Universidad de Valladolid

Escuela de Doctorado Universidad de Valladolid

PROGRAMA DE DOCTORADO EN MATEMÁTICAS

TESIS DOCTORAL:

TRIMMING METHODS FOR MODEL VALIDATION AND SUPERVISED CLASSIFICATION IN THE PRESENCE OF CONTAMINATION

Presentada por Marina Agulló Antolín para optar al grado de
Doctora por la Universidad de Valladolid

Dirigida por:
Dr. Eustasio del Barrio Tellado

Agradecimientos

En primer lugar, quiero agradecer a Tasio el haberme dado la oportunidad de realizar este trabajo con él, todo el tiempo y el trabajo que me has dedicado y lo mucho que me has enseñado a lo largo de estos años. También quiero agradecer a Carlos todo lo que me enseñaste en los primeros pasos de esta tesis. No quiero olvidarme de Jean-Michel, gracias por la acogida que me diste cuando llegué a Toulouse, por todo el tiempo que me has dedicado y por el interés que siempre has mostrado en mi trabajo. Quiero agradecer a todo el departamento de Estadística e Investigación Operativa lo bien que me han tratado a lo largo de estos años, en especial quiero agradecer a Jesús Saez toda la ayuda que me has dado.

También quiero agradecer a todos y cada uno de los maestros y profesores que he tenido a lo largo de mi vida, porque todos han puesto su granito de arena en mi formación. En especial a mi profesor de matemáticas del instituto que fue el que me metió dentro el gusanillo de las matemáticas y siempre insistió en que debía olvidarme de la psicología y estudiar matemáticas.

Quiero agradecer también a todos mis compañeros de viaje, todas esas doctorandas y doctorandos que habéis compartido estos años conmigo. No os nombro a todos porque sois muchos, pero vosotros sabéis quienes sois, los matemáticos (de la UVa y de fuera) y también los físicos. Muchas gracias por vuestra amistad, por vuestro apoyo en los momentos difíciles y por compartir mi alegría en los buenos. Aunque algunos haya sido en la distancia siempre habéis estado ahí. Voy a echar mucho de menos nuestras comidas y nuestras patatas de los jueves. Gracias a todas las personas que habéis hecho que adaptarme a una ciudad nueva no haya sido tan difícil, en especial gracias a Carmen y familia por haber tenido siempre un hueco para mi en vuestra casa, gracias a mis tios y primos y gracias a las dos alicantinas con las que he tenido la suerte de coincidir aquí. Gracias a todos mis amigos de Elche (y alrededores) que siempre habéis estado a mi lado a pesar de la distancia, por haber tenido siempre un rato para mi cuando iba de visita y por haber estado siempre al teléfono cuando lo he necesitado. Gracias a Tamara y a Patri por esa amistad eterna, no recuerdo mi vida sin vosotras a mi lado y espero no tener que hacerlo nunca.

Por último y más importante muchas gracias a mi familia, a mis padres Paco y Ventura y a mi hermano Paco. Muchas gracias por vuestro apoyo incondicional, por siempre confiar en mi y en mis posibilidades. Muchas gracias por todo lo que me habéis enseñado y por todo el amor que me habéis dado. Se que no lo digo mucho, pero os quiero mucho a los tres.

Contents

Chapter 1

Introduction

1.1 Introducción en Español

La debilidad de muchos procedimientos estadísticos clásicos en presencia de valores atípicos es un problema que ha preocupado a los estadísticos durante años. Lo ideal sería que la distribución de un estimador cambie sólo ligeramente si la distribución de las observaciones se modifica ligeramente. Este principio subyace en la llamada *Estadística Robusta* desde el trabajo pionero de Huber (1964). En el enfoque de Huber, los procedimientos estadísticos robustos son aquellos que se desempeñan relativamente bien, incluso cuando las hipótesis sólo se cumplen aproximadamente. Como una medida de la calidad de un estimador desde el punto de vista de la robustez, Hampel (1971) introdujo el punto de ruptura, basado en un concepto similar definido en Hodges (1967). A grandes rasgos, el punto de ruptura se definió como la distancia máxima de Prokhorov (véase Hampel (1971)) desde el modelo paramétrico para el que el estimador da todavía alguna indicación de la distribución original. Años más tarde, Donoho and Huber (1983) introdujo una versión más simple del punto de ruptura pensado para muestras finitas. Su versión es más parecida a la idea original de Hodges que a la definición de Hampel. Ellos consideran el punto de ruptura como la fracción más pequeña de contaminación que puede hacer que el estimador tome valores arbitrariamente grandes. Basado en esta definición Rousseeuw publicó varios trabajos sobre estimadores con un alto punto de ruptura como Rousseeuw (1985), Rousseeuw (1997) y, más recientemente, Rousseeuw and Hubert (2013). Otros autores también han trabajado sobre este tema, pongamos como ejemplo Yohai (1987) o Alfons et al. (2013).

Los métodos más robustos tienen un punto de rotura de 0,5 porque si la contaminación es superior al 50% es imposible distinguir entre la distribución contaminada y la subya-

cente. Un ejemplo muy conocido y sencillo es la mediana: para medir la tendencia central se suele utilizar la media, pero se sabe que no es un método robusto e incluso un outlier puede estropearla (tiene un punto de ruptura de $1/n$ cuando el tamaño de la muestra es n) por lo que una alternativa más robusta es la mediana cuyo punto de ruptura es de 0.5. Para una visión general sobre los métodos robustos en estadística nos referimos a Andrews et al. (1972), Hampel et al. (1986), Huber (1996) o más recientemente Maronna et al. (2006).

Se han propuesto muchas formas de robustificar estimadores como, en el caso de los estimadores de localización, cambiar la media por la mediana o, en el caso de regresión por mínimos cuadrados, cambiar la pérdida cuadrática por algo con mejores propiedades como, por ejemplo, la pérdida ℓ_1 (pero hay que tener en cuenta que esto no genera ninguna ganancia, al menos en el punto de ruptura). Entre todos estos, estamos interesados en los procedimientos de recorte.

Los procedimientos de recorte se han utilizado en estadística robusta durante muchos años, véase, por ejemplo, Bickel and Lehmann (1975). Un estimador recortado fue considerado en primer lugar como un estimador que se derivó de otro estimador excluyendo algunas de las observaciones extremas. Por ejemplo, un estimador recortado del $k\%$ fue el estimador obtenido eliminando las $k\%$ primeras y las $k\%$ últimas observaciones. Volviendo al ejemplo de tendencia central, la mediana es un estimador recortado del 50% de la media. El problema con este procedimiento surge cuando se trató de generalizarlo a variables aleatorias n-dimensionales debido a la ausencia de direcciones preferenciales para eliminar datos. Además, la forma de seleccionar la proporción de los datos a eliminar es arbitraria. Para hacer frente a estos dos problemas Rousseeuw (1984) introdujo el recorte imparcial, es decir, un procedimiento de recorte en el que es la muestra misma la que nos dice cuál es la mejor manera de recortar. Más tarde, Gordaliza (1991) introdujo el concepto de función de recorte en lugar de los conjuntos de recorte utilizados anteriormente. Se propusieron otros métodos de recorte en Cuesta-Albertos et al. (1997) o en García-Escudero et al. (2003).

Las técnicas de recorte pueden aplicarse a muchos problemas estadísticos diferentes. Entre otros, hay trabajos en análisis de datos funcionales, ver por ejemplo Fraiman and Muniz (2001) y Cuesta-Albertos and Fraiman (2006), en comparación de distribuciones, ver Álvarez-Esteban et al. (2008) y Álvarez-Esteban et al. (2012). Recortar también es muy popular como una herramienta robusta para problemas de reconocimiento de patrones. Como ejemplo de esto ya hemos mencionado su uso en regresión, donde se propuso por primera vez el procedimiento de recorte, incluyendo Rousseeuw and Driessen

(2006) y Alfons et al. (2013). Otro ejemplo del problema del reconocimiento de patrones es la clasificación. En la clasificación no supervisada podemos citar García-Escudero et al. (1999), Cuesta-Albertos et al. (2002) y García-Escudero et al. (2008). También hay cierta literatura sobre métodos de recorte en la implementación de la clasificación supervisada, véase, por ejemplo, Debruyne (2009).

En esta tesis exploramos el uso de los métodos de recorte en dos problemas estadísticos diferentes: la validación del modelos y el aprendizaje supervisado. En estas dos configuraciones propondremos y analizaremos nuevos procedimientos que se basan en el uso de recortes. Observamos en este punto que los nuevos métodos no sólo comparten un uso coincidente del recorte. De hecho, el recorte es la base de lo que podríamos llamar 'validación esencial de modelos' o 'clasificación esencial' lo que significa que estamos cambiando nuestro paradigma a través del uso de recortes y estamos tratando con nuevas versiones de la validación de modelos o del problema de clasificación. Intentaremos determinar si el generador aleatorio subyacente a una muestra puede ser asumido como una versión ligeramente contaminada de un modelo dado o identificar clasificadores simples que funcionan bien en una gran fracción de las instancias. Todo esto se hará con un uso sistemático de métodos de recorte y conceptos relacionados.

Hablando en términos generales, la clasificación supervisada es el problema de encontrar una forma automática de determinar a qué clase pertenece una observación basada en varias observaciones anteriores. Los primeros trabajos sobre este problema son Fisher (1936) y Fisher (1938) en el contexto de problemas binarios. Este trabajo condujo al análisis discriminante lineal de Fisher. Este trabajo inicial asumió que los valores dentro de cada grupo tenían una distribución normal multivariante. Posteriormente, Rao (1952) y Anderson (1958) extendieron el análisis discriminante, aún bajo suposiciones de normalidad, a más de dos grupos. Durante las últimas décadas, el problema de la clasificación binaria ha sido ampliamente estudiado y existe una gran variedad de métodos para encontrar clasificadores óptimos en diferentes entornos. Nos referimos, por ejemplo, a Devroye et al. (1996), Cristianini and Shawe-Taylor (2000) o Hastie et al. (2009) y las referencias en ellos para una visión general. Un gran avance en este campo vino con el enfoque de Vapnik basado en el principio de la minimización del riesgo empírico (ver Vapnik (1982), recientemente reimpreso en Vapnik (2006)). Este nuevo enfoque no dependía de un modelo paramétrico limitado y permitía tratar con clases de clasificadores más flexibles. Asimismo, las desigualdades de concentración recientemente desarrolladas permitieron obtener cotas precisas sobre el error de generalización de las reglas de clasificación. Pronto se reconoció que una flexibilidad excesiva en el tipo de clasificadores bajo

consideración tenía un precio.

En un escenario ideal podríamos seleccionar entre todas las funciones posibles la que mejor clasifique los datos, pero en la práctica esto es imposible en la mayoría (si no en todos) de los casos. Este problema se puede resolver mediante técnicas de selección de modelos. Brevemente, esto equivale a elegir, dados los datos, el mejor modelo estadístico entre una lista de candidatos. Aquí, por mejor modelo nos referimos a uno que equilibra el ajuste a los datos y la complejidad del modelo, logrando un buen equilibrio sesgo-varianza. Entre todo el trabajo que se ha hecho en este campo, nos gustaría citar a Massart (2007) o a el más reciente Bühlmann and van de Geer (2011) para dos estudios exhaustivos con diferentes puntos de vista.

Cuando el número de observaciones se hace grande o en un caso de alta dimensión, algunos de los datos pueden contener errores de observación y pueden considerarse datos contaminantes. La presencia de tales observaciones, si no se eliminan, dificulta la eficacia de los clasificadores, ya que muchos métodos de clasificación son muy sensibles a los valores atípicos. En realidad, si el conjunto de aprendizaje está demasiado corrompido, entrenar a un clasificador sobre este conjunto lleva a malas tasas de clasificación. Por lo tanto, existe una creciente necesidad de métodos robustos para abordar esta cuestión. Una solución para hacer frente a este problema es clasificar sólo una fracción de los datos. Este es el problema que vamos a tratar en este trabajo.

