^d$ | 50 | 45 | 0.2833 | 47 | 0.2828 | 47 | 0.2730 | 672 | 0.1882 | 45 | 0.2920 |
| | | | 100 | 88 | 0.3287 | 87 | 0.3210 | 85 | 0.3145 | 1303 | 0.2025 | 82 | 0.3270 |
| | | $2C_n^d$ | 50 | 9 | 0.9083 | 9 | 0.9093 | 9 | 0.9123 | 152 | 0.6683 | 9 | 0.9057 |
| | | | 100 | 12 | 0.9383 | 12 | 0.9397 | 11 | 0.9437 | 228 | 0.7227 | 11 | 0.9483 |

**Table A.3.:** Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$, for different values of $n, d$ and $\Sigma$ using $a$ and $b$ from Proposition 3.5.4 and $k_{\max} = 1/(1 - F(b,t) + F(a,t))$. We also show the sample mean of the required projections. This table extends Table 3.8 to $n = 100, 500$ and $d = 5$.

| $d$ | $n$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 50 | 50 | 32 | 0.0334 | 33 | 0.0350 | 32 | 0.0320 | 34 | 0.0282 | 32 | 0.0326 |
| | | 100 | 64 | 0.0340 | 66 | 0.0298 | 65 | 0.0304 | 67 | 0.0292 | 64 | 0.0314 |
| | 100 | 50 | 32 | 0.0314 | 33 | 0.0366 | 32 | 0.0344 | 34 | 0.0372 | 32 | 0.0366 |
| | | 100 | 64 | 0.0356 | 67 | 0.0288 | 63 | 0.0328 | 67 | 0.0256 | 63 | 0.0324 |
| | 500 | 50 | 32 | 0.0344 | 33 | 0.0332 | 33 | 0.0328 | 34 | 0.0320 | 32 | 0.0338 |
| | | 100 | 64 | 0.0296 | 65 | 0.0312 | 64 | 0.0322 | 66 | 0.0258 | 64 | 0.0320 |
| 50 | 50 | 50 | 32 | 0.0300 | 33 | 0.0310 | 32 | 0.0306 | 38 | 0.0130 | 33 | 0.0326 |
| | | 100 | 65 | 0.0306 | 64 | 0.0364 | 63 | 0.0344 | 75 | 0.0106 | 64 | 0.0358 |
| | 100 | 50 | 32 | 0.0298 | 33 | 0.0322 | 32 | 0.0322 | 38 | 0.0110 | 32 | 0.0352 |
| | | 100 | 64 | 0.0334 | 64 | 0.0346 | 63 | 0.0342 | 75 | 0.0096 | 63 | 0.0298 |
| | 500 | 50 | 32 | 0.0350 | 32 | 0.0326 | 32 | 0.0328 | 38 | 0.0114 | 32 | 0.0328 |
| | | 100 | 64 | 0.0328 | 64 | 0.0312 | 64 | 0.0298 | 76 | 0.0090 | 64 | 0.0304 |
| 100 | 50 | 50 | 32 | 0.0310 | 32 | 0.0326 | 32 | 0.0318 | 40 | 0.0092 | 32 | 0.0268 |
| | | 100 | 64 | 0.0316 | 65 | 0.0280 | 64 | 0.0338 | 78 | 0.0066 | 64 | 0.0326 |
| | 100 | 50 | 32 | 0.0334 | 32 | 0.0300 | 32 | 0.0336 | 39 | 0.0096 | 32 | 0.0340 |
| | | 100 | 64 | 0.0302 | 64 | 0.0326 | 64 | 0.0296 | 78 | 0.0068 | 64 | 0.0328 |
| | 500 | 50 | 32 | 0.0302 | 32 | 0.0316 | 32 | 0.0292 | 40 | 0.0088 | 32 | 0.0364 |
| | | 100 | 63 | 0.0322 | 64 | 0.0330 | 63 | 0.0316 | 79 | 0.0062 | 64 | 0.0248 |
| 500 | 50 | 50 | 32 | 0.0316 | 32 | 0.0342 | 32 | 0.0304 | 43 | 0.0044 | 32 | 0.0350 |
| | | 100 | 64 | 0.0362 | 64 | 0.0324 | 64 | 0.0326 | 85 | 0.0054 | 64 | 0.0282 |
| | 100 | 50 | 32 | 0.0288 | 32 | 0.0342 | 32 | 0.0300 | 42 | 0.0052 | 32 | 0.0306 |
| | | 100 | 64 | 0.0292 | 64 | 0.0320 | 64 | 0.0324 | 83 | 0.0042 | 64 | 0.0344 |
| | 500 | 50 | 32 | 0.0330 | 32 | 0.0376 | 32 | 0.0342 | 42 | 0.0052 | 32 | 0.0306 |
| | | 100 | 64 | 0.0324 | 64 | 0.0362 | 64 | 0.0334 | 84 | 0.0060 | 64 | 0.0316 |
| 1000 | 50 | 50 | 32 | 0.0337 | 33 | 0.0340 | 33 | 0.0373 | 43 | 0.0053 | 32 | 0.0320 |
| | | 100 | 63 | 0.0322 | 63 | 0.0377 | 64 | 0.0337 | 86 | 0.0027 | 64 | 0.0247 |
| | 100 | 50 | 32 | 0.0323 | 32 | 0.0267 | 32 | 0.0360 | 44 | 0.0040 | 32 | 0.0340 |
| | | 100 | 65 | 0.0321 | 64 | 0.0297 | 63 | 0.0373 | 86 | 0.0030 | 64 | 0.0297 |
| | 500 | 50 | 33 | 0.0337 | 32 | 0.0340 | 32 | 0.0450 | 43 | 0.0047 | 32 | 0.0303 |
| | | 100 | 64 | 0.0314 | 64 | 0.0297 | 63 | 0.0330 | 87 | 0.0043 | 62 | 0.0370 |

**Table A.4.:** Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_{\Sigma} = rC_n^d$, $r = 1.2, 2$ when we truncate by $k_{\max} = 1/(1 - F(b,t) + F(a,t))$. This table expands Table 3.9 to $n = 100, 500$ and $d = 5$ and its description applies.

| $d$ | $n$ | $\|\mathbf{X}\|_{\Sigma}$ | $k_I^1$ | $\hat{k}_I$ | $I_d$ | $\hat{k}_1$ | $\Sigma_1^d$ | $\hat{k}_2$ | $\Sigma_2^d$ | $\hat{k}_3$ | $\Sigma_3^d$ | $\hat{k}_4$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 50 | $1.2C_n^d$ | 50 | 14 | 0.7376 | 21 | 0.4722 | 17 | 0.6130 | 20 | 0.5326 | 15 | 0.7220 |
| | | | 100 | 15 | 0.7366 | 35 | 0.5724 | 27 | 0.7068 | 31 | 0.6620 | 19 | 0.8524 |
| | | $2C_n^d$ | 50 | 3 | 0.9726 | 5 | 0.9266 | 4 | 0.9588 | 5 | 0.9468 | 3 | 0.9728 |
| | | | 100 | 3 | 0.9858 | 5 | 0.9626 | 4 | 0.9776 | 5 | 0.9734 | 3 | 0.9844 |
| | 100 | $1.2C_n^d$ | 50 | 15 | 0.7388 | 21 | 0.4758 | 18 | 0.6012 | 21 | 0.5204 | 15 | 0.7282 |
| | | | 100 | 15 | 0.7396 | 36 | 0.5598 | 28 | 0.6966 | 30 | 0.6752 | 18 | 0.8424 |
| | | $2C_n^d$ | 50 | 3 | 0.9742 | 5 | 0.9228 | 4 | 0.9586 | 4 | 0.9422 | 3 | 0.9704 |
| | | | 100 | 3 | 0.9858 | 6 | 0.9650 | 4 | 0.9770 | 5 | 0.9698 | 3 | 0.9868 |
| | 500 | $1.2C_n^d$ | 50 | 15 | 0.7370 | 21 | 0.4726 | 18 | 0.6068 | 21 | 0.5298 | 15 | 0.7240 |
| | | | 100 | 15 | 0.7364 | 35 | 0.5702 | 26 | 0.7014 | 30 | 0.6688 | 19 | 0.8378 |
| | | $2C_n^d$ | 50 | 3 | 0.9730 | 5 | 0.9248 | 3 | 0.9556 | 4 | 0.9494 | 3 | 0.9726 |
| | | | 100 | 3 | 0.9828 | 5 | 0.9630 | 4 | 0.9786 | 5 | 0.9710 | 3 | 0.9856 |
| 50 | 50 | $1.2C_n^d$ | 50 | 30 | 0.2126 | 31 | 0.2050 | 31 | 0.2104 | 38 | 0.0628 | 30 | 0.2118 |
| | | | 100 | 57 | 0.2714 | 59 | 0.2338 | 59 | 0.2488 | 75 | 0.0690 | 59 | 0.2574 |
| | | $2C_n^d$ | 50 | 8 | 0.9186 | 9 | 0.9042 | 8 | 0.9182 | 23 | 0.5714 | 8 | 0.9198 |
| | | | 100 | 10 | 0.9566 | 11 | 0.9434 | 10 | 0.9540 | 40 | 0.6504 | 10 | 0.9574 |
| | 100 | $1.2C_n^d$ | 50 | 30 | 0.2164 | 30 | 0.1988 | 30 | 0.2062 | 38 | 0.0640 | 30 | 0.2042 |
| | | | 100 | 59 | 0.2600 | 58 | 0.2240 | 57 | 0.2546 | 75 | 0.0654 | 58 | 0.2620 |
| | | $2C_n^d$ | 50 | 8 | 0.9272 | 9 | 0.9092 | 8 | 0.9194 | 23 | 0.5670 | 8 | 0.9228 |
| | | | 100 | 10 | 0.9534 | 12 | 0.9404 | 11 | 0.9516 | 39 | 0.6638 | 10 | 0.9606 |
| | 500 | $1.2C_n^d$ | 50 | 30 | 0.2062 | 30 | 0.2068 | 30 | 0.2156 | 37 | 0.0648 | 31 | 0.2030 |
| | | | 100 | 58 | 0.2558 | 58 | 0.2286 | 58 | 0.2494 | 74 | 0.0708 | 58 | 0.2544 |
| | | $2C_n^d$ | 50 | 8 | 0.9258 | 9 | 0.9050 | 8 | 0.9132 | 23 | 0.5712 | 8 | 0.9190 |
| | | | 100 | 10 | 0.9544 | 12 | 0.9400 | 10 | 0.9550 | 39 | 0.6710 | 10 | 0.9560 |
| 100 | 50 | $1.2C_n^d$ | 50 | 30 | 0.2014 | 31 | 0.2010 | 30 | 0.1986 | 39 | 0.0504 | 31 | 0.1956 |
| | | | 100 | 58 | 0.2466 | 60 | 0.2256 | 59 | 0.2374 | 78 | 0.0466 | 59 | 0.2398 |
| | | $2C_n^d$ | 50 | 8 | 0.9184 | 9 | 0.9106 | 9 | 0.9106 | 27 | 0.4804 | 8 | 0.9172 |
| | | | 100 | 11 | 0.9468 | 12 | 0.9368 | 11 | 0.956 | 46 | 0.5752 | 11 | 0.9432 |
| | 100 | $1.2C_n^d$ | 50 | 31 | 0.1928 | 31 | 0.1962 | 31 | 0.2040 | 39 | 0.0514 | 31 | 0.2042 |
| | | | 100 | 58 | 0.2334 | 60 | 0.2238 | 59 | 0.2340 | 78 | 0.0528 | 58 | 0.2372 |
| | | $2C_n^d$ | 50 | 9 | 0.9208 | 9 | 0.9094 | 9 | 0.9158 | 27 | 0.4932 | 8 | 0.9206 |
| | | | 100 | 10 | 0.9462 | 12 | 0.9452 | 11 | 0.953 | 47 | 0.5694 | 11 | 0.9458 |
| | 500 | $1.2C_n^d$ | 50 | 31 | 0.1972 | 30 | 0.1920 | 31 | 0.1936 | 39 | 0.0568 | 31 | 0.2024 |
| | | | 100 | 59 | 0.2492 | 60 | 0.2262 | 58 | 0.2448 | 77 | 0.0566 | 58 | 0.2438 |
| | | $2C_n^d$ | 50 | 9 | 0.9214 | 9 | 0.9118 | 9 | 0.9142 | 27 | 0.4776 | 9 | 0.9180 |
| | | | 100 | 11 | 0.9456 | 12 | 0.9522 | 11 | 0.947 | 47 | 0.5790 | 11 | 0.9484 |
| 500 | 50 | $1.2C_n^d$ | 50 | 31 | 0.1898 | 31 | 0.1844 | 31 | 0.1878 | 43 | 0.0260 | 31 | 0.1904 |
| | | | 100 | 60 | 0.2198 | 60 | 0.2210 | 59 | 0.2134 | 84 | 0.0268 | 60 | 0.2244 |
| | | $2C_n^d$ | 50 | 9 | 0.9100 | 9 | 0.9116 | 9 | 0.9128 | 34 | 0.3262 | 9 | 0.9098 |
| | | | 100 | 11 | 0.9430 | 12 | 0.9442 | 11 | 0.9496 | 64 | 0.3834 | 11 | 0.9468 |
| | 100 | $1.2C_n^d$ | 50 | 30 | 0.1928 | 31 | 0.1946 | 31 | 0.1908 | 43 | 0.0292 | 31 | 0.1912 |
| | | | 100 | 60 | 0.2248 | 60 | 0.2224 | 60 | 0.2196 | 84 | 0.0264 | 60 | 0.2260 |
| | | $2C_n^d$ | 50 | 9 | 0.9150 | 9 | 0.9120 | 9 | 0.9104 | 33 | 0.3416 | 9 | 0.9184 |
| | | | 100 | 11 | 0.9458 | 11 | 0.9422 | 11 | 0.9418 | 62 | 0.4028 | 11 | 0.9464 |
| | 500 | $1.2C_n^d$ | 50 | 31 | 0.1855 | 31 | 0.1910 | 31 | 0.1892 | 43 | 0.0306 | 31 | 0.1906 |
| | | | 100 | 59 | 0.2268 | 60 | 0.2228 | 60 | 0.2142 | 85 | 0.0292 | 59 | 0.2178 |
| | | $2C_n^d$ | 50 | 9 | 0.9072 | 9 | 0.9054 | 9 | 0.9058 | 33 | 0.3358 | 9 | 0.9168 |
| | | | 100 | 11 | 0.9468 | 11 | 0.9490 | 11 | 0.9494 | 63 | 0.3922 | 12 | 0.9504 |
| 1000 | 50 | $1.2C_n^d$ | 50 | 31 | 0.1950 | 31 | 0.1880 | 31 | 0.1797 | 44 | 0.0230 | 31 | 0.1993 |
| | | | 100 | 58 | 0.2381 | 60 | 0.2307 | 60 | 0.2110 | 87 | 0.0227 | 59 | 0.2153 |
| | | $2C_n^d$ | 50 | 9 | 0.9160 | 9 | 0.9130 | 9 | 0.9120 | 36 | 0.2693 | 9 | 0.9097 |
| | | | 100 | 11 | 0.9450 | 12 | 0.9380 | 11 | 0.9477 | 68 | 0.3363 | 11 | 0.9477 |
| | 100 | $1.2C_n^d$ | 50 | 31 | 0.2021 | 31 | 0.1837 | 30 | 0.1913 | 44 | 0.0233 | 31 | 0.1883 |
| | | | 100 | 61 | 0.2334 | 59 | 0.2293 | 60 | 0.2160 | 88 | 0.0217 | 59 | 0.2267 |
| | | $2C_n^d$ | 50 | 9 | 0.9057 | 9 | 0.9003 | 9 | 0.9060 | 36 | 0.2927 | 9 | 0.9133 |
| | | | 100 | 11 | 0.9427 | 12 | 0.9490 | 11 | 0.9483 | 67 | 0.3347 | 11 | 0.9510 |
| | 500 | $1.2C_n^d$ | 50 | 30 | 0.1923 | 31 | 0.1827 | 31 | 0.1907 | 43 | 0.0317 | 31 | 0.1827 |
| | | | 100 | 61 | 0.2138 | 59 | 0.2247 | 61 | 0.2143 | 86 | 0.0243 | 59 | 0.2307 |
| | | $2C_n^d$ | 50 | 9 | 0.9080 | 9 | 0.9167 | 9 | 0.9093 | 35 | 0.2943 | 9 | 0.9030 |
| | | | 100 | 12 | 0.9477 | 11 | 0.9493 | 11 | 0.9517 | 68 | 0.3373 | 11 | 0.9453 |