De la mano con la selección de modelos viene la validación de modelos. La validación de modelos podría definirse como una comparación entre las predicciones del modelo y el mundo real para evaluar si el modelo es apropiado para los datos con los que estamos trabajando. Entonces, deberíamos observar que los métodos clásicos de bondad de ajuste generalmente no permiten concluir que el modelo es una representación válida de los datos. En el mejor de los casos, cuando no se rechaza el modelo, la conclusión débil obtenida es que no hay suficiente evidencia contra el modelo. Peor aún, es un hecho que desde hace mucho tiempo (ver Berkson (1938)) que una prueba formal de bondad de ajuste aplicada a una muestra suficientemente grande rechazará el modelo incluso en situaciones en las que el modelo parezca una buena descripción de los datos. Estas consideraciones están detrás de la distinción entre 'significación estadística' y 'significación práctica' en Hodges and Lehmann (1954) que aboga por relajar la hipótesis nula, sustituyéndola por una desviación controlada del modelo inicial. Esta desviación se puede medir de muchas maneras diferentes. En el caso de los modelos paramétricos, la distancia euclídea habitual puede ser una buena opción, como en Hodges and Lehmann (1954). Una visión general de las diferentes opciones para esta relajación se puede encontrar en Lindsay and Liu

(2009). Más allá de la elección de una medida de desviación, la relajación implica fijar un umbral de desviación admisible del modelo. Hay un cierto grado de arbitrariedad en la fijación de este umbral y parece aconsejable elegir una medida de desviación para la que la interpretación del umbral sea lo más simple posible. Una posibilidad simple es el nivel de contaminación que debemos admitir para considerar el modelo como una buena representación de los datos. Esto se consideró en Rudas et al. (1994) (véase también Lindsay and Liu (2009)) en un marco de datos multinomiales. En este trabajo pretendemos extender el uso de esta medida de desviación a modelos continuos. Además, abordamos el problema de la validación de modelos desde un punto de vista particular. Se acepta que el modelo propuesto no es el modelo 'verdadero', es decir, que las observaciones disponibles no proceden exactamente de un generador dentro del del modelo. Reformulamos nuestra meta y nos proponemos evaluar cuánta contaminación debemos admitir para admitir que un determinado modelo es válido. Como característica distintiva adicional, nuestra propuesta adopta el punto de vista de la selección de modelos. Consideramos los entornos de contaminación de un modelo. Estos entornos crecen con el nivel de contaminación y eventualmente incluirán el generador aleatorio de los datos. Intentaremos encontrar un equilibrio adecuado entre una medida de la desviación entre la medida empírica y un entorno de contaminación y el nivel de contaminación.

En nuestra reformulación del problema de la bondad de ajuste a un marco de selección de modelos, se necesita una medida de desviación entre la medida empírica y otras medidas de probabilidad. Nuestra elección es la métrica del coste de transporte.

Ya en el siglo 18, Monge (1781) formuló el Problema del Transporte de Masas (MTP, por sus siglas en inglés) como el problema de transportar una cierta cantidad de tierra a algunos lugares donde debería ser usada para la construcción minimizando el costo de transporte. Consideró que el coste de transporte de una unidad de suelo es la distancia Euclidea entre el punto donde se extrae y el lugar donde se va a enviar. Posteriormente, Kantorovich (1960) desarrolló herramientas de programación lineal y concibió una distancia entre dos distribuciones de probabilidad, en Kantorovich (1958), que era el costo de transporte óptimo de una distribución a la otra cuando el costo era elegido como la función de distancia. Esta distancia se conoce ahora como distancia de Kantorovich-Rubinstein o como distancia de Wasserstein (esta es la denominación que usaremos a partir de ahora). En 1975, Kantorovich ganó el Premio Nobel de economía con Koopmans.

La teoría de Monge-Kantorovich se ha aplicado a diferentes disciplinas científicas como la economía, la biología o la física y a varias ramas matemáticas como la geometría, ecuaciones diferenciales, sistemas dinámicos y matemáticas aplicadas como teoría de juegos

o registro de imágenes. Como un repaso a los avances en este problema citamos Rachev and Rüschendorf (1998) y, el más reciente, Villani (2008).

La estructura especial y las propiedades del problema de transporte han permitido desarrollar un algoritmo que, partiendo de los métodos simples habituales, resuelve el problema de una manera más eficiente. La forma estándar real del problema y una primera solución constructiva fue propuesta primero en Hitchcock (1941) y, no mucho más tarde, en Koopmans (1949) pero el método simplex para el problema de transporte tal y como lo estudiamos hoy en día fue propuesto en Dantzig (1951) y Dantzig (1963).

En este trabajo hemos tratado una variante del problema de transporte óptimo, el problema de transporte parcial. Esta variante incluye la posibilidad de cierta holgura tanto en la oferta como en la demanda, es decir, la cantidad de masa en los nodos de oferta supera la cantidad que debe ser servida y la demanda no tiene que ser satisfecha completamente, sino sólo una fracción de la misma. El problema del transporte parcial surge de forma natural cuando consideramos la métrica de los costos de transporte entre conjuntos de recortes, que son, como veremos, objetos duales a los entornos de contaminación de la estadística robusta.

El objetivo de este trabajo era desarrollar métodos estadísticos en la clasificación y validación de modelos bajo el tema común de la contaminación, tal como acabamos de exponer. Hemos propuesto métodos y analizado aspectos teóricos y prácticos. Estos dos objetivos han hecho necesario utilizar una gran variedad de herramientas de diferentes campos matemáticos, así como de estadística y computación. Entre otros, queremos destacar las desigualdades de concentración, las desigualdades oráculo, los problemas de transporte óptimo, la teoría de dualidad, la programación lineal, la optimización convexa y los algoritmos de gradiente. En aras de la legibilidad, se incluye un capítulo preliminar en el que se describen algunos de los conceptos y resultados fundamentales utilizados durante esta investigación. Además, hemos implementado en R y C algoritmos para calcular eficientemente los métodos estadísticos propuestos en esta tesis. Están disponibles bajo petición.

Dedicamos el capítulo 3 al problema del transporte parcial. Mostramos cómo esta variante del problema clásico del transporte se relaciona con los modelos de contaminación a través de la idea de recortar. Luego, con vistas a las aplicaciones estadísticas, consideramos el problema de la computación de versiones empíricas del coste parcial del transporte. Nos ocupamos de un esquema doble. En primer lugar, tratamos el caso del transporte parcial entre dos medidas empíricas. Esto resulta en un problema discreto que demostramos que puede ser reformulado como un tipo particular de problema de

transporte clásico (en el sentido de programación lineal). Discutimos métodos numéricos eficientes para la solución de este problema y proporcionamos un código C (llamable desde R) con una implementación. Una segunda opción en este capítulo es el caso del transporte parcial entre la medición empírica y un modelo continuo, que lleva a un problema semidiscreto. Para este problema proporcionamos un método de optimización estocástico basado en algoritmos de gradiente. Todos estos conceptos se aplican en este capítulo a los problemas de validación de modelos contaminados. Más allá de proporcionar métodos numéricos viables, proporcionamos desigualdades oráculo que garantizan el rendimiento de nuestro método. Para concluir el capítulo, probamos nuestro algoritmo en un estudio de simulación.

El capítulo 4 puede parecer una desviación del tema de esta tesis y en realidad lo es. Durante esta investigación resultó que las técnicas de transporte óptimas eran una herramienta útil en los problemas de registro y que los algoritmos numéricos diseñados para la versión discreta del problema de transporte parcial también eran útiles para la implementación de métodos de registro basados en el transporte óptimo. En este capítulo estudiamos un problema de alineación para distribuciones deformadas proponiendo un procedimiento para la estimación de deformaciones basado en la minimización de la distancia de Wasserstein entre distribuciones deformadas y las no deformadas. El buen funcionamiento del criterio está asegurado por sus propiedades asintóticas. Finalmente ilustramos este trabajo con algunos ejemplos y resolviendo algunos problemas simulados. El material de este capítulo se publicó en Agulló-Antolín et al. (2015).

En el capítulo 5 se aborda el problema de la clasificación parcial, que es un enfoque robusto del problema clásico de clasificación. Es práctica común en el análisis de datos eliminar los datos 'perturbadores' o 'atípicos' y aplicar los procedimientos al conjunto de datos depurado. La clasificación no es una excepción a esta regla. Pero entonces las garantías de generalización de los clasificadores entrenados de esta manera pueden ser diferentes de lo que el usuario esperaría. Con esta motivación nos proponemos buscar el clasificador que ofrezca el mejor rendimiento sobre una fracción suficientemente grande de los datos. Formulamos esto como un problema de minimización del riesgo sobre un conjunto de recortes. Dado que estos conjuntos de recortes crecen con el nivel de recorte, proponemos una versión penalizada, por la que proporcionamos garantías de su rendimiento a través de las desigualdades oráculo. Si bien la teoría que presentamos es más simple y limpia con el uso de una pérdida de 0/1, su aplicabilidad se ve obstaculizada por el hecho de que esta pérdida de 0/1 requeriría, en su aplicación práctica, la minimización de algún objetivo no convexo. Por esta razón hemos explorado la extensión de los resultados

a una función de pérdida de mejor comportamiento, la pérdida hinge usada en el SVM. Proporcionamos desigualdades oráculo y proporcionamos estrategias computacionales viables. También esbozamos algunas maneras alternativas de elegir la función de pérdida. Este capítulo incluye algunos resultados de simulación que ilustran el rendimiento de los métodos, así como el análisis de un conjunto de datos reales. Parte del material de este capítulo se presenta en Agulló-Antolín et al. (2017).

Finalizamos este documento con un breve capítulo que resume las contribuciones en este trabajo. Creemos que algunos resultados de esta tesis son contribuciones útiles hacia una metodología más completa para la validación de modelos esenciales y la clasificación esencial, que fueron los objetivos iniciales de este proyecto. Sin embargo, durante esta investigación han surgido nuevas líneas de investigación. En lugar de una tarea inacabada, nos gustaría pensar en estos problemas no resueltos como una oportunidad para el trabajo futuro.

1.2 Introduction in English

The weakness of many classical statistical procedures in the presence of atypical values is a problem that has concerned statisticians for years. Ideally, the distribution of an estimator changes only slightly if the distribution of the observations is slightly altered. This principle underlies the so called *Robust Statistics* since the pioneering work of Huber (1964). In Huber's approach robust statistical procedures are those that perform relatively well even when the assumptions are only approximately met. As a measure of the quality of an estimator from the point view of robustness, Hampel (1971) introduced the breakdown point, based on a similar concept defined in Hodges (1967). Roughly speaking, the breakdown point was defined as the maximum Prokhorov distance (see Hampel (1971)) from the parametric model for which the estimator still gives some indication of the original distribution. Years later, Donoho and Huber (1983) introduced a simpler version of the breakdown point intended for finite samples. Their version is closer to the original idea of Hodges than Hampel's definition. They consider the breakdown point as the smallest fraction of contamination that can make the estimator take arbitrarily large values. Based on this definition, Rousseeuw published several works about estimators with a high breakdown point as Rousseeuw (1985), Rousseeuw (1997) and, more recently, Rousseeuw and Hubert (2013). Other authors have worked on this topic as well, take as an example Yohai (1987) or Alfons et al. (2013).

The most robust methods have a breakdown point of 0.5 because if the contamination

is greater than 50% it is impossible to distinguish between the contaminated distribution and the underlying one. A very well-known and simple example is the median: to measure the central tendency is often used the mean, but it is known that is not a robust estimator and even one outlier can crash it (it has a breakdown point of $1/n$ when the sample size is n) so a more robust alternative is the median whose breakdown point is 0.5. For a general view on robust methods in statistics we refer to Andrews et al. (1972), Hampel et al. (1986), Huber (1996) or the more recent Maronna et al. (2006).

Many ways to robustify estimators have been proposed as, in the case of location estimators, changing the mean by the median or, in the case of least squares regression, changing the square loss by something with better properties as, for example, the ℓ_1-loss (but beware that this does not generate any gain, at least in breakdown point). Among all, we are interested in trimming procedures.

Trimming procedures have been used in robust statistics for many years, see as an example Bickel and Lehmann (1975). A trimmed estimator was firstly considered as an estimator that was derived from another one by excluding some of the extreme observations. For example an k%-trimmed estimator was the estimator obtained by removing the k% first and the k% last observations. Going back to the central tendency example, the median is a 50% trimmed estimator of the mean. The problem with this procedure arises when one tries to generalize it to n dimensional random variables due to the absence of preferential directions to remove data. In addition, the way of selecting the proportion of the data to be removed is arbitrary. To deal with this two problems Rousseeuw (1984) introduced the impartial trimming, that is, a trimming procedure in which is it the sample itself that tells us what is the best way to trim. Later, Gordaliza (1991) introduced the concept of trimming function instead of the trimming sets that were used before. Further trimming methods were proposed in Cuesta-Albertos et al. (1997) or in García-Escudero et al. (2003).

Trimming techniques can be applied to many different statistical problems. Among others, there are works in functional data analysis, see for example Fraiman and Muniz (2001) and Cuesta-Albertos and Fraiman (2006), in comparison of distributions, see Álvarez-Esteban et al. (2008) and Álvarez-Esteban et al. (2012). Trimming is also very popular as a robust tool for pattern recognition problems. As an example of this we have already mentioned its use in regression, where the trimming procedure was first proposed, including Rousseeuw and Driessen (2006) and Alfons et al. (2013). Another example of pattern recognition problem is classification. In unsupervised classification we can cite García-Escudero et al. (1999), Cuesta-Albertos et al. (2002) and García-Escudero et al.

(2008). There is also some literature on trimming methods in the setup of supervised classification, see, e.g. Debruyne (2009).

In this thesis we explore the use of trimming methods in two different statistical problems: model validation and supervised learning. In these two setups we will propose and analyze new procedures that rely on the use of trimming. We note at this point that the new methods do not share only a coincidental use of trimming. In fact, trimming is the basis for what we could call *essential model validation* or *essential classification*, meaning that we are changing our paradigm through the use of trimming and are dealing with new versions of the model validation or the classification problem. We will try to determine whether the random generator underlying a sample can be assumed to be a slightly contaminated version of a given model or to identify simple classifiers that perform well over a large fraction of the instances. All this will be done with a systematic use of trimming methods and related concepts.

Roughly speaking, supervised classification is the problem of finding an automatic way to determine to which class does an observation belong based on several previous observations. Early work on this problem is Fisher (1936) and Fisher (1938) in the context of binary problems. This work led to Fisher's linear discriminant analysis. This early work assumed that the values within each group had a multivariate normal distribution. Later Rao (1952) and Anderson (1958) extended discriminant analysis, still under normality assumptions, to more than two groups. During the last decades, the binary classification problem has been extensively studied and there exists a large variety of methods to find optimal classifiers in different settings. We refer for instance to Devroye et al. (1996), Cristianini and Shawe-Taylor (2000) or Hastie et al. (2009) and references therein for a survey. A major breakthrough in the field came with Vapnik's approach based on the principle of empirical risk minimization (see Vapnik (1982), recently reprinted in Vapnik (2006)). This new approach did not depend on a limited parametric modeling and allowed to deal with more flexible classes of classifiers. Also, the newly developed concentration inequalities enabled to obtain precise bounds about the generalization error of the classification rules. Soon it was recognized that an excessive flexibility in the kind of classifiers under consideration came at a price.

In an ideal setting we would be able to select among all the possible functions the one that better classifies the data, but in practice this is impossible in most (if not all) of the cases. This problem can be handled through model selection techniques. In short, this amounts to choosing, given the data, the best statistical model among a list of candidates. Here, by best model we refer to one that balances fit to the data and complexity of the

model, attaining a good bias-variance trade-off. Among all the work that has been done in this field we would like to cite Massart (2007) or the more recent Bühlmann and van de Geer (2011) for two comprehensive surveys with different points of view.

When the number of observations grows large or in a high dimensional case, some of the data may contain observation errors and may be considered as contaminating data. The presence of such observations, if not removed, hampers the efficiency of classifiers since many classification methods are very sensitive to outliers. Actually, if the learning set is too corrupted, training a classifier over this set leads to bad classification rates. Hence there is a growing need for robust methods to tackle such issue. A solution to cope with this issue is to allow ourselves to classify only a fraction of the data. This is the problem we are going to deal with in this work.

Hand in hand with model selection comes model validation. Model validation could be defined as a comparison between the model's predictions and the real world to assess if the model is appropriate for the data we are working with. Then, we should note that classic goodness-of-fit methods generally do not allow to conclude that the model is a valid representation of the data. In the best case scenario, when the model is not rejected, the weak conclusion obtained is that there's not enough evidence against the model. Even worse, it is a long-held fact (see Berkson (1938)) that a formal goodness-of-fit test applied on a sufficiently large sample will reject the model even in situations where the model looks like a good description of the data. These considerations are behind the distinction between 'statistical significance' and 'practical significance' in Hodges and Lehmann (1954) which advocates for relaxing the null hypothesis, replacing it by a controlled deviation from the initial model. This deviation can be measured in many different ways. In the case of parametric models the usual Euclidean distance can be a good choice, as in Hodges and Lehmann (1954). A survey of different choices for this relaxation can be found in Lindsay and Liu (2009). Beyond the choice of a measure of a deviation, the relaxation involves fixing a threshold of admissible deviation from the model. There is a certain degree of arbitrariness in fixing this threshold and it seems advisable to choose a deviation measure for which the interpretation of the threshold is as simple as possible. A simple possibility is the contamination level that we should admit to consider the model as a good representation of the data. This was considered in Rudas et al. (1994) (see also Lindsay and Liu (2009)) in the setup of multinomial data. In this work we aim at extending the use of this measure of deviation to continuous models. Additionally, we address the problem of model validation from a particular point of view. It is accepted that the proposed model is not the 'true' model, i.e., that the observations available do not exactly come

from a generator within the scope of the model. We reformulate our goal and we propose to evaluate how much contamination we must accept in order to admit a given model as valid. As a further distinctive feature, our proposal adopts the point of view of model selection. We consider contamination neighbourhoods of a model. These neighbourhoods grow with the contamination level and will eventually include the random generator of the data. We will try to find a right balance between some measure of deviation between the empirical measure and a contamination neighbourhood and the contamination level.

In our reformulation of the goodness-of-fit problem to the setup of model selection, a measure of deviation between the empirical measure and other probability measures is needed. Our choice is the transportation cost metric.

Back in the 18th century Monge (1781) formulated the Mass Transportation Problem (MTP) as the problem of transporting a certain amount of soil to some places where it should be used for construction minimizing the transportation cost. He considered that the transport cost of a unit of soil is the Euclidean distance between the point where it is extracted and the place where it is going to be shipped. Later, Kantorovich (1960) developed tools for linear programming and conceived a distance between two probability distributions, in Kantorovich (1958), which was the optimal transportation cost from one distribution to the other when the cost was chosen as the distance function. This distance is now known as Kantorovich-Rubinstein distance or as Wasserstein distance (this is the denomination we will use in the following). In 1975 Kantorovich won the Nobel Prize of economics with Koopmans.

Monge-Kantorovich theory has been applied to different science disciplines as economy, biology or physics and to several mathematic branches such as geometry, nonlinear partial differential equations, dynamical systems and applied mathematics as game theory or image registration. As a review of the advances in this problem we cite Rachev and Rüschendorf (1998) and, more recently, Villani (2008).

The special structure and properties of the transportation problem has made it possible to develop an algorithm that, starting from the usual simplex methods, solves the problem in a more efficient way. The actual standard form of the problem and a first constructive solution was first proposed in Hitchcock (1941) and, not much later, in Koopmans (1949) but the simplex method for transportation problem as we study it today was proposed in Dantzig (1951) and Dantzig (1963).

In this work he have dealt with a variant of the optimal transportation problem, the partial transportation problem. This variant includes the possibility of some slackness both in supply and demand, that is, the amount of mass at supply nodes exceeds the

amount that has to be served and the demand does not have to be completely met, but only a fraction of it. The partial transportation problem arises in a natural way when we consider the transportation cost metric to sets of trimmings, which are, as we will see, dual objects to the contamination neighbourhoods of robust statistics.

The purpose of this work was to develop statistical methods in classification and model validation under the common topic of contamination, in the way we have just discussed. We have proposed methods and analyzed theoretical and practical aspects. These two objectives have made necessary to use a great variety of tools from different mathematical fields, as well as from statistics and computation. Among others we want to point out concentration inequalities, oracle inequalities, optimal transportation problems, duality theory, linear programming, convex optimization and gradient algorithms. For the sake of readability, a preliminary chapter is included in which we describe some of the concepts and fundamental results used during this research. Besides, we have implemented, in R and C, algorithms to efficiently compute the statistical methods proposed in this thesis. They are available upon request.

We devote chapter 3 to the partial transportation problem. We show how this variant of the classical transportation problem is related to contamination models through the idea of trimming. Then, with a view towards statistical applications we consider the problem of computing empirical versions of the partial transportation cost. We deal with a double setup. First, we handle the case of partial transportation between two empirical measures. This results in a discrete problem which we show that can be recast as a particular kind of classical transportation problem (in the sense of linear programming). We discuss efficient numerical methods for the solution of this problem and provide C code (callable from R) with an implementation. A second setup in this chapter is the case of partial transportation between empirical measure and a continuous model, leading to a semidiscrete problem. For this problem we provide a stochastic optimization method based on gradient algorithms. All these concepts are applied in this chapter to contaminated model validation problems. Beyond providing feasible numerical methods, we provide oracle inequalities that guarantee the performance of our method. To conclude the chapter, we tested our algorithm in a simulation study.

Chapter 4 may seem a deviation of the topic of this thesis and actually it is. It turned out during this research that optimal transportation techniques were a useful tool in registration problems and that the numerical algorithms designed for the discrete version of the partial transportation problem were also useful for the partial implementation of registration methods based on optimal transportation. In this chapter we study an align-

ment problem for warped distributions proposing a procedure for deformation estimation based on the minimization of Wasserstein's distance between warped distributions and the unwarped ones. The good performance of the criterion is assured by its asymptotic properties. Finally we illustrate this work with some examples and solving some simulated problems. The material in this chapter was published in Agulló-Antolín et al. (2015).

Chapter 5 deals with the partial classification problem which is a robust approach to the classical classification problem. It is common practice in data analysis to remove 'disturbing' or 'atypical' data and apply established procedures to the cleaned data set. Classification is not an exception to this rule. But then the generalization guarantees of the classifiers trained in this way may be different from what the user would expect. With this motivation we propose to look for the classifier that offers the best performance over a large enough fraction of the data. We formulate this as a problem of risk minimization over a set of trimmings. Since these sets of trimmings grow with the trimming level, we propose a penalized version, for which we provide performance guarantees through oracle inequalities. While the theory that we present is simpler and cleaner with the use of a 0/1 loss, its applicability is hampered by the fact that this 0/1 loss would require, in its practical application, the minimization of a non-convex target. For this reason we have explored the extension of the results to a better behaved loss function, the hinge loss from SVM. We provide oracle inequalities and provide feasible computational strategies. We also outline some alternative ways of choosing loss function. This chapter includes some simulation results illustrating the performance of the methods as well as the analysis of a real data set. Part of the material in this chapter is submitted in Agulló-Antolín et al. (2017).

We end this document with a short chapter summarizing the contributions in this work. We believe that some results in this thesis are useful contributions towards a more complete methodology for *essential model validation* and *essential classification*, that were the initial goals of this project. However, during this research new lines have arisen. Rather than an unfinished task, we would like to think of these unsolved problems as an opportunity for future work.

Chapter 2

Preliminaries

This chapter describes several concepts and results that are fundamental in the development of this thesis and that will be essential for proving some results. There in essentially no new result in this chapter, but it is included for the sake of reference and readability for the other chapters in this document. We start by introducing trimming methods that will be the principal nexus of the problems treated in this thesis. Section 2.2 deals with the classification problem; a robust approach to this problem is the main subject of this chapter. Section 2.3 addresses the optimal transportation problem and Wasserstein metrics that result from this problem. Combining trimming methods with the problem of optimal transportation presents the partial transport problem. Its resolution by means of the Wasserstein distance between trimming sets will be the first topic we will tackle in this paper. Trimming methods will give a new vision to the problems of validating contaminated models and aligning models with deformations. We will treat them in sections 2.1 and 2.4. To solve some of these problems, in addition to the theoretical approach, we will propose algorithms that solve them efficiently. These algorithms will be based on other existing algorithms whose description is covered in the last section of this chapter.

2.1 Statistical methods based on trimming

In statistical practice, it is usual to treat in a particular way observations that deviate from the main trend or that significantly affect the credibility of a particular model. One possibility is to eliminate or trim such atypical observations or outliers. More specifically, as mentioned in the introduction, impartial trimming methods were introduced in Rousseeuw (1984) as a robustifying method for statistical estimators and have received

extensive attention in the literature (see Gordaliza (1991), Cuesta-Albertos et al. (1997), Fraiman and Muniz (2001), Álvarez-Esteban et al. (2008) o Alfons et al. (2013) among others).