## A.2 Additional tables of Chapter 4

Table A.5.: Obtained values of $(a, b)$ when $\Sigma = I_d$ for different values of $n, d$ and $k_I^1$ and $C_n^d \equiv C_n^d(0.05)$. This is the counterpart of Table 4.1 for the non-robust version.

| | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $k_I^1 = 50$ | | $k_I^1 = 100$ | | $k_I^1 = 50$ | | $k_I^1 = 100$ | | $k_I^1 = 50$ | | $k_I^1 = 100$ | |
| $d$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ | $a$ | $b$ |
| 5 | 0.0585 | 5.1122 | 0.0293 | 5.2463 | 0.0597 | 4.9761 | 0.0306 | 5.0557 | 0.0642 | 5.0333 | 0.0318 | 5.1007 |
| 50 | 0.0326 | 4.5159 | 0.0163 | 4.7797 | 0.0328 | 4.3838 | 0.0165 | 4.6543 | 0.0337 | 4.3975 | 0.0168 | 4.6273 |
| 100 | 0.0299 | 4.2508 | 0.0150 | 4.5520 | 0.0294 | 4.1322 | 0.0151 | 4.3927 | 0.0305 | 4.1120 | 0.0154 | 4.3485 |
| 500 | 0.0269 | 3.8995 | 0.0134 | 4.1358 | 0.0265 | 3.7599 | 0.0134 | 3.9849 | 0.0264 | 3.6948 | 0.0131 | 3.8957 |
| 1000 | 0.0260 | 3.7950 | 0.0130 | 4.0536 | 0.0258 | 3.6550 | 0.0132 | 3.8897 | 0.0260 | 3.5884 | 0.0130 | 3.7879 |

Table A.6.: Approximated values of $b_\Sigma$ for $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$, and different values of $n$ and $d$. The $a$'s are the values obtained in Table 4.1 for $l_I^1 = 50, 100$. The values of $\hat{l}_1^1$ are also shown.

| | | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $l_I^1$=50 | | $l_I^1$=100 | | $l_I^1$=50 | | $l_I^1$=100 | | $l_I^1$=50 | | $l_I^1$=100 | |
| $d$ | $\Sigma$ | $b_\Sigma$ | $\hat{l}_1^1$ | $b_\Sigma$ | $\hat{l}_1^1$ | $b_\Sigma$ | $\hat{l}_1^1$ | $b_\Sigma$ | $\hat{l}_1^1$ | $b_\Sigma$ | $\hat{l}_1^1$ | $b_\Sigma$ | $\hat{l}_1^1$ |
| 5 | $\Sigma_1^d$ | 5.0184 | 55 | 5.1241 | 108 | 4.9185 | 54 | 4.9781 | 107 | 5.0139 | 53 | 5.0735 | 109 |
| | $\Sigma_2^d$ | 6.0112 | 52 | 6.3086 | 102 | 5.4728 | 47 | 5.6175 | 101 | 5.1618 | 50 | 5.2590 | 100 |
| | $\Sigma_3^d$ | 6.0573 | 73 | 6.4004 | 157 | 5.4914 | 78 | 5.6877 | 161 | 5.1718 | 76 | 5.2804 | 156 |
| | $\Sigma_4^d$ | 6.0346 | 49 | 6.3624 | 102 | 5.4755 | 51 | 5.6643 | 99 | 5.1693 | 50 | 5.2697 | 102 |
| 50 | $\Sigma_1^d$ | 5.0236 | 49 | 5.3996 | 99 | 4.6351 | 51 | 4.9236 | 97 | 4.4517 | 48 | 4.7039 | 100 |
| | $\Sigma_2^d$ | 4.9995 | 50 | 5.4076 | 100 | 4.6284 | 48 | 4.9289 | 99 | 4.4522 | 48 | 4.6932 | 101 |
| | $\Sigma_3^d$ | 5.1413 | 189 | 5.4932 | 389 | 4.6374 | 192 | 4.9504 | 384 | 4.4439 | 180 | 4.6858 | 381 |
| | $\Sigma_4^d$ | 5.0236 | 51 | 5.3798 | 125 | 4.6194 | 49 | 4.9075 | 103 | 4.4520 | 51 | 4.7080 | 99 |
| 100 | $\Sigma_1^d$ | 4.7425 | 51 | 5.0891 | 101 | 4.3529 | 51 | 4.6494 | 93 | 4.1477 | 50 | 4.4083 | 99 |
| | $\Sigma_2^d$ | 4.7425 | 50 | 5.1213 | 100 | 4.3538 | 52 | 4.6387 | 99 | 4.1471 | 49 | 4.4056 | 98 |
| | $\Sigma_3^d$ | 4.8813 | 260 | 5.1857 | 541 | 4.3539 | 269 | 4.6494 | 525 | 4.1399 | 260 | 4.3691 | 530 |
| | $\Sigma_4^d$ | 4.7523 | 48 | 5.1052 | 110 | 4.3497 | 51 | 4.6467 | 98 | 4.1458 | 49 | 4.3922 | 100 |
| 500 | $\Sigma_1^d$ | 4.3012 | 51 | 4.6556 | 100 | 3.9438 | 49 | 4.2325 | 99 | 3.7279 | 48 | 3.9533 | 102 |
| | $\Sigma_2^d$ | 4.3244 | 48 | 4.6361 | 99 | 3.9701 | 49 | 4.2069 | 99 | 3.7421 | 51 | 3.9509 | 97 |
| | $\Sigma_3^d$ | 4.3244 | 580 | 4.6946 | 1162 | 4.0248 | 590 | 4.2460 | 1180 | 3.7421 | 580 | 3.9143 | 1195 |
| | $\Sigma_4^d$ | 4.3219 | 49 | 4.6166 | 100 | 3.9613 | 51 | 4.2216 | 101 | 3.7509 | 49 | 3.9475 | 102 |
| 1000 | $\Sigma_1^d$ | 4.2058 | 50 | 4.5254 | 100 | 3.8623 | 48 | 4.1146 | 101 | 3.6276 | 49 | 3.8192 | 98 |
| | $\Sigma_2^d$ | 4.2272 | 50 | 4.5385 | 100 | 3.8442 | 48 | 4.1094 | 101 | 3.6236 | 50 | 3.8363 | 101 |
| | $\Sigma_3^d$ | 4.3129 | 830 | 4.6166 | 1678 | 3.9221 | 805 | 4.1094 | 1610 | 3.6080 | 834 | 3.8168 | 1620 |
| | $\Sigma_4^d$ | 4.2272 | 49 | 4.5385 | 102 | 3.8442 | 51 | 4.1094 | 100 | 3.6080 | 50 | 3.8326 | 101 |

Table A.7.: Approximated values of $b_\Sigma$ for $\Sigma = \Sigma_i^d$ with $i = 1, \ldots, 4$, and different values of $n$ and $d$. The $a$'s are the values obtained in Table A.5 for $k_I^1 = 50, 100$. The values of $\hat{k}_1^1$ are also shown. This table is the non-robust counterpart of Table A.6.