Shortly, trimming a sample consists in removing a fraction of the sample's points. That is, to replace the empirical measure

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$$

by a new measure

$$\frac{1}{n}\sum_{i=1}^{n}b_i\delta_{x_i} \tag{2.1}$$

where $b_i = 0$ for the observations we want to delete and $b_i = \frac{n}{n-k}$ for the rest of the observations with $k \leq n\alpha$ for the number of observations trimmed and $\alpha \in (0,1)$ the maximum proportion of points we can remove. An alternative to eliminating some observations and leaving others is to divide the weights so that we decrease the weight of the points we suspect may be outliers instead of eliminating them and increase the weight of others. In this case, for all points of the sample we will have $0 \leq b_i \leq \frac{1}{n}\frac{1}{(1-\alpha)}$ in such a way that $\sum_{i=1}^{n} b_i = 1$.

The theoretical counterpart of trimmings in a sample are the trimmed distributions. Suppose we are in the measurable space $(\mathcal{X}, \beta)$, we will call $\mathcal{P}(\mathcal{X}, \beta)$ the set of probability measures in the measurable space. The *trimming set of level α or set of α-trimmings* of a probability $P \in \mathcal{P}(\mathcal{X}, \beta)$ is defined, given $0 \leq \alpha \leq 1$, as

$$\mathcal{R}_\alpha(P) = \left\{ Q \in \mathcal{P}(\mathcal{X}, \beta) : \quad Q \ll P, \quad \frac{dQ}{dP} \leq \frac{1}{1-\alpha} \quad P\text{-a.s.} \right\}$$

Equivalently, we can say that $Q \in \mathcal{R}_\alpha(P)$ if and only if $Q \ll P$ y $\frac{dQ}{dP} = \frac{1}{1-\alpha} f$ with $0 \leq f \leq 1$. When $f = I_A$ with A a measurable set such that $P(A) = 1 - \alpha$, trimming is just a conditional probability.

Now we present some useful properties of trimming sets. This proposition combines the results of Propositions 3.5 and 3.6 in Álvarez-Esteban (2009).

Proposition 2.1. *Let $\alpha, \alpha_1, \alpha_2 \in (0,1)$ and $P \in \mathcal{P}(\mathcal{X}, \beta)$ be a probability measure, then trimming sets satisfy:*

(a) $\alpha_1 \leq \alpha_2 \Rightarrow \mathcal{R}_{\alpha_1}(P) \subset \mathcal{R}_{\alpha_2}(P)$.

(b) $\mathcal{R}_\alpha(P)$ *is a convex set.*

(c) $Q \in \mathcal{R}_\alpha(P)$ *if and only if* $Q(A) \leq \frac{1}{1-\alpha}P(A)$ *for all* $A \in \beta$.

(d) If $(\mathcal{X}, \beta)$ *is a separable metric space then* $\mathcal{R}_\alpha(P)$ *is closed for the weak convergence topology in* $\mathcal{P}(\mathcal{X}, \beta)$. *If, in addition,* $\mathcal{X}$ *is complete, then* $\mathcal{R}_\alpha(P)$ *is compact.*

Trimming sets relate to contamination models that are an essential part of the statistical robustness theory introduced by Huber (Huber (1964), Huber (1996), Huber (1981)). Briefly, robust statistics is an alternative approach to classic statistical methods that aim to obtain estimators that are not affected by small variations in model assumptions. In addition to limited sensitivity to small variations in the model, it is also desirable for a robust method that larger deviations do not completely ruin the model.

We say that $P \in \mathcal{P}(\mathcal{X}, \beta)$ is an *α-contaminated version* of Q if

$$P = (1-\alpha)Q + \alpha P'$$

with $Q, P' \in \mathcal{P}(\mathcal{X}, \beta)$. Furthermore, when we consider a whole class of probabilities $\mathcal{F}$ instead of a single probability, we say that $\mathcal{F}_\alpha$ is the *α-contaminated neighbourhood* of $\mathcal{F}$ if

$$\mathcal{F}_\alpha = \{(1-\alpha)Q + \alpha R : \ Q \in \mathcal{F} \text{ and } R \text{ a probability}\}.$$

That is, the α-contaminated neighbourhood of $\mathcal{F}$ is the set of all α-contaminated versions of the probability measures that form the class $\mathcal{F}$.

A related concept is $\alpha-$similarity (Álvarez-Esteban et al. (2012)). Two probabilities $P, Q \in \mathcal{P}(\mathcal{X}, \beta)$ are *α-similar* if both are α-contaminated versions of the same distribution $R \in \mathcal{P}(\mathcal{X}, \beta)$, that is, if

$$\begin{cases} P = (1-\alpha)R + \alpha P' \\ \\ Q = (1-\alpha)R + \alpha Q' \end{cases}$$

where $P', Q' \in \mathcal{P}(\mathcal{X}, \beta)$.

The following propositions, coming from Proposition 2 in Álvarez-Esteban et al. (2008), reflect the mentioned relationship between trimming sets and contamination models.

Proposition 2.2. *Let* $P, P', Q \in \mathcal{P}(\mathcal{X}, \beta)$ *and* $\alpha \in [0, 1)$ *then*

$$Q \in \mathcal{R}_\alpha(P) \iff P = (1-\alpha)Q + \alpha P'.$$

Proposition 2.3. *Let* $P, P', Q, Q', R \in \mathcal{P}(\mathcal{X}, \beta)$ *and* $\alpha \in [0, 1)$ *then*

$$\begin{cases} P = (1-\alpha)R + \alpha P' \\ \\ Q = (1-\alpha)R + \alpha Q' \end{cases} \iff \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q) \neq \emptyset \iff d_{TV}(P, Q) \leq \alpha,$$

where ($d_{TV}(P,Q) = \sup_{A\in\beta} |P(A) - Q(A)|$) *stands for the distance in total variation which is the largest difference between the probability that the two distributions can give to the same set.*

Contamination neighbourhoods are also an essential element in the theory of statistical robustness. Propositions 2.2 and 2.3 provide a dual formulation of contamination models that will be exploited in Chapter 3. More specifically, if we consider a metric, d, over $\mathcal{P}(\mathcal{X}, \beta)$, or over an appropriate subset so that $\mathcal{R}_\alpha(P)$ is closed for d, then another distribution $Q \in \mathcal{R}_\alpha(P)$ is a contaminated version of P if and only if $d(Q, \mathcal{R}_\alpha(P)) = 0$. Equivalently, the same characterization is given for Proposition 2.3, two distributions P and Q are contaminated versions of the same distribution if and only if $d(\mathcal{R}_\alpha(Q), \mathcal{R}_\alpha(P)) = 0$.

2.2 The classic classification problem

The classification problem is one of the classic problems in Statistics. Briefly, it consists in searching, from a data set (the *training set*) in which group belonging has been observed along with additional attributes, for rules that predict belonging to one of those groups of future cases in which only these additional attributes will have been observed. There is extensive literature on this issue. For an overview, see Hastie et al. (2009). Mention should also be made of Lugosi (2002), Boucheron et al. (2005), Massart (2007) and the third chapter of del Barrio et al. (2007), which give a complete overview of the current state of the art in classification with a focus on obtaining oracle inequalities for model selection. Finally, mention should also be made of Cristianini and Shawe-Taylor (2000) for a more applied and computational view of the problem. To make it easier to read and fix the notation, we include here a brief description of the main elements of the problem.

Usually we find samples formed by n pairs $\xi_i = (Y_i, X_i)$ where Y_i is the *label* assigned to an individual i whose *attributes* are X_i. While X_i is usually a vector in $\mathbb{R}^d$, the domain of Y_i will vary depending on the problem we are working with:

- $Y \in \{0, 1\}$ or $Y \in \{-1, 1\}$: Binary classification
- $Y \in \{1, \ldots, m\}$: Multiclass classification

The sample $S = ((Y_1, X_1), \ldots, (Y_n, X_n))$ is called the *training set*.

The goal of classification is to find a function $g : \mathbb{R}^d \mapsto \mathrm{Dom}(Y)$ that predicts from an attribute which will be its label. This function is obtained from the training sample, selecting it in such a way that it classifies correctly as many pairs of the sample as possible.

In this work we will focus on binary classification. We assume that the pairs in the sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ are i.i.d. observations with $Y_i \in \{0, 1\}$ (or $Y \in \{-1, 1\}$) and $X_i \in \mathbb{R}^d$. We are going to consider that all variables in this section are defined in the space $(\Omega, \mathcal{F}, \mathbb{P})$. In this context a *classification rule* is a function $g : \mathbb{R}^d \to \{0, 1\}$. If (Y, X) is a new observation, with the same distribution as (Y_i, X_i) then the prediction given by the rule will be $g(X)$. The rule is *correct* if $g(X) = Y$.

The law of (Y_i, X_i) is a probability in $\{0, 1\} \times \mathbb{R}^d$ which will be denoted by P. Given $A \subset \{0, 1\} \times \mathbb{R}^d$ we say $A_i = \{x \in \mathbb{R}^d : (i, x) \in A\}$, $i = 0, 1$. Obviously $A = (\{0\} \times A_0) \cup (\{1\} \times A_1)$ and the union is disjoint, so that for each $A \subset \{0, 1\} \times \mathbb{R}^d$ measurable,

$$P(A) = p_0 P_0(A_0) + p_1 P_1(A_1), \tag{2.2}$$

where $p_0 = P(\{0\} \times \mathbb{R}^d)$, $p_1 = 1 - p_0$, $P_0(A_0) = P(\{0\} \times A_0)/p_0$ and $P_1(A_1) = P(\{1\} \times A_1)/p_1$. P_0 and P_1 are probabilities in $\mathbb{R}^d$. Conversely, from $p_0 \in [0, 1]$ and probabilities P_0 and P_1 in $\mathbb{R}^d$ equation (2.2) defines a probability in $\{0, 1\} \times \mathbb{R}^d$ and the relationship is one-to-one (except for the degenerate cases $p_0 = 0$ or $p_0 = 1$), so we can identify each probability P with the object (p_0, P_0, P_1). We will keep this identification along the work.

For each sample we can find infinite classification functions and we want to choose among them the one that best classifies the sample. In order to make this choice we need a measure of the error we make. We have chosen to use the *generalization error* which is the probability of misclassification of future observations, that is,

$$R(g) := \mathbb{P}(g(X) \neq Y) = P\{(y, x) : g(x) \neq y\}.$$

Since, in the case we are considering, a classification rule is a function with values in $\{0, 1\}$ we can express it as the indicator of a set, $g = I_A$. Bad classification occurs for the point $(0, x)$ if $x \in A$ and for the point $(1, x')$ if $x' \notin A$. In terms of the notation (p_0, P_0, P_1) the classification error is then

$$P\{(y, x) : g(x) \neq y\} = p_0 P_0(A) + (1 - p_0) P_1(A^C).$$

Similarly, if μ is a measure for which P_0 and P_1 are absolutely continuous with densities f_0, f_1 respectively, then the generalization error is

$$\int_{\mathbb{R}^d} (p_0 f_0 I_A + (1 - p_0) f_1 I_{A^C}) d\mu.$$

With this last expression we see that there is a classification rule that minimizes the generalization error, this rule is obtained by taking $A = \{x \in \mathbb{R}^d : p_0 f_0(x) \leq (1 - p_0) f_1(x)\}$, which produces the generalization error

$$\mathrm{Err}(P) = \mathrm{Err}(p_0, P_0, P_1) = \int_{\mathbb{R}^d} \min(p_0 f_0, (1 - p_0) f_1) d\mu.$$

The above rule is known as *Bayes rule* and $\mathrm{Err}(P)$ as Bayes error. Its interest is theoretical because it is impossible to calculate it unless the distribution of the sample is known a priori. It is the least possible generalization error; on the other hand, it depends on p_0, f_0 and f_1, which are unknown. Given the impossibility of calculating Bayes rule, we usually restrict ourselves to looking for the best possible classification rule within a class.

The selection of a suitable class is very important when it comes to getting good classifiers. To know if a class is adequate, we need to measure the error that is made by selecting the class and to do this we are going to use the *generalization error of a class* $\mathcal{G}$ which is the minimum error that we can achieve with classifiers in the class, that is to say,

$$R(\mathcal{G}) := \min_{g \in \mathcal{G}} R(g). \tag{2.3}$$

The generalization error is defined in terms of a probability P that we will rarely know, which makes impossible to calculate $R(g)$. Instead, an estimator of this amount is used, usually an empirical version of $R(g)$. The proportion of errors made with a rule g is known as the *empirical error or empirical risk* of the rule,

$$R_n(g) := \frac{1}{n} \sum_{j=1}^{n} I_{(g(x_j) \neq y_j)}. \tag{2.4}$$

Through the following inequality, inspired by Lemma 1.1 in Lugosi (2002), we can assess how appropriate the choice of $\mathcal{G}$ is. Given a class of classifiers $\mathcal{G}$, we define

$$g_B = \arg\min_{g} R(g),\ \hat{g_0} = \arg\min_{g \in \mathcal{G}} R(g) \text{ and } \hat{g_n} = \arg\min_{g \in \mathcal{G}} R_n(g),$$

the Bayes classifier, the best classifier in the class and the classifier that provides the minimum empirical error in the class respectively.

Proposition 2.4. *Let* $\mathcal{E}(\mathcal{G}) := \min_{g \in \mathcal{G}} R(g) - R(g_B)$ *be the* excess of risk *of* $\mathcal{G}$*, then*

$$R(\hat{g}_n) - R(g_B) \leq 2 \sup_{g \in \mathcal{G}} |R_n(g) - R(g)| + \mathcal{E}(\mathcal{G}). \tag{2.5}$$

Proof.

$$\begin{aligned} R(\hat{g}_n) - R(g_B) &= R(\hat{g}_n) - R_n(\hat{g}_n) + R_n(\hat{g}_n) - R(\hat{g}_0) + R(\hat{g}_0) - R(g_B) \\ &\leq (R(\hat{g}_n) - R_n(\hat{g}_n)) + (R_n(\hat{g}_0) - R(\hat{g}_0)) + (R(\hat{g_0}) - R(g_B)) \\ &\leq 2 \sup_{g \in \mathcal{G}} |R_n(g) - R(g)| + \mathcal{E}(\mathcal{G}). \end{aligned}$$

□

As happened before with the Bayes rule, we can not calculate $\hat{g}$, and therefore $R(\hat{g})$, without a priori knowledge of the distribution. But we can estimate the value of the generalization error by means of the empirical generalization error defined in (2.4) and, therefore, approximate $\hat{g}$ by means of $\hat{g}_n := \arg\min_{g \in \mathcal{G}} R_n(g)$ which will be the rule that minimizes the empirical error within the class we are considering.

Proposition 2.4 helps us to understand what would be an appropriate choice of the class $\mathcal{G}$. The two terms on the right-hand side have a different role: the second term is a bias term and the first one is a variance term. If we choose a class that is too large, we will have a small bias but a very large variance, which translates into classifiers that are over-adjusted to the training sample and can also be computationally expensive to obtain. On the other hand, choosing a small class will reduce variance but increases bias, in this case we will get bad classifiers. In order to find a balance between bias size and variance, several techniques have been proposed, among which we will focus on the model selection techniques derived from Vapnik's method for minimizing structural risk based on concentration inequalities presented in Vapnik (1982).

Since choosing a class that is too small, even if it reduces variance, provides poor classifiers, the tendency is to choose large classes at the risk of getting over-adjusted rules that may not work well with samples different from the training set. In order to avoid selecting classes that are too large, a penalty is introduced on the size of the class to the generalization error that we want to minimize. From now on we are going to work with the following objective function

$$\min_{m \in \mathbb{N}} \left[\min_{g \in \mathcal{G}_m} R(g) + pen(\mathcal{G}_m) \right], \tag{2.6}$$

with $(\mathcal{G}_m)_{m \in \mathbb{N}}$ a family of classifiers class.

Choosing an appropriate penalty is crucial for a good class selection. A very large penalty means that very small classes are chosen which, as we have already seen, leads to bad classifiers. On the other hand, a very small penalty leads to very large classes that cause over-adjustment and that is precisely what we were trying to avoid. To ensure that the chosen penalty is appropriate, oracle inequalities are used. They guarantee that if the sample is large enough, the error of generalization we make with $\hat{g}_{\hat{m}}$ (that is the classifier in which

$$\min_{m \in \mathbb{N}} \left[\min_{g \in \mathcal{G}_m} R_n(g) + pen(\mathcal{G}_m) \right]$$

is attained) will be small.

An oracle inequality relates the performance of a real classifier to an ideal classifier that relies on perfect information provided by a superior being (or oracle) and is not available

in practice. The best model (the oracle) will be the one that minimizes (2.6), so it is said that a model selector $\hat{m}$ satisfies an oracle inequality if

$$E(R(\hat{g}_{\hat{m}})) \leq K \inf_{m \in \mathcal{M}} \left(R(\mathcal{G}_m) + pen(\mathcal{G}_m) + o(n^{-r}) \right) + o(n^{-s}),$$

for certain $r, s > 0$ and $\mathcal{M}$ a subset of $\mathbb{N}$.

As we have already said, the penalty depends on the size of the chosen class. Since a class will usually consist of infinite classifiers, the form we will use to quantify the size of a class will be its Vapnik-Chervonenkis dimension. Together with this dimension we are going to use a series of concentration inequalities when it comes to obtaining our own oracle inequalities.

2.2.1 Vapnik-Chervonenkis theory

Vapnik-Chervonenkis theory was developed during the 70's to explain the learning process from a statistical point of view. In addition to statistical learning theory, it is related to empirical processes. The basic elements of this theory are explained below, with statements drawn mainly from Devroye et al. (1996), Devroye and Lugosi (2001) and Vapnik (1999).

Given a collection of measurable sets $\mathcal{A}$, a *Vapnik-Chervonenkis shatter coefficient* is

$$S_{\mathcal{A}}(n) = \max_{x_1, \ldots, x_n \in \mathbb{R}^d} |\{\{x_1, \ldots, x_n\} \cap A; A \in \mathcal{A}\}|.$$

If $\mathcal{A}$ is a set collection with $|\mathcal{A}| \geq 2$, the *Vapnik-Chervonenkis dimension* (or VC dimension) of collection $\mathcal{A}$ is the biggest integer $k \geq 0$ for which $S_{\mathcal{A}}(k) = 2^k$. We will denote it by $V_{\mathcal{A}}$. If $S_{\mathcal{A}}(n) = 2^n \ \forall n$, we say that $V_{\mathcal{A}} = \infty$.

A relevant example is given by the collection $\mathcal{A}$ of subsets of $\mathbb{R}^d$ of the form $\{x : a^T x - b \geq 0\}$ with $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$, then $V_{\mathcal{A}} = d+1$ and $S_{\mathcal{A}}(n) = 2 \sum_{i=0}^{d} \binom{n-1}{i} \leq 2(n-1)^m + 2$ (see Corollary 13.1 in Devroye et al. (1996)).

Usually Vapnik-Chervonenkis shatter coefficients are difficult to calculate. A crucial result to ease the control of the size of $S_{\mathcal{A}}(n)$ is Sauer's Lemma, which we reproduce below. The lemma guarantees that $S_{\mathcal{A}}(n)$ grows at most polynomically if the collection $\mathcal{A}$ has finite VC dimension. A proof can be found at Devroye et al. (1996) (Theorem 13.2).

Theorem 2.5 (Sauer's lemma)**.** *Let $\mathcal{A}$ be a collection of sets with Vapnik-Chervonenkis dimension $V_{\mathcal{A}} < \infty$. Then for all n,*

$$S_{\mathcal{A}}(n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}.$$

Shatter coefficients are a useful tool to control the variance term in decomposition (2.5). As an example, by symmetrization and randomization arguments we can prove (see Theorem 3.1 in Devroye and Lugosi (2001)) that

$$E\left(\sup_{A\in\mathcal{A}}|P(A)-P_n(A)|\right)\leq 2\sqrt{\frac{\ln(2)S_{\mathcal{A}}}{n}},$$

where P_n stands for the empirical version of P, that is, $P_n=\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$, $X_1,\ldots,X_n$ are random vectors i.i.d. with values in $\mathbb{R}^d$ and $\mathcal{A}$ is a class of measurable sets in $\mathbb{R}^d$. If class $\mathcal{A}$ has finite VC dimension then the following consequence is obtained (see Devroye and Lugosi (2001), Corollary 4.1),

$$E\left(\sup_{A\in\mathcal{A}}|P(A)-P_n(A)|\right)\leq 2\sqrt{\frac{V_{\mathcal{A}}\ln(n+1)+\ln(2)}{n}}. \tag{2.7}$$

From the point of view of the 0/1 loss, a classifier g can be identified with the indicator of the set $A_g=\{(y,x):g(x)\neq y\}$, so that $R_n(g)=P_n(A_g)$. In the same way, we can identify a class $\mathcal{G}$ with the indicator of the set of points $\mathcal{A}=\{(Y,X):g(X)\neq Y;g\in\mathcal{G}\}$ and now $R_n(\mathcal{G})=\inf_{g\in\mathcal{G}}P_n(A_g)$. As a result, we can identify the VC dimension of the class $\mathcal{G}$, $V_{\mathcal{G}}$, with the VC dimension of the collection of sets $\mathcal{A}$. Therefore, the bound (2.7) is also valid for the expected value of the difference between theoretical and empirical generalization error. If $V_{\mathcal{G}}<\infty$, then

$$E\left(\sup_{g\in\mathcal{G}}|R(g)-R_n(g)|\right)\leq 2\sqrt{\frac{V_{\mathcal{G}}\ln(n+1)+\ln(2)}{n}}. \tag{2.8}$$

2.2.2 Loss functions, SVM and LASSO

For computational and other reasons it can be of interest to replace the 0/1 loss and consider different loss functions and, consequently, different risk measures. When we consider the 0/1 loss all poorly ranked points are equally penalized regardless how far or close they are from being well ranked. This loss function is a piecewise constant function and is neither convex nor continuous. This means that, in practice, the calculation of the minimum is very expensive, making it impossible to calculate the optimal rule even for small samples. The reason why we consider its study interesting is because of the good statistical properties it has. We will analyze these properties in depth in section 5.2. Due to this computational problem, other loss functions with better calculation properties such as convex functions will be considered. Throughout section 5.3 we will study this type of functions focusing mainly on the hinge function which is a convex function of the form

$l(x) = (1-x)_+$. This function not only considers whether an observation is well or poorly classified, in case of misclassification it also gives a measure of how bad this classification is. The hinge functions is well-known because it is the one used in SVM that we will see later in this section.

On the other hand, it is worth mentioning the L_1 loss function used in LASSO algorithms. This method was originally proposed in Tibshirani (1996) for regression problems but the proposed techniques have also been used to obtain optimal rules for the classification problem. Shortly, LASSO estimator is the one that, given a training sample $(y_1, x_1), \ldots, (y_n, x_n)$ with $x_1, \ldots, x_n$ independent and $(y_i, x_i) \in \mathbb{R}^{d+1}$, is defined by

$$\hat{\beta}_{LASSO} := \begin{array}{c} \arg\min_{\beta \in \mathcal{R}^d} \sum_{i=1}^{n} (Y_i - \beta^T x_i)^2 \\ \text{s.t.} \sum_{i=1}^{d} |\beta_i| \leq \lambda \end{array}$$

where $\lambda > 0$ is an adjustment parameter. The resulting optimization problem is a quadratic programming problem with linear constraints. There are many efficient and stable algorithms for resolving this type of problem, such as LARS (Efron et al. (2004)) or sparse LTS (Alfons et al. (2013)), which are discussed in more detail at the end of this chapter.

As has been observed, LASSO methods consider only linear rules. If f is a linear function of the form $f(x) = x^T\beta + \beta_0$ where $\beta \in \mathbb{R}^d$, $x \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$,

$$g(x) = \begin{cases} -1 & if \quad g(x) < 0 \\ \\ 1 & if \quad g(x) \geq 0 \end{cases}.$$

Within the linear classifiers we highlight *support vector classifier* (see, for example, Cristianini and Shawe-Taylor (2000)) which is obtained by maximizing the distance of each observation to the separation hyperplane. These classifiers are those provided by *Support Vector Machines*. SVMs are a set of supervised learning algorithms developed by Vapnik and used in classification and regression problems to obtain optimal classifiers and regressors. Intuitively, a SVM is a model that separates the points of each class into semispaces as wide as possible by means of a separation hyperplane (note that by the term wide we mean the distance between the points and the hyperplane). The optimization

problem that solves SVMs is the following

$$\begin{aligned} \text{(SVM)} \ \max_{\beta,\beta_0} \quad & C \\ \text{s.t.} \quad & y_i(x^T\beta + \beta_0) \geq C(1-\zeta_i) \ i = 1, \ldots, n \\ & \parallel \beta \parallel = 1 \\ & \zeta_i \geq 0 \\ & \sum_{i=1}^{n} \zeta_i \leq K, \end{aligned}$$

where K is a constant that limits the maximum number of classification errors and $\zeta = (\zeta_1, \ldots, \zeta_n)$ are slack variables that represent the proportional amount by which the prediction is on the wrong side of its margin (they will be worth 0 when the observation is on the correct side).