| | | $n = 50$ | | | | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k_I^1$=50 | | $k_I^1$=100 | | $k_I^1$=50 | | $k_I^1$=100 | | $k_I^1$=50 | | $k_I^1$=100 | |
| $d$ | $\Sigma$ | $b_\Sigma$ | $\hat{k}_1^1$ | $b_\Sigma$ | $\hat{k}_1^1$ | $b_\Sigma$ | $\hat{k}_1^1$ | $b_\Sigma$ | $\hat{k}_1^1$ | $b_\Sigma$ | $\hat{k}_1^1$ | $b_\Sigma$ | $\hat{k}_1^1$ |
| 5 | $\Sigma_1^d$ | 5.0184 | 55 | 5.1241 | 108 | 4.9185 | 54 | 4.9781 | 107 | 5.0139 | 53 | 5.0735 | 109 |
| | $\Sigma_2^d$ | 5.0727 | 52 | 5.1954 | 103 | 4.9281 | 53 | 5.0164 | 102 | 5.0309 | 53 | 5.0908 | 106 |
| | $\Sigma_3^d$ | 5.0687 | 79 | 5.1814 | 161 | 4.9569 | 80 | 5.0360 | 154 | 5.0333 | 78 | 5.1004 | 160 |
| | $\Sigma_4^d$ | 5.0727 | 50 | 5.2247 | 100 | 4.9617 | 51 | 5.0544 | 100 | 5.0285 | 51 | 5.1004 | 100 |
| 50 | $\Sigma_1^d$ | 4.5155 | 49 | 4.7796 | 99 | 4.3796 | 50 | 4.6453 | 98 | 4.3933 | 49 | 4.6271 | 100 |
| | $\Sigma_2^d$ | 4.4896 | 50 | 4.7704 | 100 | 4.3796 | 49 | 4.6362 | 98 | 4.3972 | 49 | 4.6183 | 99 |
| | $\Sigma_3^d$ | 4.4984 | 188 | 4.7611 | 390 | 4.3796 | 193 | 4.6453 | 386 | 4.3890 | 188 | 4.6228 | 380 |
| | $\Sigma_4^d$ | 4.4984 | 51 | 4.7727 | 127 | 4.3833 | 49 | 4.6452 | 102 | 4.3954 | 50 | 4.6228 | 101 |
| 100 | $\Sigma_1^d$ | 4.2495 | 50 | 4.5343 | 100 | 4.1321 | 51 | 4.3799 | 94 | 4.1040 | 50 | 4.3443 | 99 |
| | $\Sigma_2^d$ | 4.2507 | 50 | 4.5343 | 101 | 4.1319 | 51 | 4.3831 | 99 | 4.1115 | 49 | 4.3379 | 98 |
| | $\Sigma_3^d$ | 4.2497 | 263 | 4.5166 | 540 | 4.1161 | 270 | 4.3585 | 527 | 4.1120 | 265 | 4.3316 | 529 |
| | $\Sigma_4^d$ | 4.2466 | 49 | 4.5471 | 117 | 4.1241 | 51 | 4.3876 | 99 | 4.1118 | 50 | 4.3474 | 98 |
| 500 | $\Sigma_1^d$ | 3.8919 | 50 | 4.1353 | 100 | 3.7597 | 50 | 3.9772 | 99 | 3.6877 | 49 | 3.8948 | 103 |
| | $\Sigma_2^d$ | 3.8919 | 49 | 4.1197 | 99 | 3.7563 | 51 | 3.9772 | 98 | 3.6877 | 51 | 3.8954 | 102 |
| | $\Sigma_3^d$ | 3.8988 | 584 | 4.1277 | 1162 | 3.7595 | 594 | 3.9791 | 1183 | 3.6877 | 593 | 3.8806 | 1199 |
| | $\Sigma_4^d$ | 3.8919 | 50 | 4.1197 | 99 | 3.7581 | 51 | 3.9808 | 99 | 3.6805 | 50 | 3.8955 | 103 |
| 1000 | $\Sigma_1^d$ | 3.7922 | 50 | 4.0427 | 100 | 3.6539 | 49 | 3.8745 | 102 | 3.5745 | 50 | 3.7806 | 97 |
| | $\Sigma_2^d$ | 3.7949 | 50 | 4.0516 | 100 | 3.6532 | 49 | 3.8745 | 98 | 3.5849 | 50 | 3.7856 | 99 |
| | $\Sigma_3^d$ | 3.7803 | 833 | 4.0378 | 1688 | 3.6549 | 816 | 3.8594 | 1618 | 3.5745 | 832 | 3.7732 | 1624 |
| | $\Sigma_4^d$ | 3.7929 | 50 | 4.0477 | 101 | 3.6515 | 50 | 3.8859 | 98 | 3.5858 | 50 | 3.7875 | 101 |

Table A.8.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$ with the non-robust procedure, for $n = 50$ and several values of $d$ and $\Sigma$. We also show the sample means of $K_n$. This table is the non-robust version of Table 4.6.

| $d$ | $k_I^1$ | $\hat{k}_I^1$ | $I_d$ | $\hat{k}_1^1$ | $\Sigma_1^d$ | $\hat{k}_2^1$ | $\Sigma_2^d$ | $\hat{k}_3^1$ | $\Sigma_3^d$ | $\hat{k}_4^1$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 50 | 0.0464 | 50 | 0.0533 | 49 | 0.0491 | 188 | 0.0537 | 50 | 0.0473 |
| | 100 | 96 | 0.0516 | 99 | 0.0528 | 98 | 0.0550 | 384 | 0.0519 | 100 | 0.0514 |
| 100 | 50 | 50 | 0.0478 | 50 | 0.0509 | 51 | 0.0529 | 270 | 0.0560 | 50 | 0.0538 |
| | 100 | 101 | 0.0509 | 97 | 0.0504 | 100 | 0.0509 | 535 | 0.0484 | 100 | 0.0558 |
| 500 | 50 | 50 | 0.0517 | 49 | 0.0482 | 50 | 0.0491 | 591 | 0.0530 | 49 | 0.0527 |
| | 100 | 99 | 0.0504 | 99 | 0.0512 | 99 | 0.0486 | 1175 | 0.0468 | 100 | 0.0500 |
| 1000 | 50 | 51 | 0.0517 | 50 | 0.0520 | 51 | 0.0522 | 824 | 0.0496 | 51 | 0.0553 |
| | 100 | 101 | 0.0498 | 98 | 0.0533 | 102 | 0.0500 | 1653 | 0.0510 | 101 | 0.0492 |

**Table A.9.:** Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = C_n^d$. We also show the sample means of $L_n$. This table is Table 4.6 expanded to $n = 100, 500$ and its description applies.

| $d$ | $n$ | $l_I^1$ | $\hat{l}_I^1$ | $I_d$ | $\hat{l}_1^1$ | $\Sigma_1^d$ | $\hat{l}_2^1$ | $\Sigma_2^d$ | $\hat{l}_3^1$ | $\Sigma_3^d$ | $\hat{l}_4^1$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | 50 | 51 | 0.0528 | 49 | 0.0571 | 49 | 0.0541 | 186 | 0.0668 | 50 | 0.0569 |
| | | 100 | 98 | 0.0560 | 99 | 0.0558 | 103 | 0.0553 | 366 | 0.0580 | 99 | 0.0572 |
| | 100 | 50 | 50 | 0.0497 | 52 | 0.0528 | 50 | 0.0511 | 186 | 0.0566 | 49 | 0.0477 |
| | | 100 | 102 | 0.0478 | 101 | 0.0535 | 100 | 0.0535 | 382 | 0.0510 | 101 | 0.0539 |
| | 500 | 50 | 50 | 0.0501 | 50 | 0.0537 | 50 | 0.0484 | 194 | 0.0537 | 50 | 0.0558 |
| | | 100 | 99 | 0.0533 | 99 | 0.0484 | 99 | 0.0492 | 387 | 0.0499 | 98 | 0.0487 |
| 100 | 50 | 50 | 49 | 0.0507 | 48 | 0.0496 | 50 | 0.0501 | 249 | 0.0628 | 50 | 0.0489 |
| | | 100 | 100 | 0.0538 | 101 | 0.0519 | 100 | 0.0526 | 526 | 0.0603 | 98 | 0.0494 |
| | 100 | 50 | 50 | 0.0535 | 51 | 0.0496 | 48 | 0.0548 | 251 | 0.0560 | 49 | 0.0528 |
| | | 100 | 99 | 0.0479 | 99 | 0.0495 | 99 | 0.0490 | 508 | 0.0540 | 99 | 0.0481 |
| | 500 | 50 | 50 | 0.0547 | 49 | 0.0538 | 50 | 0.0538 | 264 | 0.0573 | 50 | 0.0536 |
| | | 100 | 97 | 0.0557 | 98 | 0.0572 | 100 | 0.0523 | 516 | 0.0494 | 97 | 0.0507 |
| 500 | 50 | 50 | 49 | 0.0481 | 50 | 0.0507 | 50 | 0.0518 | 552 | 0.0628 | 50 | 0.0483 |
| | | 100 | 100 | 0.0520 | 102 | 0.0509 | 99 | 0.0545 | 1111 | 0.0589 | 101 | 0.0538 |
| | 100 | 50 | 49 | 0.0543 | 50 | 0.0532 | 50 | 0.054 | 571 | 0.0588 | 50 | 0.0527 |
| | | 100 | 101 | 0.0522 | 102 | 0.0518 | 100 | 0.0511 | 1136 | 0.0549 | 99 | 0.0550 |
| | 500 | 50 | 51 | 0.0508 | 50 | 0.0530 | 50 | 0.0502 | 596 | 0.0538 | 50 | 0.0506 |
| | | 100 | 103 | 0.0520 | 100 | 0.0540 | 102 | 0.0470 | 1199 | 0.0489 | 102 | 0.0505 |
| 1000 | 50 | 50 | 50 | 0.0496 | 50 | 0.0538 | 49 | 0.0534 | 790 | 0.0586 | 50 | 0.0500 |
| | | 100 | 100 | 0.0520 | 101 | 0.0476 | 102 | 0.0507 | 1601 | 0.0549 | 99 | 0.0553 |
| | 100 | 50 | 49 | 0.0564 | 50 | 0.0556 | 50 | 0.0509 | 775 | 0.0599 | 50 | 0.0513 |
| | | 100 | 100 | 0.0516 | 101 | 0.0518 | 101 | 0.0534 | 1650 | 0.0545 | 101 | 0.0571 |
| | 500 | 50 | 51 | 0.0588 | 50 | 0.0508 | 49 | 0.0536 | 830 | 0.0543 | 49 | 0.0539 |
| | | 100 | 100 | 0.0500 | 100 | 0.0529 | 98 | 0.0569 | 1688 | 0.0444 | 100 | 0.0568 |

**Table A.10.:** Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d, r = 1.2, 2$ with the non-robust procedure with $n = 50$. We also show the sample means of $K_n$. It is the non-robust version of Table 4.7.

| $d$ | $\|X\|_\Sigma$ | $k_I^r$ | $\hat{k}_I^r$ | $I_d$ | $\hat{k}_1^r$ | $\Sigma_1^d$ | $\hat{k}_2^r$ | $\Sigma_2^d$ | $\hat{k}_3^r$ | $\Sigma_3^d$ | $\hat{k}_4^r$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | $1.2C_n^d$ | 50 | 46 | .2630 | 50 | .2398 | 48 | .2486 | 163 | .1774 | 48 | .2908 |
| | | 100 | 92 | .2990 | 94 | .2500 | 93 | .2864 | 325 | .1772 | 89 | .2924 |
| | $2C_n^d$ | 50 | 10 | .9106 | 11 | .8975 | 10 | .9054 | 39 | .6933 | 9 | .9138 |
| | | 100 | 12 | .9374 | 14 | .9293 | 13 | .9361 | 61 | .7377 | 12 | .9467 |
| 100 | $1.2C_n^d$ | 50 | 45 | .2764 | 46 | .2542 | 47 | .2538 | 218 | .1808 | 46 | .2664 |
| | | 100 | 90 | .2914 | 91 | .2730 | 91 | .2792 | 448 | .1770 | 89 | .2846 |
| | $2C_n^d$ | 50 | 10 | .9053 | 10 | .8998 | 10 | .9055 | 56 | .6736 | 10 | .9073 |
| | | 100 | 13 | .9338 | 14 | .9275 | 14 | .9350 | 90 | .7124 | 13 | .9337 |
| 500 | $1.2C_n^d$ | 50 | 47 | .2532 | 46 | .2412 | 46 | .2532 | 495 | .1732 | 48 | .2496 |
| | | 100 | 89 | .2952 | 88 | .2772 | 89 | .2980 | 977 | .1844 | 89 | .2948 |
| | $2C_n^d$ | 50 | 10 | .9014 | 10 | .8968 | 11 | .8993 | 127 | .6536 | 11 | .8973 |
| | | 100 | 13 | .9391 | 14 | .9366 | 14 | .9357 | 199 | .6942 | 14 | .9418 |
| 1000 | $1.2C_n^d$ | 50 | 46 | .253 | 46 | .261 | 49 | .285 | 769 | .158 | 48 | .241 |
| | | 100 | 90 | .277 | 95 | .306 | 92 | .262 | 1430 | .194 | 96 | .301 |
| | $2C_n^d$ | 50 | 10 | .8964 | 11 | .8956 | 10 | .8997 | 178 | .6486 | 10 | .9023 |
| | | 100 | 14 | .9361 | 14 | .9356 | 14 | .9360 | 285 | .6907 | 14 | .9346 |

Table A.11.: Estimation of the probability of declaring as an outlier a vector such that $\|\mathbf{X}\|_\Sigma = rC_n^d$ with $r = 1.2, 2$, for several values of $n, d$ and $\Sigma$. It is the expansion of part of the Table 4.7 to $n = 100$ and $n = 500$. We also show the sample means of $L_n$.

| $d$ | $n$ | $\|\mathbf{X}\|_\Sigma$ | $l_I^1$ | $\hat{l}_I^{1.2}$ | $I_d$ | $\hat{l}_1^{1.2}$ | $\Sigma_1^d$ | $\hat{l}_2^{1.2}$ | $\Sigma_2^d$ | $\hat{l}_3^{1.2}$ | $\Sigma_3^d$ | $\hat{l}_4^{1.2}$ | $\Sigma_4^d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 50 | $1.2C_n^d$ | 50 | 48 | 0.2378 | 48 | 0.2247 | 48 | 0.2338 | 163 | 0.1752 | 48 | 0.2333 |
|  |  |  | 100 | 93 | 0.2729 | 96 | 0.2412 | 95 | 0.2617 | 313 | 0.1867 | 92 | 0.2639 |
|  |  | $2C_n^d$ | 50 | 12 | 0.8817 | 13 | 0.8660 | 12 | 0.8830 | 47 | 0.6575 | 12 | 0.8912 |
|  |  |  | 100 | 16 | 0.9259 | 19 | 0.9061 | 16 | 0.9153 | 74 | 0.6985 | 16 | 0.9229 |
|  | 100 | $1.2C_n^d$ | 50 | 47 | 0.2618 | 48 | 0.2403 | 47 | 0.2530 | 163 | 0.1764 | 47 | 0.2626 |
|  |  |  | 100 | 89 | 0.3057 | 91 | 0.2731 | 88 | 0.2935 | 307 | 0.1968 | 88 | 0.3006 |
|  |  | $2C_n^d$ | 50 | 50 | 0.9101 | 11 | 0.8880 | 10 | 0.9009 | 40 | 0.6801 | 10 | 0.9053 |
|  |  |  | 100 | 12 | 0.9349 | 15 | 0.9257 | 13 | 0.9372 | 62 | 0.7393 | 13 | 0.9375 |
|  | 500 | $1.2C_n^d$ | 50 | 44 | 0.3079 | 45 | 0.2702 | 44 | 0.2897 | 152 | 0.1957 | 45 | 0.2991 |
|  |  |  | 100 | 82 | 0.3471 | 87 | 0.2985 | 84 | 0.3306 | 296 | 0.2054 | 82 | 0.3412 |
|  |  | $2C_n^d$ | 50 | 8 | 0.9193 | 9 | 0.9116 | 9 | 0.9185 | 35 | 0.7134 | 8 | 0.9205 |
|  |  |  | 100 | 10 | 0.9506 | 12 | 0.9391 | 11 | 0.9509 | 51 | 0.7658 | 10 | 0.9502 |
| 100 | 50 | $1.2C_n^d$ | 50 | 48 | 0.2235 | 49 | 0.2093 | 48 | 0.2146 | 223 | 0.1729 | 49 | 0.2191 |
|  |  |  | 100 | 97 | 0.2387 | 97 | 0.2236 | 95 | 0.2320 | 460 | 0.1723 | 96 | 0.2487 |
|  |  | $2C_n^d$ | 50 | 13 | 0.8829 | 13 | 0.8678 | 13 | 0.8734 | 70 | 0.6289 | 13 | 0.8743 |
|  |  |  | 100 | 18 | 0.9150 | 19 | 0.9081 | 18 | 0.9115 | 113 | 0.6738 | 18 | 0.9160 |
|  | 100 | $1.2C_n^d$ | 50 | 46 | 0.2501 | 47 | 0.2331 | 47 | 0.2527 | 218 | 0.1803 | 47 | 0.2571 |
|  |  |  | 100 | 90 | 0.2795 | 90 | 0.2605 | 91 | 0.2766 | 418 | 0.1832 | 90 | 0.2884 |
|  |  | $2C_n^d$ | 50 | 10 | 0.9009 | 11 | 0.8924 | 11 | 0.9019 | 56 | 0.6631 | 10 | 0.9003 |
|  |  |  | 100 | 13 | 0.9369 | 15 | 0.9251 | 14 | 0.9351 | 90 | 0.7050 | 14 | 0.9353 |
|  | 500 | $1.2C_n^d$ | 50 | 45 | 0.2930 | 46 | 0.2746 | 45 | 0.2863 | 209 | 0.1917 | 45 | 0.3002 |
|  |  |  | 100 | 84 | 0.3319 | 85 | 0.3080 | 85 | 0.3220 | 409 | 0.2026 | 84 | 0.3296 |
|  |  | $2C_n^d$ | 50 | 9 | 0.9240 | 9 | 0.9109 | 9 | 0.9134 | 49 | 0.6858 | 9 | 0.9182 |
|  |  |  | 100 | 11 | 0.9480 | 12 | 0.9379 | 11 | 0.9434 | 74 | 0.7329 | 11 | 0.9446 |
| 500 | 50 | $1.2C_n^d$ | 50 | 50 | 0.2160 | 48 | 0.2132 | 49 | 0.2168 | 518 | 0.1711 | 50 | 0.2198 |
|  |  |  | 100 | 97 | 0.2454 | 99 | 0.2375 | 96 | 0.2399 | 973 | 0.1761 | 97 | 0.2412 |
|  |  | $2C_n^d$ | 50 | 13 | 0.8771 | 13 | 0.8617 | 13 | 0.8780 | 150 | 0.6139 | 13 | 0.8726 |
|  |  |  | 100 | 18 | 0.9166 | 18 | 0.9185 | 18 | 0.9075 | 249 | 0.6513 | 18 | 0.9090 |
|  | 100 | $1.2C_n^d$ | 50 | 46 | 0.2647 | 47 | 0.2575 | 47 | 0.2569 | 474 | 0.1766 | 47 | 0.2558 |
|  |  |  | 100 | 91 | 0.2795 | 91 | 0.2810 | 90 | 0.2737 | 963 | 0.1821 | 91 | 0.2766 |
|  |  | $2C_n^d$ | 50 | 11 | 0.8985 | 11 | 0.8992 | 11 | 0.8981 | 124 | 0.6488 | 11 | 0.8949 |
|  |  |  | 100 | 14 | 0.9355 | 15 | 0.9307 | 14 | 0.9273 | 204 | 0.6901 | 14 | 0.9313 |
|  | 500 | $1.2C_n^d$ | 50 | 46 | 0.2690 | 46 | 0.2701 | 46 | 0.2831 | 483 | 0.1844 | 46 | 0.2709 |
|  |  |  | 100 | 89 | 0.3196 | 88 | 0.3139 | 87 | 0.3191 | 961 | 0.1954 | 86 | 0.3104 |
|  |  | $2C_n^d$ | 50 | 9 | 0.9089 | 9 | 0.9133 | 9 | 0.9137 | 114 | 0.6708 | 9 | 0.9113 |
|  |  |  | 100 | 12 | 0.9451 | 12 | 0.9421 | 12 | 0.9450 | 173 | 0.7119 | 12 | 0.9410 |
| 1000 | 50 | $1.2C_n^d$ | 50 | 49 | 0.2202 | 51 | 0.2136 | 49 | 0.2159 | 700 | 0.1632 | 49 | 0.2156 |
|  |  |  | 100 | 98 | 0.2470 | 97 | 0.2338 | 97 | 0.2429 | 1383 | 0.1616 | 96 | 0.2366 |
|  |  | $2C_n^d$ | 50 | 13 | 0.8797 | 13 | 0.8728 | 13 | 0.8729 | 214 | 0.6128 | 13 | 0.8674 |
|  |  |  | 100 | 19 | 0.9116 | 19 | 0.9134 | 19 | 0.9093 | 360 | 0.6551 | 19 | 0.9124 |
|  | 100 | $1.2C_n^d$ | 50 | 46 | 0.2545 | 46 | 0.2575 | 47 | 0.2536 | 651 | 0.1720 | 46 | 0.2583 |
|  |  |  | 100 | 92 | 0.2883 | 92 | 0.2828 | 91 | 0.2809 | 1373 | 0.1778 | 90 | 0.2835 |
|  |  | $2C_n^d$ | 50 | 11 | 0.8994 | 11 | 0.8986 | 11 | 0.8947 | 182 | 0.6508 | 10 | 0.8956 |
|  |  |  | 100 | 14 | 0.9315 | 15 | 0.9300 | 14 | 0.9300 | 283 | 0.6894 | 14 | 0.9288 |
|  | 500 | $1.2C_n^d$ | 50 | 45 | 0.2790 | 46 | 0.2726 | 46 | 0.2723 | 689 | 0.1823 | 45 | 0.2805 |
|  |  |  | 100 | 88 | 0.3172 | 85 | 0.3210 | 84 | 0.3213 | 1333 | 0.1963 | 86 | 0.3173 |
|  |  | $2C_n^d$ | 50 | 9 | 0.9143 | 9 | 0.9105 | 9 | 0.9136 | 154 | 0.6683 | 9 | 0.9080 |
|  |  |  | 100 | 12 | 0.9409 | 12 | 0.9412 | 12 | 0.9477 | 238 | 0.7128 | 12 | 0.9432 |

Table A.12.: Proportion of outliers found in a clean data set for several covariance matrices.

| | | MDP | | | | PCOut | | | | RP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ |
| 50 | 50 | .2460 | .1372 | .2509 | .1348 | .1118 | .1026 | .1082 | .1043 | .1077 | .1160 | .1326 | .1095 |
| | 500 | .0884 | .0579 | .3748 | .1291 | .0963 | .0974 | .0964 | .0946 | .1104 | .1104 | .1368 | .1099 |
| | 1000 | — | — | — | — | .0978 | .1043 | .0977 | .0970 | .1144 | .1145 | .1462 | .1109 |
| 100 | 50 | .2209 | .0758 | .0797 | .0746 | .1111 | .1013 | .1056 | .1029 | .1019 | .1046 | .1088 | .1027 |
| | 500 | .0702 | .0552 | .2310 | .0833 | .0803 | .0833 | .0794 | .0788 | .1069 | .1106 | .1228 | .1090 |
| | 1000 | — | — | — | — | .0813 | .0809 | .0806 | .0784 | .1128 | .1121 | .1249 | .1116 |

Table A.13.: Samples contain 10% of real outliers. Columns show the proportion of them correctly identified.

| | | MDP | | | | PCOut | | | | RP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ | $\Sigma_1^d$ | $\Sigma_2^d$ | $\Sigma_3^d$ | $\Sigma_4^d$ |
| 50 | 50 | .2545 | .1865 | .2886 | .1942 | .2064 | .1636 | .1724 | .1596 | .2736 | .2844 | .2412 | .2868 |
| | 500 | .0933 | .0803 | .1826 | .1859 | .1196 | .1224 | .1104 | .1352 | .1636 | .1680 | .1776 | .1604 |
| | 1000 | — | — | — | — | .1136 | .1312 | .1184 | .1248 | .1440 | .1452 | .1612 | .1592 |
| 100 | 50 | .2241 | .1282 | .2581 | .1320 | .2736 | .2360 | .2482 | .2376 | .2812 | .3020 | .2310 | .2986 |
| | 500 | .0747 | .0935 | .3330 | .1419 | .0996 | .0964 | .0990 | .0952 | .1636 | .1638 | .1584 | .1760 |
| | 1000 | — | — | — | — | .0874 | .0982 | .0892 | .0972 | .1548 | .1504 | .1598 | .1476 |