2.3 Optimal transportation problem and Wasserstein metrics

Optimal transportation problem is a widely studied mathematical problem due to its usefulness in many branches of science and social science. In this section we will focus on studying the continuous formulation of the problem that is used to calculate the cost of mass transportation between probabilities. We would like to outline Bickel and Freedman (1981), where you can find proofs of the statements reproduced in this section and the books Villani (2003) and Villani (2008) in which we can find an overview of the evolution of the problem.

Generally speaking, the transportation problem consists in transferring a mass quantity between two distributions so that the transportation cost is minimal. If P and Q are two probabilities in a separable metric space and c is a measurable function in such a way that P represents the distribution of mass at source and Q at destination and c is a cost function where $c(x, y)$ represents the cost of moving a unit of mass from the location x to the location y, the transportation problem can be formulated as

$$\inf_{\pi \in \mathcal{M}(P,Q)} \int c(x,y) d\pi(x,y),$$

where $\mathcal{M}(P, Q)$ is the set of probability measures with marginals P and Q respectively.

This problem is directly related to Wasserstein metrics that were designed to quantify the distance between probability distributions.

Although transportation problem and Wasserstein metrics can be defined in more general spaces, in this work we limit ourselves to handling the concepts in $\mathbb{R}^d$. Let $\mathcal{P}_p(\mathbb{R}^d)$ with $p \geq 1$ be the set of all (Borel) probabilities in $\mathbb{R}^d$ with finite moment of order p. We denote by $\mathcal{W}_p(P, Q)$, the Wasserstein L_p distance between two probability measures $P, Q \in \mathcal{P}_p(\mathbb{R}^d)$. This distance is

$$\mathcal{W}_p^p(P,Q) := \inf_{\pi \in \mathcal{M}(P,Q)} \left\{ \int \|x-y\|^p d\pi(x,y) \right\} \tag{2.9}$$

$$= \min\{E((X-Y)^p) : X, Y \text{ r.v. with marginals } P \text{ and } Q\} \tag{2.10}$$

where $\mathcal{M}(P, Q)$ is, as before, the set of probability measures in $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. In Bickel and Freedman (1981) is proved that $\mathcal{W}_p$ actually defines a distance over $\mathcal{P}_p(\mathbb{R}^d)$.

If $d = 1$ this distance can be calculated easily thanks to the following result that corresponds to Lemma 8.2 in Bickel and Freedman (1981).

Proposition 2.6. *If $P, Q \in \mathcal{P}_p(\mathbb{R})$, let F and G be the distribution functions of P and Q respectively and F^{-1} and G^{-1} their quantile functions, then*

$$\mathcal{W}_p^p(P,Q) = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right|^p dt.$$

If $d > 1$ there is not a equivalent expression that allows us to calculate this distance. Here arises the interest in numerical methods that allow us to approximate its value as the discretization of the problem through empirical versions of the distributions (we will study this in section 2.5.1) or the stochastic approximation in section 3.3.

The following proposition reproduces an interesting characterization of Wasserstein's distance convergence that can be found in Bickel and Freedman (1981).

Proposition 2.7. *Let $\{P_n\}_{n\in\mathbb{N}}$ be a sequence in*

$$\mathcal{W}_2(\mathbb{R}^d) := \left\{ P \in \mathbb{R}^d : \int_{\mathbb{R}^d} \|x\|^2 dP(x) < \infty \right\}$$

and $P \in \mathcal{W}_2(\mathbb{R}^d)$ then

$$\mathcal{W}_2(P_n, P) \xrightarrow{n\to\infty} 0 \text{ if and only if } \begin{cases} P_n \rightharpoonup P \\ \displaystyle\int_{\mathbb{R}^d} \|x\|^2 dP_n(x) \xrightarrow{n\to\infty} \int_{\mathbb{R}^d} \|x\|^2 dP(x) \end{cases}, \tag{2.11}$$

where $\rightharpoonup$ denotes weak convergence in the classical way.

From this characterization it is easy to prove that $\mathcal{R}_\alpha(P)$ is compact for $\mathcal{W}_p$ (see Álvarez-Esteban et al. (2011)). As a consequence, the equivalent forms of Propositions 2.2 and 2.3 will be valid.

2.4 Deformation models. Alignment

Sometimes we come across several samples that we know they come from the same distribution but have suffered transformations that make it difficult to compare them. In order to undo these transformations and be able to compare the samples, the alignment problem arises. The following example illustrates the problem. Suppose we have 5 samples all from a normal distribution. One follows a $N(0,1)$, but the others have suffered translations, change of scale or both, coming now from distributions $N(1, 0.75)$, $N(2,2)$, $N(0,5)$ and $N(5,1)$. The graphic on the left in figure 2.1 represents the empirical distribution functions of the 5 samples while the graphic on the right represents the empirical distribution functions for the standardized data. In this easy way the distributions have been aligned and made more comparable.

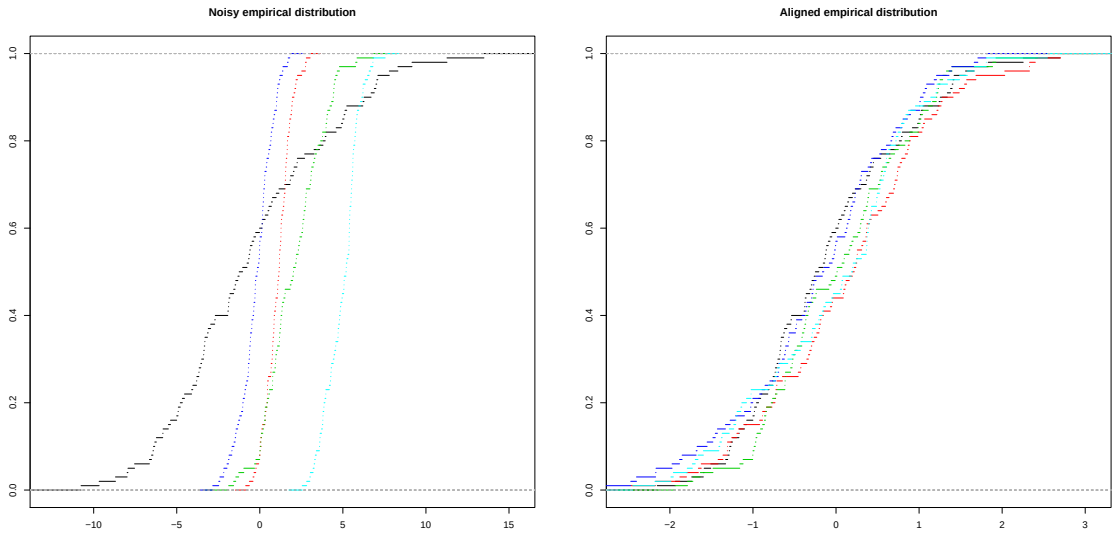


Figure 2.1: Example of normal distribution alignment

The distribution alignment problem often arises in biology, for example, when considering gene expression data obtained with microarray technologies. In this case, the alignment is known as normalization, see for example Bolstad et al. (2003) and related work Gallon et al. (2013). Here we are going to consider the extension of semi-parametric alignment methods, as in Gamboa et al. (2007) or in Vimond (2010), to the problem of estimating a distribution of random variables observed in a deformation frame.

Assume that we have n samples of J independent random variables X_{ij} with distribution μ_j, where $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Each sample is extracted from one mean distribution μ with some variations, i. e. there are some unobserved deformation functions

of φ such that, for all j, we have $\mu_j = \mu \circ \varphi^{-1}$. In order to deal with this problem we will consider that the deformations have a known shape that depends on parameters specific for each sample. Therefore, we have parameters $\theta^* = (\theta_1^*, \ldots, \theta_J^*)$ such that $\varphi_j = \varphi_{\theta_j^*}$ for every $j = 1, \ldots, J$. Each θ_j^* represents the deformation of the j-th sample that must be removed by reversing the deformation operator to recover the unknown distribution. So, the observed model is

$$X_{ij} = \varphi_{\theta_j^*}(\varepsilon_{ij}) \quad 1 \leq i \leq n, \quad 1 \leq j \leq J,$$

where ε_{ij} are independent identically distributed unobserved random variables with an unknown distribution μ. We aim to build an estimator for the parameters θ_j^*.

Alignment problems can be seen from two perspectives:

- The solution is provided by choosing an observation as a reference and aligning the rest according to this chosen pattern.
- We obtain the solution by aligning the sample with the mean of the deformed distributions.

The second option is more robust and less sensitive to such a previous choice than the first. This case has been studied for the case of regression in Vimond (2010). In chapter 4 we will generalize this work to the case of distribution deformation that, besides, will allow us to deal with problems in which we have multidimensional deformation parameters.

2.5 Algorithms

In section 2.3 we saw the relationship between continuous transportation problem and Wasserstein metrics that measure the distance between probability distributions. As we have already seen, for a dimension greater than 1 there is not an expression that allows us to calculate this distance exactly. Hence the interest in numerical methods that enable us to approximate it arises. We will study a stochastic approximation in section 3.3 and also we will calculate the distance between the empirical versions of the distributions for large enough samples. In this latter case, the empirical distribution of the first probability will play the role of origin and the second one will play the role of destination. As transportation cost we will take a power of the norm of the difference between each point of the first distribution and each point of the second. In this way, the problem of optimal transportation becomes a linear programming problem. Although the number of variables grows fast as the sample sizes increase (as a linear programming problem we will

have $n \times m$ variables with $n + m$ restrictions) there are specialized implementations of the simplex method that are able to efficiently solve the problem for relatively large values of n and m (see Bazaraa et al. (2010)). However, the statistical problems considered in this thesis lead to a different version of the classic transportation problem: the optimal partial transportation problem. This problem naturally arises when studying properties of trimming sets with Wasserstein metric and will be addressed in chapter 3. The empirical version of this partial transportation problem leads to a discrete version of the transportation problem that does not fit into the classic linear programming problem. For this reason, an adaptation of a classic algorithm capable of solving this new problem has been developed in this memory. This adaptation is described in chapter 3. Here, in section 2.5.1, we describe the classical algorithm on which we have relied in order to make the document easier to read.

The other part of this section explores another type of algorithms that could potentially lead to efficient methods for calculating optimal classifiers for certain loss functions. As we observed before, despite its good theoretical properties, the lack of algorithms to obtain the optimal classifier when the loss function is $0/1$ limits in practice its application. This brings us to investigate other loss functions that are convex and can be efficiently minimized with existing algorithms. Among these loss functions we highlight the quadratic and hinge functions which with appropriate penalties lead to convex optimization problems. Despite convexity, we often encounter problems of very high dimensions, which result in very slow optimization methods. In order to find efficient algorithms, related to the LASSO estimator, arose the coordinate descent algorithm. This algorithm is used to find the classifier (or regressor) that is used to minimize the LASSO objective function (for more details see, for example, Bühlmann and van de Geer (2011)).

Again, the aim of this work is to apply trimming techniques to find more robust solutions to problems such as classification. Therefore, instead of considering only that we want to minimize the classification error of each observation, we will have a double minimization since we will also consider that each observation will be accompanied by a weight that indicates whether we should take it into account for classification or not. Although we do not have an algorithm for calculating double minimization directly, by combining the above-mentioned coordinate descent and a concentration algorithm as proposed in Rousseeuw and Driessen (2006) we can approximate the solutions to such double minimization problems.

In order to facilitate the presentation of algorithms of this type introduced in this thesis, we include here a subsection dedicated to describing the method of coordinate

descent (2.5.2) and another dedicated to presenting concentration algorithms of the type of Rousseeuw and Driessen (2006)(2.5.3).

2.5.1 Optimal transportation algorithms

The discrete optimal transportation problem is one that minimizes the cost of transporting a certain amount of goods from n origins to m destinations. A graphical representation of this problem is shown in figure 2.2.

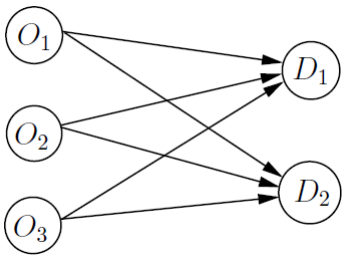


Figure 2.2: Graph of a transportation problem with 3 origin nodes and 3 destination nodes.