## A.3 Additional tables and further simulations of Chapter 5

We perform next an independent simulation study to elucidate the reasons of the somehow surprising conclusion *(viii)* in Section 5.3.2. The tests of the simulation study in such section are benchmarked with respect to the Invariant Likelihood Ratio Test (ILRT) for testing uniformity against the alternative (5.58). If $f_0$ denotes the uniform pdf on $\Omega^{d-1}$, the ILRT for testing uniformity against (5.58) for a specified $0 < \kappa < 1$, that is, for testing

$$\mathbf{H}_0 : f = f_0 \quad \text{vs.} \quad \mathbf{H}_{1,\kappa} : f \in \{f_{\boldsymbol{\mu},\kappa} : \boldsymbol{\mu} \in \Omega^{d-1}\} \tag{A.1}$$

is the test that rejects for large values of the ILRT statistic:

$$\mathrm{L}_\kappa := \int_{\Omega^{d-1}} \prod_{i=1}^{n} f_{\boldsymbol{\mu},\gamma}(\mathbf{X}_i)\, \omega_{d-1}(\mathrm{d}\gamma).$$

We focus on the simplest DGP (5.58) among CvM, AD, and Rt that admits a tractable ILRT. This DGP is Rt with $t = 1/2$ for $d = 2$, therefore coinciding with [4]'s "semicircle deviation". In this setting, each sample observation can be parametrized as $\mathbf{X}_i = (\cos\Theta_i, \sin\Theta_i)'$ for $\Theta_i \in [0, 2\pi)$ and $f^{\mathrm{Rt}}(z) = 1_{\{z\geq 0\}} + 1/2$. Thus the ILRT statistic becomes

$$\mathrm{L}_\kappa = \int_0^{2\pi} \prod_{i=1}^n g_\kappa(\Theta_i - \theta)\,\mathrm{d}\theta = \sum_{j=1}^{2n} \int_{I_j} \prod_{i=1}^n g_\kappa(\Theta_i - \theta)\,\mathrm{d}\theta = \sum_{j=1}^{2n} \prod_{i=1}^n g_\kappa(\theta_j)\ell_j,$$

where $g_\kappa(\varphi) := \frac{1}{2\pi}\left\{1 + \kappa\left(1_{\{\cos(\varphi)\geq 0\}} - \frac{1}{2}\right)\right\}$, $\{I_j\}_{j=1}^{2n}$ are certain intervals defined below, and $\ell_j$ is the length of $I_j$ and $\theta_j$ its midpoint. The intervals $\{I_j\}_{j=1}^{2n}$ are constructed by first augmenting the sample $\{\Theta_i\}_{i=1}^n$ to $\{\tilde{\Theta}_i\}_{i=1}^{2n}$, where $\tilde{\Theta}_i = (\Theta_i - \pi/2) \mod 2\pi$ and $\tilde{\Theta}_{i+n} = (\Theta_i + \pi/2) \mod 2\pi$ for $i = 1, \ldots, n$, and then setting $I_j := [\tilde{\Theta}_{(j)}, \tilde{\Theta}_{(j+1)})$, $j = 1, \ldots, 2n$, where $\tilde{\Theta}_{(2n+1)} := \tilde{\Theta}_{(1)} + 2\pi$.

We consider $M = 10^8$ Monte Carlo replicates to reduce the Monte Carlo noise and capture smaller power effects. We employ the tests considered in Section 5.3.2 (Ajne is omitted since it coincides with Rt for $t = 1/2$) plus the ILRT for (A.1). We use sample size $n = 50$ and the local deviations $\kappa = 0.05k$, $k = 0, \ldots, 20$. As in Section 5.3.2, the statistics are calibrated under the null hypothesis by Monte Carlo. The obtained empirical powers are collected in Figures A.1 and A.2 and give the following conclusions:

($a$). The optimality of the ILRT is verified and evidenced to be smaller than $10^{-3}$ for the investigated $\kappa$'s (Figure A.1). Therefore, the power gap between the optimal test for (A.1) and other tests is fairly small, as is also reflected in the virtual equivalence of the powers shown in Figure A.2. The Monte Carlo noise explains that the empirical power of Rt is larger than the power of the ILRT.

($b$). The Rt test is locally equivalent to the ILRT for $\kappa \approx 0$, both being indistinguishable (at the $95\%$ confidence) within the Monte Carlo noise until $\kappa$ approaches $0.10$. The Rt test clearly outperforms the remaining tests except the ILRT.

($c$). An apparently high number of Monte Carlo replicates such as $10^6$ is still insufficient to fully capture optimalities in the investigated DGP. We conjecture this is a prevalent issue with all the alternatives (5.58) investigated in Section 5.3.2.

($d$). Unsurprisingly, the Bingham and Giné tests are blind against this alternative and have the nominal significance level as power. A difference in power is evidenced for the Rayleigh and Ajne test (here acting as the Rt test), yet again it is fairly small.

We conclude mentioning that this kind DGP was already considered in Stephens [126]. In particular, his Table 3 compares the powers of Ajne, Watson, and Kuiper [92] tests for the circle at significance level $10\%$ using $5000$ Monte Carlo replicates. However, his study does not show that the Ajne test is significantly (with a 95% confidence) more powerful than the competing tests for this alternative.

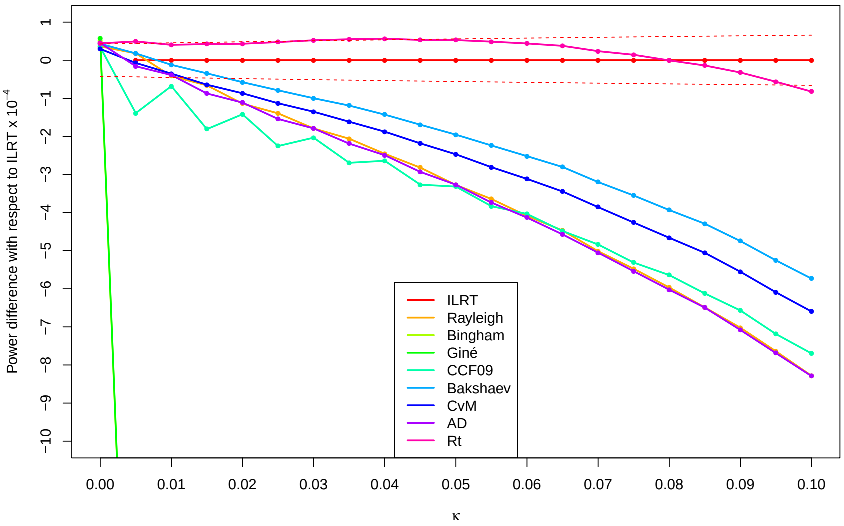


**Figure A.1.:** Difference of empirical powers with respect to the ILRT for different deviations $\kappa$. The dashed lines represent the 99% confidence interval about the ILRT power. For $\kappa = 0$, the testing problem (A.1) is undefined, and its power is replaced by the significance level, 5%. The vertical axis is on the scale $10^{-4}$.

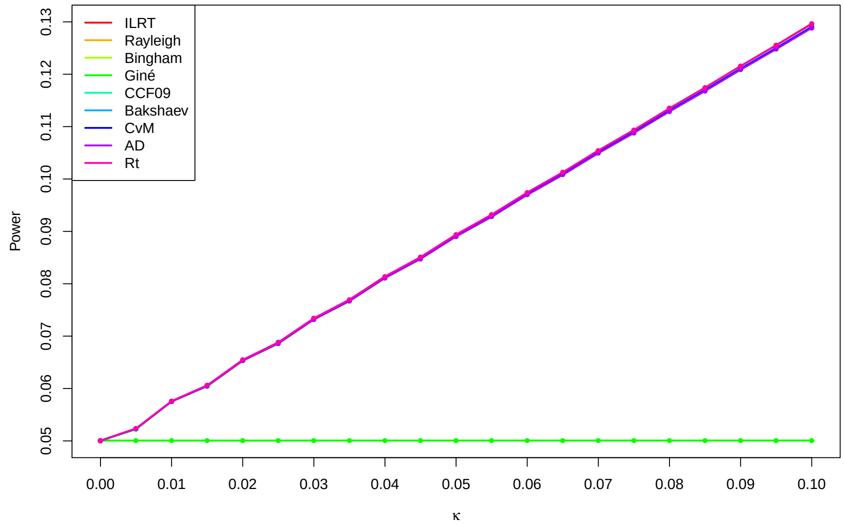


**Figure A.2.:** Empirical powers for different deviations $\kappa$.

Table A.14.: Empirical powers for the uniformity tests on $\Omega^1$. The description of Table 5.4 applies.

| DGP | $n$ | $\kappa$ | Rayleigh | Bingham | Ajne | Giné | CCF09 | Bakshaev | CvM | AD | Rt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CvM | 50 | 0.25 | 0.0746 | 0.0557 | 0.0749 | 0.0559 | 0.0735 | 0.0755 | 0.0757 | **0.0760** | 0.0755 |
| | | 0.50 | 0.1570 | 0.0743 | 0.1580 | 0.0749 | 0.1512 | 0.1619 | 0.1627 | **0.1635** | 0.1617 |
| | | 0.75 | 0.3077 | 0.1073 | 0.3104 | 0.1096 | 0.2949 | 0.3201 | 0.3222 | **0.3244** | 0.3197 |
| | 100 | 0.25 | 0.1006 | 0.0619 | 0.1010 | 0.0619 | 0.0980 | 0.1025 | **0.1027** | **0.1028** | 0.1024 |
| | | 0.50 | 0.2773 | 0.1004 | 0.2793 | 0.1021 | 0.2653 | 0.2879 | 0.2897 | **0.2918** | 0.2879 |
| | | 0.75 | 0.5580 | 0.1709 | 0.5628 | 0.1760 | 0.5398 | 0.5806 | 0.5845 | **0.5885** | 0.5801 |
| | 200 | 0.25 | 0.1571 | 0.0744 | 0.1582 | 0.0751 | 0.1509 | 0.1614 | 0.1620 | **0.1628** | 0.1613 |
| | | 0.50 | 0.5063 | 0.1566 | 0.5113 | 0.1609 | 0.4886 | 0.5284 | 0.5321 | **0.5365** | 0.5282 |
| | | 0.75 | 0.8599 | 0.3066 | 0.8650 | 0.3194 | 0.8484 | 0.8802 | 0.8833 | **0.8866** | 0.8799 |
| AD | 50 | 0.25 | 0.1897 | 0.0933 | 0.1924 | 0.0960 | 0.1884 | 0.1998 | 0.2020 | **0.2053** | 0.1995 |
| | | 0.50 | 0.6134 | 0.2429 | 0.6235 | 0.2578 | 0.6117 | 0.6480 | 0.6545 | **0.6635** | 0.6468 |
| | | 0.75 | 0.9377 | 0.4936 | 0.9424 | 0.5255 | 0.9351 | 0.9521 | 0.9542 | **0.9568** | 0.9514 |
| | 100 | 0.25 | 0.3419 | 0.1409 | 0.3478 | 0.1471 | 0.3399 | 0.3647 | 0.3696 | **0.3765** | 0.3642 |
| | | 0.50 | 0.9002 | 0.4463 | 0.9071 | 0.4776 | 0.9019 | 0.9234 | 0.9271 | **0.9319** | 0.9225 |
| | | 0.75 | 0.9989 | 0.8023 | 0.9992 | 0.8361 | 0.9989 | 0.9994 | 0.9995 | **0.9996** | 0.9994 |
| | 200 | 0.25 | 0.6113 | 0.2438 | 0.6227 | 0.2586 | 0.6144 | 0.6509 | 0.6588 | **0.6698** | 0.6501 |
| | | 0.50 | 0.9966 | 0.7517 | 0.9973 | 0.7914 | 0.9972 | 0.9985 | 0.9986 | **0.9989** | 0.9984 |
| | | 0.75 | **1.0000** | 0.9818 | **1.0000** | 0.9894 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| Rt | 50 | 0.25 | 0.0867 | 0.0588 | 0.0867 | 0.0590 | 0.0845 | **0.0875** | **0.0875** | **0.0876** | **0.0875** |
| | | 0.50 | 0.2136 | 0.0870 | 0.2136 | 0.0880 | 0.2099 | 0.2179 | 0.2187 | **0.2190** | **0.2191** |
| | | 0.75 | 0.4514 | 0.1383 | 0.4511 | 0.1411 | 0.4701 | 0.4707 | 0.4742 | **0.4780** | 0.4756 |
| | 100 | 0.25 | 0.1271 | 0.0678 | 0.1268 | 0.0682 | 0.1233 | **0.1286** | **0.1287** | **0.1285** | **0.1287** |
| | | 0.50 | 0.4020 | 0.1282 | 0.4014 | 0.1303 | 0.4124 | 0.4175 | 0.4203 | **0.4225** | 0.4213 |
| | | 0.75 | 0.7776 | 0.2380 | 0.7775 | 0.2453 | **0.8305** | 0.8125 | 0.8186 | 0.8250 | 0.8199 |
| | 200 | 0.25 | 0.2154 | 0.0871 | 0.2152 | 0.0881 | 0.2122 | 0.2205 | 0.2210 | 0.2213 | **0.2221** |
| | | 0.50 | 0.7091 | 0.2158 | 0.7091 | 0.2217 | **0.7506** | 0.7402 | 0.7455 | **0.7510** | 0.7470 |
| | | 0.75 | 0.9795 | 0.4381 | 0.9798 | 0.4540 | **0.9938** | 0.9894 | 0.9905 | 0.9917 | 0.9907 |
| vMF | 50 | 0.25 | **0.1816** | 0.0504 | 0.1814 | 0.0506 | 0.1642 | 0.1804 | 0.1792 | 0.1770 | 0.1795 |
| | | 0.50 | **0.5842** | 0.0562 | 0.5830 | 0.0562 | 0.5301 | 0.5797 | 0.5767 | 0.5700 | 0.5778 |
| | | 0.75 | **0.9112** | 0.0816 | 0.9105 | 0.0813 | 0.8747 | 0.9085 | 0.9067 | 0.9026 | 0.9076 |
| | 100 | 0.25 | **0.3302** | 0.0511 | 0.3291 | 0.0511 | 0.2946 | 0.3272 | 0.3251 | 0.3203 | 0.3257 |
| | | 0.50 | **0.8867** | 0.0634 | 0.8859 | 0.0634 | 0.8451 | 0.8837 | 0.8816 | 0.8769 | 0.8823 |
| | | 0.75 | **0.9979** | 0.1164 | 0.9978 | 0.1155 | 0.9951 | 0.9977 | 0.9976 | 0.9973 | 0.9976 |
| | 200 | 0.25 | **0.5991** | 0.0521 | 0.5979 | 0.0522 | 0.5433 | 0.5943 | 0.5910 | 0.5843 | 0.5927 |
| | | 0.50 | **0.9956** | 0.0782 | **0.9956** | 0.0778 | 0.9912 | 0.9954 | 0.9952 | 0.9949 | 0.9953 |
| | | 0.75 | **1.0000** | 0.1925 | **1.0000** | 0.1909 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| SC | 50 | 0.25 | 0.3017 | 0.0969 | 0.3014 | 0.0964 | 0.2891 | 0.3080 | 0.3088 | 0.3085 | **0.3092** |
| | | 0.50 | 0.7906 | 0.1780 | 0.7906 | 0.1765 | 0.7987 | 0.8170 | 0.8211 | **0.8237** | 0.8216 |
| | | 0.75 | 0.9738 | 0.2405 | 0.9741 | 0.2380 | 0.9823 | 0.9849 | 0.9861 | **0.9869** | 0.9862 |
| | 100 | 0.25 | 0.5643 | 0.1498 | 0.5634 | 0.1486 | 0.5584 | 0.5844 | 0.5875 | **0.5887** | **0.5883** |
| | | 0.50 | 0.9843 | 0.3243 | 0.9841 | 0.3212 | 0.9891 | 0.9910 | 0.9918 | **0.9922** | 0.9919 |
| | | 0.75 | 0.9999 | 0.4487 | 0.9999 | 0.4446 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| | 200 | 0.25 | 0.8767 | 0.2640 | 0.8760 | 0.2615 | 0.8845 | 0.8982 | 0.9013 | **0.9033** | 0.9022 |
| | | 0.50 | 1.0000 | 0.5913 | 1.0000 | 0.5868 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| | | 0.75 | **1.0000** | 0.7605 | **1.0000** | 0.7561 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| W | 50 | 0.25 | 0.0514 | **0.2580** | 0.0518 | 0.2559 | 0.0888 | 0.0683 | 0.0746 | 0.0854 | 0.0742 |
| | | 0.50 | 0.0540 | **0.7809** | 0.0551 | 0.7763 | 0.2482 | 0.1438 | 0.1906 | 0.2662 | 0.1906 |
| | | 0.75 | 0.0575 | **0.9854** | 0.0594 | 0.9846 | 0.5442 | 0.3535 | 0.4981 | 0.6568 | 0.5024 |
| | 100 | 0.25 | 0.0513 | **0.4808** | 0.0516 | 0.4767 | 0.1394 | 0.0901 | 0.1075 | 0.1374 | 0.1073 |
| | | 0.50 | 0.0538 | **0.9785** | 0.0547 | 0.9773 | 0.5358 | 0.3560 | 0.4916 | 0.6396 | 0.4946 |
| | | 0.75 | 0.0573 | **1.0000** | 0.0590 | **1.0000** | 0.9283 | 0.8734 | 0.9493 | 0.9826 | 0.9510 |
| | 200 | 0.25 | 0.0512 | **0.7959** | 0.0514 | 0.7918 | 0.2689 | 0.1576 | 0.2122 | 0.2948 | 0.2131 |
| | | 0.50 | 0.0535 | **0.9999** | 0.0545 | **0.9999** | 0.9146 | 0.8528 | 0.9350 | 0.9751 | 0.9370 |
| | | 0.75 | 0.0569 | **1.0000** | 0.0589 | **1.0000** | 0.9999 | 0.9999 | 1.0000 | **1.0000** | 1.0000 |

**Table A.15.:** Empirical powers for the uniformity tests on $\Omega^2$. The description of Table 5.4 applies.

| DGP | $n$ | $\kappa$ | Rayleigh | Bingham | Ajne | Giné | CCF09 | Bakshaev | CvM | AD | Rt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CvM | 50 | 0.25 | 0.0700 | 0.0537 | 0.0700 | 0.0539 | 0.0672 | **0.0704** | **0.0704** | **0.0705** | **0.0705** |
| | | 0.50 | 0.1360 | 0.0645 | 0.1362 | 0.0649 | 0.1233 | 0.1383 | 0.1383 | **0.1390** | 0.1379 |
| | | 0.75 | 0.2622 | 0.0840 | 0.2630 | 0.0851 | 0.2303 | 0.2686 | 0.2686 | **0.2700** | 0.2673 |
| | 100 | 0.25 | 0.0905 | 0.0567 | 0.0905 | 0.0568 | 0.0844 | **0.0914** | **0.0914** | **0.0916** | 0.0913 |
| | | 0.50 | 0.2367 | 0.0794 | 0.2377 | 0.0801 | 0.2083 | 0.2424 | 0.2424 | **0.2434** | 0.2414 |
| | | 0.75 | 0.4918 | 0.1216 | 0.4941 | 0.1237 | 0.4330 | 0.5043 | 0.5043 | **0.5068** | 0.5022 |
| | 200 | 0.25 | 0.1359 | 0.0638 | 0.1362 | 0.0640 | 0.1230 | 0.1377 | 0.1377 | **0.1380** | 0.1374 |
| | | 0.50 | 0.4411 | 0.1118 | 0.4432 | 0.1137 | 0.3870 | 0.4528 | 0.4528 | **0.4555** | 0.4500 |
| | | 0.75 | 0.8096 | 0.2070 | 0.8124 | 0.2124 | 0.7499 | 0.8231 | 0.8231 | **0.8259** | 0.8200 |
| AD | 50 | 0.25 | 0.1607 | 0.0759 | 0.1615 | 0.0766 | 0.1462 | 0.1656 | 0.1656 | **0.1674** | 0.1645 |
| | | 0.50 | 0.5367 | 0.1691 | 0.5408 | 0.1735 | 0.4824 | 0.5555 | 0.5555 | **0.5607** | 0.5515 |
| | | 0.75 | 0.8969 | 0.3460 | 0.8993 | 0.3568 | 0.8543 | 0.9075 | 0.9075 | **0.9096** | 0.9054 |
| | 100 | 0.25 | 0.2887 | 0.1038 | 0.2910 | 0.1056 | 0.2560 | 0.3004 | 0.3004 | **0.3038** | 0.2979 |
| | | 0.50 | 0.8507 | 0.3110 | 0.8542 | 0.3213 | 0.8037 | 0.8670 | 0.8670 | **0.8710** | 0.8637 |
| | | 0.75 | 0.9969 | 0.6342 | 0.9971 | 0.6510 | 0.9931 | 0.9977 | 0.9977 | **0.9978** | 0.9975 |
| | 200 | 0.25 | 0.5361 | 0.1669 | 0.5403 | 0.1715 | 0.4824 | 0.5578 | 0.5578 | **0.5640** | 0.5525 |
| | | 0.50 | 0.9919 | 0.5815 | 0.9924 | 0.5995 | 0.9853 | 0.9941 | 0.9941 | **0.9946** | 0.9937 |
| | | 0.75 | **1.0000** | 0.9225 | **1.0000** | 0.9320 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| Rt | 50 | 0.25 | **0.0785** | 0.0540 | **0.0785** | 0.0541 | 0.0749 | **0.0787** | **0.0787** | **0.0786** | **0.0787** |
| | | 0.50 | 0.1778 | 0.0657 | 0.1777 | 0.0664 | 0.1636 | **0.1788** | **0.1788** | **0.1787** | **0.1790** |
| | | 0.75 | 0.3765 | 0.0865 | 0.3763 | 0.0883 | 0.3543 | 0.3825 | 0.3825 | **0.3831** | 0.3821 |
| | 100 | 0.25 | 0.1095 | 0.0575 | 0.1093 | 0.0578 | 0.1013 | **0.1100** | **0.1100** | 0.1098 | **0.1101** |
| | | 0.50 | 0.3362 | 0.0820 | 0.3360 | 0.0833 | 0.3127 | **0.3413** | **0.3413** | **0.3414** | **0.3413** |
| | | 0.75 | 0.7019 | 0.1292 | 0.7019 | 0.1336 | 0.6965 | 0.7192 | 0.7192 | **0.7232** | 0.7174 |
| | 200 | 0.25 | 0.1798 | 0.0651 | 0.1796 | 0.0655 | 0.1646 | **0.1809** | **0.1809** | **0.1808** | **0.1807** |
| | | 0.50 | 0.6288 | 0.1186 | 0.6286 | 0.1221 | 0.6158 | 0.6428 | 0.6428 | **0.6456** | 0.6411 |
| | | 0.75 | 0.9612 | 0.2282 | 0.9612 | 0.2402 | 0.9696 | 0.9701 | 0.9701 | **0.9722** | 0.9691 |
| vMF | 50 | 0.25 | **0.1180** | 0.0506 | 0.1176 | 0.0506 | 0.1065 | 0.1172 | 0.1172 | 0.1167 | 0.1177 |
| | | 0.50 | **0.3622** | 0.0533 | 0.3614 | 0.0532 | 0.3124 | 0.3595 | 0.3595 | 0.3563 | 0.3610 |
| | | 0.75 | **0.7091** | 0.0645 | 0.7080 | 0.0642 | 0.6346 | 0.7048 | 0.7048 | 0.7003 | 0.7066 |
| | 100 | 0.25 | **0.1980** | 0.0507 | 0.1977 | 0.0504 | 0.1725 | 0.1966 | 0.1966 | 0.1950 | 0.1973 |
| | | 0.50 | **0.6648** | 0.0558 | 0.6638 | 0.0556 | 0.5895 | 0.6606 | 0.6606 | 0.6560 | 0.6628 |
| | | 0.75 | **0.9602** | 0.0809 | 0.9597 | 0.0805 | 0.9286 | 0.9585 | 0.9585 | 0.9568 | 0.9593 |
| | 200 | 0.25 | **0.3698** | 0.0508 | 0.3690 | 0.0509 | 0.3174 | 0.3662 | 0.3662 | 0.3629 | 0.3675 |
| | | 0.50 | **0.9396** | 0.0620 | 0.9391 | 0.0617 | 0.8998 | 0.9374 | 0.9374 | 0.9353 | 0.9383 |
| | | 0.75 | **0.9998** | 0.1140 | 0.9998 | 0.1132 | 0.9990 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |
| SC | 50 | 0.25 | 0.1797 | 0.0672 | 0.1793 | 0.0669 | 0.1624 | 0.1805 | 0.1805 | 0.1801 | **0.1808** |
| | | 0.50 | 0.5316 | 0.0942 | 0.5313 | 0.0939 | 0.4932 | 0.5402 | 0.5402 | **0.5412** | 0.5388 |
| | | 0.75 | 0.8378 | 0.1135 | 0.8380 | 0.1127 | 0.8187 | 0.8521 | 0.8521 | **0.8548** | 0.8495 |
| | 100 | 0.25 | 0.3410 | 0.0857 | 0.3403 | 0.0855 | 0.3077 | **0.3458** | **0.3458** | **0.3460** | **0.3456** |
| | | 0.50 | 0.8716 | 0.1489 | 0.8711 | 0.1480 | 0.8545 | 0.8858 | 0.8858 | **0.8887** | 0.8836 |
| | | 0.75 | 0.9942 | 0.1951 | 0.9941 | 0.1937 | 0.9943 | 0.9964 | 0.9964 | **0.9968** | 0.9961 |
| | 200 | 0.25 | 0.6389 | 0.1293 | 0.6378 | 0.1285 | 0.6015 | 0.6529 | 0.6529 | **0.6557** | 0.6504 |
| | | 0.50 | 0.9962 | 0.2788 | 0.9961 | 0.2765 | 0.9962 | 0.9977 | 0.9977 | **0.9980** | 0.9975 |
| | | 0.75 | **1.0000** | 0.3846 | **1.0000** | 0.3813 | **1.0000** | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| W | 50 | 0.25 | 0.0514 | **0.1540** | 0.0514 | 0.1528 | 0.0597 | 0.0591 | 0.0591 | 0.0632 | 0.0570 |
| | | 0.50 | 0.0539 | **0.5682** | 0.0545 | 0.5640 | 0.1081 | 0.0947 | 0.0947 | 0.1231 | 0.0824 |
| | | 0.75 | 0.0576 | **0.9315** | 0.0592 | 0.9294 | 0.2457 | 0.1940 | 0.1940 | 0.3124 | 0.1466 |
| | 100 | 0.25 | 0.0513 | **0.2807** | 0.0514 | 0.2782 | 0.0705 | 0.0677 | 0.0677 | 0.0774 | 0.0634 |
| | | 0.50 | 0.0536 | **0.8850** | 0.0543 | 0.8823 | 0.2031 | 0.1643 | 0.1643 | 0.2570 | 0.1265 |
| | | 0.75 | 0.0573 | **0.9991** | 0.0588 | 0.9990 | 0.5830 | 0.5265 | 0.5265 | 0.7639 | 0.3614 |
| | 200 | 0.25 | 0.0509 | **0.5377** | 0.0510 | 0.5343 | 0.0988 | 0.0884 | 0.0884 | 0.1138 | 0.0772 |
| | | 0.50 | 0.0531 | **0.9966** | 0.0536 | 0.9964 | 0.4843 | 0.4183 | 0.4183 | 0.6500 | 0.2816 |
| | | 0.75 | 0.0567 | **1.0000** | 0.0581 | **1.0000** | 0.9695 | 0.9741 | 0.9741 | 0.9973 | 0.8987 |