First of all, we are going to see the formulation of the linear optimization problem in order to fix notation and make it easier to understand chapter 3. We denote by o_i the amount offered in source node i and by d_j the amount demanded by destination node j. Let C_{ij} be the cost of transporting a unit from source i to destination j and call X to the matrix $n \times m$ in which each of its components x_{ij} corresponds to the variable that indicates the amount transported from the source node i to the destination node j, then the formulation of the optimal transport problem will be

$$\begin{aligned}
(\text{OTP}) \quad \min_{X} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}C_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{m} x_{ij} \leq o_i,\ i = 1,\dots,n && (2.12)\\
& \sum_{i=1}^{n} x_{ij} \geq d_j,\ j = 1,\dots,m && (2.13)\\
& x_{ij} \geq 0,\ i = 1,\dots,n,\ j = 1,\dots,m.
\end{aligned}$$

For this problem to have a feasible solution, it must necessarily happen that

$$\sum_{j=1}^{m} d_j \leq \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \leq \sum_{i=1}^{n} o_i,$$

that is, total offer must always be greater or equal than total demand.

Among all optimal transportation problems, those that are balanced (those with $\sum_{i=1}^{n} o_i = \sum_{j=1}^{m} d_j$) are of special importance to us. When we are in this situation, we can replace the inequalities in the restrictions (2.12) and (2.13) with equalities and the problem (OTP) will be equivalent to

$$\begin{aligned} \text{(BTP)} \ \min_X \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}C_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{m} x_{ij} = o_i, \ i = 1, \ldots, n \\ & \sum_{i=1}^{n} x_{ij} = d_j, \ j = 1, \ldots, m \\ & x_{ij} \geq 0, \ i = 1, \ldots, n, \ j = 1, \ldots, m. \end{aligned}$$

The importance of these problems is that the algorithm we are going to describe in this section can only be applied when the problem we want to solve is balanced. Occasionally we find problems that are not balanced, i.e., $\sum_{i=1}^{n} o_i > \sum_{j=1}^{m} d_j$. In these cases we can transform the unbalanced problem into an equivalent balanced problem. To do this, an artificial demand node with zero costs is added: $c_{i(m+1)} = 0$ for $i = 1; \ldots, n$ with demand $d_{m+1} = \sum_{i=1}^{n} o_i - \sum_{j=1}^{m} d_j$ and the problem can be written in a standard form.

$$\begin{aligned} \min_X \quad & \sum_{i=1}^{n}\sum_{j=1}^{m+1} x_{ij}C_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{m+1} x_{ij} = o_i, \ i = 1, \ldots, n \\ & \sum_{i=1}^{n} x_{ij} = d_j, \ j = 1, \ldots, m+1 \\ & x_{ij} \geq 0, \ i = 1, \ldots, n, \ j = 1, \ldots, m+1. \end{aligned}$$

Below we are going to see that it is possible to estimate the Wasserstein distance between two probability distributions by calculating the optimal transport plan between their empirical distributions.

If we denote by Π_{ij} the amount transported between x_i and y_j, we take

$$\begin{aligned}\sum_{j=1}^{m}\Pi_{ij} &= p_i;\ i=1,\dots,n,\\ \sum_{i=1}^{n}\Pi_{ij} &= q_j;\ j=1,\dots,m\end{aligned}$$

and assume that

$$\Pi_{ij}\geq 0;\ i=1,\dots,n;\ j=1,\dots,m,$$

then, writing the expected value in (2.10) as

$$E((X-Y)^p)=\sum_{i=1}^{n}\sum_{j=1}^{m}|x_i-y_j|^p\Pi_{ij},$$

we have

$$\begin{aligned}\mathcal{W}_p^p(P,Q)=\min_{\Pi}\quad & \sum_{i=1}^{n}\sum_{j=1}^{m}|x_i-y_j|^p\Pi_{ij}\\ \text{s.t.}\quad & \sum_{j=1}^{m}\Pi_{ij}=p_i;\ i=1,\dots,n\\ & \sum_{i=1}^{n}\Pi_{ij}=q_j;\ j=1,\dots,m\\ & \Pi_{ij}\geq 0;\ i=1,\dots,n;\ j=1,\dots,m.\end{aligned}$$

that is a problem of the same type as (BTP).

Now we describe the algorithm. We start from a balanced problem (an unbalanced problem can be easily reformulated into this case). We need a feasible initial solution to start working with. This solution is obtained using a heuristic algorithm. There are already several heuristic algorithms described in the literature and we will see below some of them. This solution is known as a feasible initial solution. In order to reach the optimum from the initial solution, we must obtain new solutions in an iterative way that bring us closer and closer to the optimum (that is to say, we will obtain solutions whose total cost is lower and lower) until we reach it. These new solutions will be obtained by the method called MODI (Bazaraa et al. (2010)). This procedure is known as the simplex method for transportation problem.

For calculating the initial basic solution there are several heuristics in the literature, the best known are the *northwest corner* denoted as NC and *Vogel approximation method* (known as VAM). We are also going to consider a modification of this last method which,

as we will see later, is the one that obtains the best resolution times, especially when the problem we want to solve is a big one.

To apply these heuristics we have to start by building what is known as a transportation tableau. To elaborate this table we have to put in the first column the origin nodes and in the first row the target nodes, each square of the table will have in the upper right corner a rectangle with the cost of going from the origin node of the row in which we are to the destination node of the corresponding column. Finally, outside the table we write the offer vector on a column to the right and the demand vector on a row below the table. An example of transportation tableau is shown in Figure 2.3.

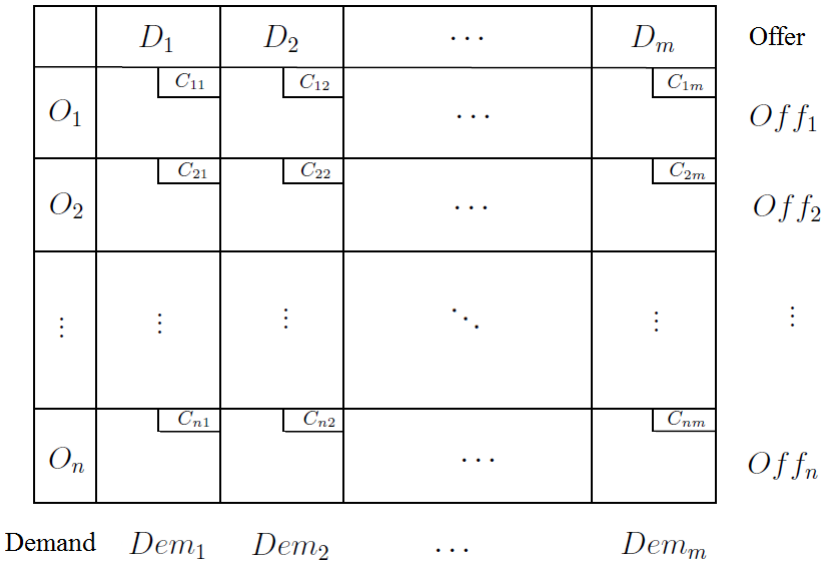


Figure 2.3: Transportation tableau.

This table will be filled according to the heuristic algorithm that we choose until we obtain the basic initial solution. The number that will appear in each cell will be the quantity of goods that we transport between the source node and the destination node of the row and the column in which the cell is.

The northwest corner heuristic (see Bazaraa et al. (2010)) is the easiest to implement and the fastest to get an initial feasible solution. The problem with this algorithm is that it only takes into account offer and demand and does not take transport costs into account. This means that, in most cases, the initial solution is very far from the optimal solution, which will require many iterations to reach the final solution. The number of iterations will increase with the size of the problem and the time required for each iteration will also increase making the use of this unfeasible to solve big problems. Algorithm 1 shows this

Algorithm 1: Northwest corner

1. Start with the cell in the upper left corner, which corresponds to the value of x_{11}, and give the cell the maximum value possible that will be the minimum between the supply on the first source node and the demand of the first destination node.

2. Subtract from the corresponding elements of the offer and demand vectors the amount already allocated, which will cause at least one of the two to become 0. If the one that is canceled is demand, fill the rest of the column with 0, if it is supply, do the same with the row. In the event that both of them are canceled, we fill the row with 0 and in the offer vector, instead of putting a 0 in the corresponding position, we will put a very small amount ε (small enough so the solution of the problem will not be affected).

3. Ignore the row or column filled with 0 and repeat steps 1 and 2 for the new table that will have one row or column less than the previous one. Repeat this until all the demand and therefore all the offer has been delivered.

method.

On the other hand, we have the heuristic VAM (see Reinfeld and Vogel (1960)) which is slower than the northwest corner because the calculations we must make to get the initial feasible solution are more complicated. The advantage of this algorithm is that it does take into account transportation costs and therefore the solution will be closer to the optimum. This reduction in the number of iterations required will be reflected in a reduction in the total resolution time. We can see a diagram of this heuristics in Algorithm 2.

Algorithm 2: Vogel's approximation method

1. Calculate the penalty for each row and column, this is done by subtracting the smallest cost from the second smallest cost of the row and the same way for the columns.
2. Select the row or column for which the penalty is the highest, in case of a tie we will keep the lowest index.
3. We assign to the empty cell that has the lowest cost of the row or column selected the largest amount of goods we can (as was done in the northwest corner this amount will be the minimum between demand and supply in the corresponding node).
4. Subtract the quantity already allocated to the corresponding elements of the vectors offer and demand. One of them will be canceled, if it is demand fill with 0 the rest of the elements of the column, if it is supply do the same with the row. If both are canceled, fill the row with 0 and assign an offer of ε.
5. Repeat the previous 4 steps until we have $m + n - 1$ boxes with positive value.

A slight modification of the VAM algorithm is described in Korukoğlu and Ballı (2011) which provides a better initial solution and is called IVAM. The advantage of using this heuristic is that the initial solution is slightly better than VAM's. This means that the

number of iterations required to reach the optimum will be smaller and, for large problems, will significantly reduce the computation time of the problem.

This improvement has two differences with the original method. The first is that instead of working with the given cost matrix, it works with what is known as the TOC matrix or total opportunity cost matrix, which is obtained by adding the matrices obtained by subtracting the lowest cost within each row (ROC) and the second by doing the same thing per column (COC). If we denote by $C_i = (C_{i1}, \ldots, C_{im})$ with $i = 1, \ldots, n$ and by $C'_j = (C_{1j}, \ldots, C_{nj})^T$ with $j = 1, \ldots, m$, clearly $C = [C_1, \ldots, C_n]^T = [C'_1, \ldots, C'_m]$. So if we define

$$\begin{aligned} \tilde{C}_i &= C_i - \min_{1 \leq j \leq m} \{C_{ij}\} \text{ with } i = 1, \ldots, n \\ \tilde{C}'_j &= C'_j - \min_{1 \leq i \leq n} \{C_{ij}\} \text{ with } j = 1, \ldots, m. \end{aligned}$$

we have that $ROC = [\tilde{C}_1, \ldots, \tilde{C}_n]^T$ and $COC = [\tilde{C}'_1, \ldots, \tilde{C}'_m]$ and, as we have already said, $TOC = ROC + COC$. The second difference, is that instead of selecting the row or column with the highest penalty, it takes into consideration the three rows or columns with the highest penalty. In Algorithm 3 we find a schematic explanation of this heuristic.

Algorithm 3: Improved Vogel's approximation method

1. Calculate the penalty for each row and column, this is done by taking the smallest cost out of the row and to the second smallest cost out of the row and the same way for the columns.
2. Select the three rows or columns with the three highest penalties, in case of ties we will select those with the lowest indices.
3. For each of the three selected rows or columns, it calculates the transport cost by assigning to the empty cell with the lowest cost of the selected row or column the largest quantity of goods that we can (as always this amount will be the minimum between demand and supply in the corresponding nodes).
4. Choose the minimum of the three transportation costs chosen in step 3. Break any possible ties by choosing the cell corresponding to the row or column with the lowest index.
5. Subtract the quantity already allocated to the corresponding elements of the vectors offer and demand. One of them will be canceled, if it is demand fill with 0 the rest of the elements of the column, if it is offer do the same with the row.
6. Until we have $m + n - 1$ boxes with a positive value check if the row or column deleted was providing the minimum for any column or row. If so, go to step 1, otherwise go to step 2.

In the literature we can find evidence that both the northwest corner algorithm and Vogel's algorithm provide an initial basic feasible solution. To our knowledge, there is no explicit proof of this property for the IVAM algorithm. We include here a short result covering this fact.