Table A.16.: Empirical powers for the uniformity tests on $\Omega^3$. The description of Table 5.4 applies.

| DGP | $n$ | $\kappa$ | Rayleigh | Bingham | Ajne | Giné | CCF09 | Bakshaev | CvM | AD | Rt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CvM | 50 | 0.25 | 0.0663 | 0.0522 | **0.0665** | 0.0522 | 0.0648 | **0.0668** | 0.0667 | **0.0667** | 0.0665 |
| | | 0.50 | 0.1214 | 0.0589 | 0.1218 | 0.0589 | 0.1126 | **0.1230** | **0.1230** | **0.1231** | 0.1224 |
| | | 0.75 | 0.2288 | 0.0711 | 0.2295 | 0.0718 | 0.2039 | 0.2327 | 0.2325 | **0.2330** | 0.2314 |
| | 100 | 0.25 | 0.0835 | 0.0546 | 0.0834 | 0.0547 | 0.0796 | **0.0839** | **0.0839** | **0.0839** | **0.0837** |
| | | 0.50 | 0.2064 | 0.0692 | 0.2068 | 0.0696 | 0.1839 | 0.2098 | 0.2095 | **0.2104** | 0.2087 |
| | | 0.75 | 0.4378 | 0.0965 | 0.4387 | 0.0975 | 0.3835 | 0.4449 | 0.4445 | **0.4460** | 0.4428 |
| | 200 | 0.25 | 0.1198 | 0.0586 | 0.1198 | 0.0588 | 0.1110 | **0.1209** | **0.1209** | **0.1211** | 0.1205 |
| | | 0.50 | 0.3895 | 0.0899 | 0.3903 | 0.0907 | 0.3414 | 0.3965 | 0.3961 | **0.3976** | 0.3942 |
| | | 0.75 | 0.7583 | 0.1524 | 0.7599 | 0.1549 | 0.6906 | 0.7676 | 0.7671 | **0.7688** | 0.7649 |
| AD | 50 | 0.25 | 0.1434 | 0.0675 | 0.1442 | 0.0679 | 0.1317 | 0.1466 | 0.1464 | **0.1471** | 0.1455 |
| | | 0.50 | 0.4837 | 0.1333 | 0.4862 | 0.1353 | 0.4298 | 0.4962 | 0.4952 | **0.4984** | 0.4922 |
| | | 0.75 | 0.8620 | 0.2639 | 0.8636 | 0.2688 | 0.8070 | 0.8698 | 0.8694 | **0.8707** | 0.8676 |
| | 100 | 0.25 | 0.2547 | 0.0874 | 0.2557 | 0.0883 | 0.2268 | 0.2619 | 0.2613 | **0.2637** | 0.2594 |
| | | 0.50 | 0.8092 | 0.2378 | 0.8113 | 0.2426 | 0.7497 | 0.8218 | 0.8209 | **0.8241** | 0.8180 |
| | | 0.75 | 0.9942 | 0.5102 | 0.9943 | 0.5193 | 0.9866 | 0.9950 | 0.9950 | **0.9951** | 0.9948 |
| | 200 | 0.25 | 0.4832 | 0.1316 | 0.4852 | 0.1333 | 0.4290 | 0.4973 | 0.4964 | **0.5004** | 0.4927 |
| | | 0.50 | 0.9860 | 0.4602 | 0.9865 | 0.4698 | 0.9733 | 0.9884 | 0.9883 | **0.9888** | 0.9877 |
| | | 0.75 | **1.0000** | 0.8404 | **1.0000** | 0.8483 | 1.0000 | **1.0000** | **1.0000** | **1.0000** | **1.0000** |
| Rt | 50 | 0.25 | 0.0732 | 0.0523 | **0.0734** | 0.0525 | 0.0713 | **0.0735** | **0.0735** | 0.0732 | **0.0734** |
| | | 0.50 | 0.1548 | 0.0586 | **0.1551** | 0.0589 | 0.1464 | **0.1552** | **0.1553** | 0.1547 | **0.1552** |
| | | 0.75 | 0.3247 | 0.0702 | 0.3252 | 0.0709 | 0.3094 | **0.3272** | **0.3272** | 0.3265 | 0.3269 |
| | 100 | 0.25 | **0.0988** | 0.0544 | **0.0987** | 0.0548 | 0.0946 | **0.0988** | **0.0989** | **0.0987** | **0.0988** |
| | | 0.50 | 0.2902 | 0.0686 | 0.2903 | 0.0694 | 0.2756 | **0.2924** | **0.2924** | **0.2923** | **0.2922** |
| | | 0.75 | 0.6357 | 0.0951 | 0.6360 | 0.0973 | 0.6285 | 0.6453 | 0.6446 | **0.6467** | 0.6431 |
| | 200 | 0.25 | 0.1558 | 0.0586 | 0.1556 | 0.0589 | 0.1472 | 0.1562 | **0.1564** | 0.1560 | 0.1562 |
| | | 0.50 | 0.5639 | 0.0893 | 0.5639 | 0.0911 | 0.5528 | 0.5718 | 0.5714 | **0.5727** | 0.5699 |
| | | 0.75 | 0.9377 | 0.1508 | 0.9380 | 0.1562 | 0.9450 | 0.9458 | 0.9453 | **0.9473** | 0.9437 |
| vMF | 50 | 0.25 | **0.0918** | 0.0497 | **0.0920** | 0.0498 | 0.0868 | **0.0918** | **0.0919** | 0.0913 | **0.0919** |
| | | 0.50 | **0.2477** | 0.0513 | **0.2474** | 0.0513 | 0.2208 | 0.2461 | 0.2464 | 0.2446 | 0.2469 |
| | | 0.75 | **0.5260** | 0.0570 | 0.5255 | 0.0568 | 0.4667 | 0.5224 | 0.5231 | 0.5193 | 0.5244 |
| | 100 | 0.25 | **0.1406** | 0.0507 | 0.1403 | 0.0507 | 0.1287 | 0.1399 | 0.1400 | 0.1393 | 0.1403 |
| | | 0.50 | **0.4806** | 0.0530 | 0.4798 | 0.0532 | 0.4247 | 0.4774 | 0.4779 | 0.4749 | 0.4791 |
| | | 0.75 | **0.8577** | 0.0659 | 0.8570 | 0.0659 | 0.8012 | 0.8548 | 0.8553 | 0.8525 | 0.8565 |
| | 200 | 0.25 | **0.2513** | 0.0498 | 0.2507 | 0.0498 | 0.2242 | 0.2492 | 0.2497 | 0.2476 | 0.2503 |
| | | 0.50 | **0.8142** | 0.0558 | 0.8133 | 0.0559 | 0.7535 | 0.8109 | 0.8115 | 0.8083 | 0.8127 |
| | | 0.75 | **0.9944** | 0.0817 | **0.9944** | 0.0815 | 0.9868 | 0.9941 | 0.9942 | 0.9939 | 0.9943 |
| SC | 50 | 0.25 | **0.1305** | 0.0578 | **0.1305** | 0.0578 | 0.1223 | **0.1305** | **0.1306** | 0.1301 | **0.1305** |
| | | 0.50 | 0.3750 | 0.0695 | 0.3753 | 0.0694 | 0.3463 | 0.3766 | **0.3768** | 0.3755 | 0.3764 |
| | | 0.75 | 0.6760 | 0.0769 | 0.6765 | 0.0766 | 0.6424 | **0.6818** | 0.6816 | 0.6811 | 0.6804 |
| | 100 | 0.25 | 0.2337 | 0.0670 | 0.2334 | 0.0670 | 0.2138 | **0.2350** | **0.2350** | 0.2347 | 0.2347 |
| | | 0.50 | 0.7111 | 0.0933 | 0.7104 | 0.0933 | 0.6767 | 0.7192 | 0.7187 | **0.7202** | 0.7170 |
| | | 0.75 | 0.9612 | 0.1109 | 0.9612 | 0.1106 | 0.9532 | 0.9663 | 0.9660 | **0.9671** | 0.9648 |
| | 200 | 0.25 | 0.4579 | 0.0860 | 0.4569 | 0.0858 | 0.4238 | 0.4634 | 0.4632 | **0.4637** | 0.4617 |
| | | 0.50 | 0.9685 | 0.1503 | 0.9683 | 0.1496 | 0.9615 | 0.9732 | 0.9729 | **0.9740** | 0.9718 |
| | | 0.75 | 0.9999 | 0.1942 | 0.9999 | 0.1929 | 0.9999 | 0.9999 | 0.9999 | **0.9999** | 0.9999 |
| W | 50 | 0.25 | 0.0510 | **0.1019** | 0.0511 | 0.1015 | 0.0546 | 0.0547 | 0.0543 | 0.0561 | 0.0532 |
| | | 0.50 | 0.0529 | **0.3584** | 0.0533 | 0.3555 | 0.0739 | 0.0726 | 0.0701 | 0.0811 | 0.0640 |
| | | 0.75 | 0.0560 | **0.7738** | 0.0570 | 0.7706 | 0.1322 | 0.1170 | 0.1083 | 0.1508 | 0.0884 |
| | 100 | 0.25 | 0.0508 | **0.1677** | 0.0508 | 0.1668 | 0.0582 | 0.0584 | 0.0575 | 0.0615 | 0.0553 |
| | | 0.50 | 0.0526 | **0.6728** | 0.0529 | 0.6698 | 0.1065 | 0.0980 | 0.0916 | 0.1216 | 0.0769 |
| | | 0.75 | 0.0559 | **0.9814** | 0.0567 | 0.9808 | 0.2860 | 0.2339 | 0.2027 | 0.3514 | 0.1374 |
| | 200 | 0.25 | 0.0502 | **0.3189** | 0.0501 | 0.3167 | 0.0674 | 0.0665 | 0.0645 | 0.0735 | 0.0595 |
| | | 0.50 | 0.0517 | **0.9511** | 0.0520 | 0.9497 | 0.2119 | 0.1747 | 0.1547 | 0.2540 | 0.1106 |
| | | 0.75 | 0.0548 | **1.0000** | 0.0557 | **1.0000** | 0.7051 | 0.6403 | 0.5543 | 0.8363 | 0.3176 |

Table A.17.: Empirical powers for the uniformity tests on $\Omega^{10}$. The description of Table 5.4 applies.

| DGP | $n$ | $\kappa$ | Rayleigh | Bingham | Ajne | Giné | CCF09 | Bakshaev | CvM | AD | Rt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CvM | 50 | 0.25 | 0.0585 | 0.0506 | **0.0586** | 0.0507 | 0.0557 | **0.0585** | **0.0585** | **0.0585** | **0.0585** |
| | | 0.50 | 0.0867 | 0.0524 | **0.0868** | 0.0524 | 0.0751 | **0.0870** | 0.0869 | **0.0870** | **0.0869** |
| | | 0.75 | 0.1453 | 0.0554 | 0.1456 | 0.0554 | 0.1124 | **0.1462** | 0.1460 | **0.1462** | 0.1458 |
| | 100 | 0.25 | **0.0668** | 0.0517 | **0.0669** | 0.0518 | 0.0619 | **0.0670** | 0.0669 | **0.0671** | **0.0669** |
| | | 0.50 | 0.1326 | 0.0554 | **0.1327** | 0.0554 | 0.1049 | **0.1328** | **0.1328** | **0.1328** | **0.1328** |
| | | 0.75 | 0.2735 | 0.0620 | 0.2736 | 0.0621 | 0.1978 | **0.2745** | 0.2743 | **0.2746** | 0.2742 |
| | 200 | 0.25 | 0.0874 | 0.0528 | 0.0874 | 0.0527 | 0.0757 | **0.0877** | **0.0877** | **0.0877** | 0.0876 |
| | | 0.50 | 0.2444 | 0.0604 | 0.2446 | 0.0605 | 0.1777 | 0.2458 | **0.2459** | **0.2460** | 0.2453 |
| | | 0.75 | 0.5525 | 0.0751 | 0.5528 | 0.0752 | 0.4012 | 0.5548 | 0.5546 | **0.5550** | 0.5541 |
| AD | 50 | 0.25 | 0.0999 | 0.0550 | 0.1001 | 0.0549 | 0.0836 | 0.1005 | 0.1003 | **0.1008** | 0.1002 |
| | | 0.50 | 0.3142 | 0.0735 | 0.3146 | 0.0734 | 0.2273 | 0.3170 | 0.3164 | **0.3174** | 0.3157 |
| | | 0.75 | 0.6854 | 0.1106 | 0.6858 | 0.1108 | 0.5219 | 0.6878 | 0.6875 | **0.6880** | 0.6870 |
| | 100 | 0.25 | 0.1636 | 0.0611 | 0.1638 | 0.0611 | 0.1251 | 0.1649 | 0.1646 | **0.1650** | 0.1644 |
| | | 0.50 | 0.6201 | 0.1021 | 0.6204 | 0.1023 | 0.4625 | 0.6241 | 0.6234 | **0.6246** | 0.6224 |
| | | 0.75 | 0.9624 | 0.1935 | 0.9624 | 0.1943 | 0.8790 | **0.9630** | 0.9629 | **0.9630** | 0.9628 |
| | 200 | 0.25 | 0.3166 | 0.0725 | 0.3168 | 0.0727 | 0.2272 | 0.3199 | 0.3194 | **0.3204** | 0.3185 |
| | | 0.50 | 0.9341 | 0.1716 | 0.9342 | 0.1724 | 0.8257 | 0.9361 | 0.9358 | **0.9363** | 0.9353 |
| | | 0.75 | **0.9999** | 0.3941 | 0.9999 | 0.3953 | 0.9979 | **0.9999** | **0.9999** | **0.9999** | **0.9999** |
| Rt | 50 | 0.25 | **0.0619** | 0.0504 | **0.0620** | 0.0505 | 0.0583 | **0.0619** | **0.0618** | **0.0620** | 0.0619 |
| | | 0.50 | **0.1026** | 0.0522 | **0.1027** | 0.0520 | 0.0884 | **0.1025** | 0.1024 | **0.1025** | **0.1026** |
| | | 0.75 | **0.1915** | 0.0539 | **0.1916** | 0.0540 | 0.1533 | **0.1914** | **0.1915** | **0.1913** | **0.1915** |
| | 100 | 0.25 | **0.0745** | 0.0515 | **0.0744** | 0.0517 | 0.0680 | **0.0745** | **0.0745** | **0.0745** | **0.0745** |
| | | 0.50 | **0.1744** | 0.0543 | **0.1745** | 0.0544 | 0.1410 | 0.1742 | 0.1742 | 0.1740 | **0.1744** |
| | | 0.75 | 0.4020 | 0.0594 | 0.4021 | 0.0596 | 0.3161 | 0.4023 | **0.4025** | 0.4021 | **0.4026** |
| | 200 | 0.25 | 0.1053 | 0.0524 | **0.1054** | 0.0524 | 0.0904 | 0.1053 | **0.1055** | 0.1053 | **0.1054** |
| | | 0.50 | 0.3553 | 0.0584 | 0.3554 | 0.0587 | 0.2797 | 0.3557 | **0.3560** | 0.3555 | **0.3559** |
| | | 0.75 | 0.7788 | 0.0698 | 0.7791 | 0.0699 | 0.6686 | **0.7810** | **0.7809** | **0.7810** | 0.7803 |
| vMF | 50 | 0.25 | **0.0579** | 0.0498 | **0.0580** | 0.0497 | 0.0555 | **0.0580** | **0.0579** | **0.0580** | **0.0579** |
| | | 0.50 | **0.0840** | 0.0498 | **0.0842** | 0.0497 | 0.0739 | 0.0838 | 0.0839 | 0.0838 | **0.0841** |
| | | 0.75 | **0.1372** | 0.0501 | **0.1373** | 0.0501 | 0.1099 | 0.1370 | 0.1370 | 0.1370 | 0.1371 |
| | 100 | 0.25 | **0.0657** | 0.0505 | **0.0658** | 0.0505 | 0.0610 | **0.0657** | **0.0657** | **0.0656** | **0.0657** |
| | | 0.50 | **0.1265** | 0.0503 | **0.1264** | 0.0503 | 0.1031 | 0.1261 | 0.1262 | 0.1259 | **0.1264** |
| | | 0.75 | **0.2585** | 0.0509 | 0.2583 | 0.0510 | 0.1927 | 0.2573 | 0.2576 | 0.2569 | 0.2580 |
| | 200 | 0.25 | **0.0847** | 0.0500 | **0.0846** | 0.0500 | 0.0747 | **0.0846** | **0.0847** | **0.0846** | **0.0847** |
| | | 0.50 | **0.2315** | 0.0506 | **0.2314** | 0.0507 | 0.1749 | 0.2309 | **0.2313** | 0.2307 | **0.2314** |
| | | 0.75 | **0.5305** | 0.0519 | 0.5302 | 0.0519 | 0.3967 | 0.5287 | 0.5296 | 0.5282 | 0.5301 |
| SC | 50 | 0.25 | 0.0655 | 0.0508 | **0.0657** | 0.0508 | 0.0609 | **0.0657** | **0.0656** | **0.0657** | **0.0656** |
| | | 0.50 | 0.1161 | 0.0516 | **0.1163** | 0.0515 | 0.0968 | **0.1160** | **0.1161** | 0.1160 | **0.1162** |
| | | 0.75 | 0.2096 | 0.0519 | **0.2098** | 0.0519 | 0.1624 | 0.2088 | 0.2090 | 0.2085 | 0.2094 |
| | 100 | 0.25 | **0.0846** | 0.0511 | **0.0846** | 0.0511 | 0.0744 | **0.0845** | **0.0845** | **0.0845** | **0.0846** |
| | | 0.50 | **0.2092** | 0.0526 | **0.2092** | 0.0525 | 0.1612 | 0.2086 | 0.2088 | 0.2083 | **0.2090** |
| | | 0.75 | **0.4415** | 0.0533 | **0.4413** | 0.0532 | 0.3315 | 0.4400 | 0.4405 | 0.4394 | 0.4411 |
| | 200 | 0.25 | **0.1283** | 0.0527 | **0.1283** | 0.0526 | 0.1050 | 0.1282 | **0.1283** | **0.1282** | **0.1284** |
| | | 0.50 | **0.4360** | 0.0557 | 0.4359 | 0.0558 | 0.3280 | 0.4354 | **0.4360** | 0.4351 | **0.4361** |
| | | 0.75 | 0.8191 | 0.0571 | 0.8190 | 0.0571 | 0.6824 | 0.8191 | **0.8196** | 0.8188 | **0.8196** |
| W | 50 | 0.25 | 0.0499 | **0.0533** | 0.0500 | **0.0533** | 0.0499 | 0.0502 | 0.0500 | 0.0503 | 0.0500 |
| | | 0.50 | 0.0501 | **0.0655** | 0.0502 | 0.0653 | 0.0504 | 0.0511 | 0.0508 | 0.0514 | 0.0505 |
| | | 0.75 | 0.0506 | **0.0964** | 0.0507 | 0.0962 | 0.0517 | 0.0531 | 0.0523 | 0.0536 | 0.0517 |
| | 100 | 0.25 | 0.0502 | **0.0561** | 0.0502 | **0.0560** | 0.0507 | 0.0506 | 0.0505 | 0.0507 | 0.0504 |
| | | 0.50 | 0.0504 | **0.0832** | 0.0505 | **0.0831** | 0.0515 | 0.0523 | 0.0518 | 0.0526 | 0.0512 |
| | | 0.75 | 0.0510 | **0.1581** | 0.0511 | 0.1578 | 0.0536 | 0.0559 | 0.0545 | 0.0569 | 0.0531 |
| | 200 | 0.25 | 0.0504 | **0.0633** | 0.0505 | **0.0632** | 0.0508 | 0.0513 | 0.0511 | 0.0514 | 0.0509 |
| | | 0.50 | 0.0507 | **0.1263** | 0.0507 | 0.1259 | 0.0526 | 0.0544 | 0.0535 | 0.0552 | 0.0524 |
| | | 0.75 | 0.0512 | **0.3150** | 0.0512 | 0.3144 | 0.0572 | 0.0615 | 0.0587 | 0.0637 | 0.0555 |

$$= n\mathrm{E}_{\gamma}\left(\int_{-1}^{1}\{F_{n.\gamma}(x) - F_{d-1}(x)\}^{2}$$

$$V(F_{d-1}(x))\Bigg)$$

$$\lim_{\varepsilon\to 0}\left\{\int_{-1+\varepsilon}^{1-\varepsilon}\frac{F_{d-1}(x)^{2}\mathrm{d}F_{d-1}(x)}{F_{d-1}(x)(1-F_{d-1}(x))} + \mathrm{lo}\right.$$

$$P_{n,d-1}^{W} := n\mathrm{E}_{\gamma}\left(\int_{-1}^{1}\{F_{n,\gamma}(x) - F_{d-1}(x)\}^{2}\,\mathrm{d}W(\right.$$

$$\log\left(\frac{F_{d-1}(t)}{1-F_{d-1}(t)}\right)\left(1 - F_{d-2}\left(\frac{t\tan(\theta/}{(1-t^{2})}\right.\right.$$

$$P_{n.d-1}^{W} = \frac{1}{n}\sum_{i\neq j}\psi_{d-1}^{W}(\theta_{ij}) + \int_{-1}^{1}F_{d-1}(x)(1$$

$$\psi_{d-1}^{\mathrm{AD}}(\theta) = -\log(4) + 4\int_{0}^{\cos(\theta/2)}\log\left(\right.$$

$$\Bigg)\left(1 - F_{d-2}\left(\frac{t\tan(\theta/2)}{(1-t^{2})^{1/2}}\right)\right)\mathrm{d}F_{d-}$$