Proposition 2.8. *The solution provided by the IVAM algorithm is a basic feasible solution.*

Proof. This follows from Theorem 1, p.303 in Dantzig (1963), since the algorithm proceeds as in cases 1, 2 and 3 in this reference. □

Once we have a feasible basic initial solution, we have to iterate this solution in order to reach the optimal solution, so that in each step, the feasible solution will have smaller cost. To do this we will use the method known as MODI. An scheme of this algorithm can be found in algorithm 4.

Algorithm 4: MODI

1. Look for the non-basic cell whose contribution to cost is the lowest.
2. If the contribution to cost is equal or greater than 0 the current solution is optimal, otherwise continue to step 3.
3. Find the basic cell that needs to be removed from the current solution so that we can introduce the not basic cell selected in step 1.
4. Redistribute the goods between the boxes to get a new basic feasible solution and return to step 1.

We will say that a cell is a basic cell if it is part of the current solution and non-basic if it is not. The first step is to search through all the non-basic cells to find out which one would reduce the transportation cost of the problem the most. For this we will calculate the net unit contribution of each cell to the cost, this contribution can be either negative or positive (if it is 0 we consider it as positive). If the contribution is positive it means that the solution is better without that cell than with it and when the contribution of all cells is positive we will have reached the optimal solution. This may occur in the first iteration and the initial feasible basic solution may be optimal, but the larger the problem is, the less likely it is to occur.

We will denote the contribution of each cell by e_{ij}. To calculate it we have to solve a system of equations. To each row we will assign a variable u_i, $(i = 1, \ldots, n)$ and to each column we will assign the variable v_j, $(j = 1, \ldots, m)$. For each basic cell (i, j) we have the equation $u_i + v_j = C_{ij}$. This would make a total of $n + m - 1$ equations and $n + m$ unknowns, in order to get a system with a unique solution we need to add the equation $u_1 = 0$. Once the system is solved, we will obtain the contribution of each non-basic cell with

$$e_{ij} = C_{ij} - u_i - v_j. \tag{2.14}$$

Among all the contributions we keep the minimum, as long as this minimum is negative we will have to do another iteration and get the next feasible solution. In this case we have to keep the cell that provides the minimum (in case of a tie we keep the lowest index)

to introduce it in the new solution. As the current solution already has all the offer and demand distributed among its cells, we have to remove a basic cell, so a new cell can enter, and redistribute the goods. To determine which cell will be the one that comes out, we will have to find a loop that includes the new cell and several basic ones.

A *loop* is a path that begins and ends in the same cell. We consider as cells of the path the initial cell and the sequence of all the basic cells in the path. For a path to be considered a loop it must have the following properties:

1. The starting and ending cell must be the same.
2. Two consecutive cells must be either in the same column or in the same row.
3. There cannot be three consecutive cells in the same row or column.
4. There must be an even number of cells along the path.

In Figure 2.4 we see two examples of paths that do constitute a loop and in Figure 2.5 we see two other examples of paths that do not form a loop: the first is not a loop because it fulfills neither the second nor the fourth property and the second because it fulfills neither the third one nor the fourth.

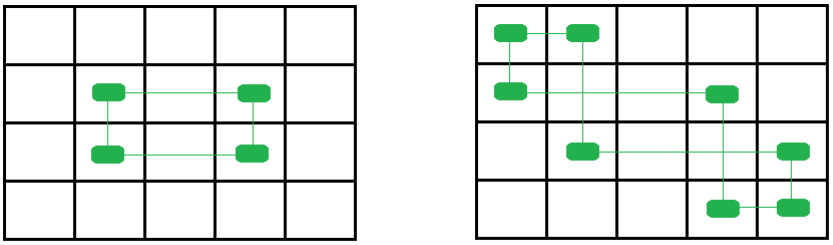


Figure 2.4: Examples of paths that constitute a loop.

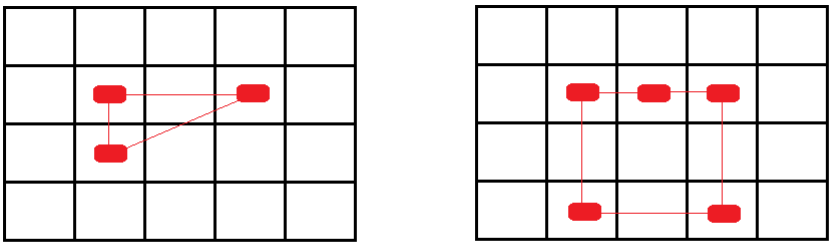


Figure 2.5: Examples of paths that are not loops.

It is known that any non-basic cell forms a loop with a subset of basic cells (see, for example, Bazaraa et al. (2010)). Finally, we only need to redistribute the goods in such a way that the new basic cell gets some goods and one of the basic cells in the previous solution runs out of goods. We are going to number the cell of the path starting from the initial cell and we are going to calculate the minimum number of goods that has an even cell assigned. We will take this minimum from the goods assigned to all even cells and we will add it to all odd cells. This means that the basic cell that is going to be left out of the new solution is the one that gives us that minimum (in case there is more than one that provides that minimum we will only leave out the first of them and to the rest of them we will assign ε goods, where ε is a very small amount that does not affect the final cost). We now have a new, feasible basic solution to begin the whole process jut described with it. Algorithm 5 describes schematically the simple algorithm for the transportation problem.

Algorithm 5: Simplex algorithm for optimal transportation problem.

1. Check if the problem is balanced and if not, balance it.
2. Construct the transportation tableau and obtain an initial solution using algorithm 3.
3. Propose and solve the system of equations.
4. Calculate the unit contribution of each non-basic cell with the formula (2.14) and get the minimum of all contributions.
   - If the minimum is positive or 0 we have the optimal solution and we have finished.
   - If the minimum is negative, proceed to step 5.
5. Get a loop.
6. Redistribute the goods between the cells of the loop. Return to step 3 with the new feasible basic solution.

In chapter 3 we propose an efficient algorithm, based on the ideas described here, for solving the discrete problem of partial transportation.

2.5.2 Gradient methods

Gradient descent algorithms are numerical methods for minimizing functions that can be used in many problems. In a statistical or machine learning context, part of the interest of these methods is their adaptability to minimize functions defined in high dimensional spaces, reducing sometimes the problem to the resolution of one-dimensional problems. As a reference on the application of these gradient methods in the context of machine

learning we can cite Bubeck (2015). Here we briefly summarize some basic aspects of the method.

Gradient methods are numerical minimization methods that arise in response to the problem of minimizing high dimensional functions by reducing them to one-dimensional problems. To describe the method, we will assume that $\mathcal{X} \subset \mathbb{R}^d$ is a convex, compact set and $Q : \mathcal{X} \to \mathbb{R}$ is a differentiable function that reaches its minimum value in $x^* \in \mathcal{X}$. Let ∇f be the gradient of Q, η the step size and $x_0 \in \mathbb{R}^d$ the starting point, the algorithm calculates in each iteration t

$$x_{t+1} = x_t - \eta \nabla Q(x_t).$$

Although these methods can get stuck in local minimums, if Q is convex and soft enough, the gradient algorithm produces a sequence of iterants converging to x^*. For example, if ∇Q is β-Lipschitz and we take $\eta = \frac{1}{\beta}$, then

$$Q(x_t) - Q(x^*) \leq \frac{2\beta \|x_0 - x^*\|^2}{t},\ t \geq 1,$$

(See Theorem 3.3 in Bubeck (2015)). We observe that the bound is independent of the dimension d, which may be interesting for solving high-dimensional problems (although the dimension often influences the Lipschitz constant β, associated with Q). Sometimes the speed of convergence towards the minimizer can be improved by choosing the appropriate step width. With variable-pass methods (η_t depending on t) an exponential bound can be achieved by means of *Nesterov's acceleration* (see Theorem 3.18 in Bubeck (2015)).

An important aspect in gradient methods is, as we see, the choice of step width. Once the direction in which we explore the space is chosen (the one given by the gradient) the ideal displacement would be the one that takes us to the point where the minimum is reached along that direction. Searching for this optimal displacement may be computationally inadvisable.

An alternative strategy is to change search direction and limit it to directions along which it is easy (computationally) to find the optimal displacement. This principle is behind the coordinate descent method. This is an iterative algorithm for fast resolution of large scale problems where classic minimization algorithms are not feasible. This heuristic algorithm produces a sequence that, under certain conditions, converges to the minimizer (see Friedman et al. (2007)). Shortly, the algorithm minimizes in each iteration in the direction of one of the coordinates given the current value of the other coordinates. The minimization will end when the improvement of the objective value between two iterations is very small.

The optimization problem to be solved by this algorithm is

$$\min_{\beta \in \mathbb{R}^d} Q(\beta). \tag{2.15}$$

In addition to function Q, the algorithm will need the gradient of this function in each coordinate. The gradient will be denoted by $G_j(\beta)$. If there is no gradient, a subgradient will suffice.

The algorithm works as follows. We start from an initial value for $\beta^{[0]}$ and $m = 0$. On each iteration it takes $m = m + 1$ and $j = m(\mod p)$ and calculates $G_j(\beta_{-j}^{[m-1]})$ where $\beta_{-j}^{[m-1]} = (\beta_1^{[m-1]}, \ldots, \beta_{j-1}^{[m-1]}, 0, \beta_{j+1}^{[m-1]}, \ldots, \beta_p^{[m-1]})$. We take

$$\beta_j^{[m]} = \begin{cases} 0 & \text{if } G_j(\beta_{-j}^{[m-1]}) = 0 \\ \underset{\beta_j \in \mathbb{R}}{\arg\min}\, Q(\beta_{+j}^{[m-1]}) & \text{other case} \end{cases},$$

with $\beta_{+j}^{[m-1]} = (\beta_1^{[m-1]}, \ldots, \beta_{j-1}^{[m-1]}, \beta_j, \beta_{j+1}^{[m-1]}, \ldots, \beta_p^{[m-1]})$ (as we said at the beginning only one of the coordinates changes in each iteration). For every $\beta^{[m]}$ we calculate $Q_m := Q(\beta^{[m]})$ and when the difference between Q_m and Q_{m-1} is small enough the algorithm ends.

Algorithm 6: Coordinate descent minimization

1. Choose $\beta^{[0]}$ and set $m = 0$.
2. Set $m = m + 1$ and $j = m \mod p$
3. Calculate $G_j(\beta_{-j}^{[m-1]})$. If its value is 0 then $\beta_j^{[m]} = 0$ in other case $\beta_j^{[m]} = \underset{\beta_j \in \mathbb{R}}{\arg\min}\, Q(\beta_{+j}^{[m-1]})$.
4. Repeat steps 2 and 3 until convergence.

Although the number of iterations that the algorithm needs in order to reach the optimum can be very large, especially in high-dimensional problems, the simplicity that, in some problems, has each one of them, makes that the resolution of the problem (2.15) is much faster than with other methods. A relevant example of this situation is LASSO. In this case step 3 in Algorithm 6 is calculated using a simple explicit formula, making the coordinate descent algorithm one of the usual methods used to calculate the LASSO estimator.

2.5.3 Concentration algorithms

In the first section of this chapter we saw that trimming a sample consists in removing a series of observations from it. A trimmed estimator is one obtained from this trimmed

sample. In general, sample trimming is made in such a way that the observations that are eliminated are those with the worst impact over the fit to a certain model. In fact, one of the motivations for considering trimming methods is the fact observed in many contexts that a single observation can completely alter the fit to a model and thus the conclusions of a data analysis, including the setup of classification (see, for example, Alfons et al. (2013) for a recent survey). From a practical point of view, the adequacy of observations to a model will be measured according to some criterion function. The concentration step algorithm or C-steps is an iterative method for getting an approximate solution for minimization problems of this type. The algorithm, in an iterative way, removes observations from the sample, recovering those that are less harmful to the model and removing those that are more harmful.

We illustrate the idea by outlining its application to a linear regression problem with quadratic loss, as in Rousseeuw and Driessen (2006). Suppose we have n observations of the form $(x_i, y_i) = (x_{i1}, \ldots, x_{id}, y_i) \in \mathbb{R}^{d+1}$ and that we can estimate the parameter vector $\beta = (\beta_1, \ldots, \beta_d)$ in such a way that

$$y = x_1\beta_1 + \ldots + x_d\beta_d + \varepsilon.$$

The presence of atypical observations or outliers can seriously affect the estimation of β. In order to limit the impact of possible outliers on the estimate, the LTS estimator (Least trimmed squares) is

$$\hat{\beta}_{LTS,\alpha} = \arg\min_{\beta\in\mathbb{R}^d} \min_{H\in\mathcal{H}} \sum_{i\in H} (y_i - x_i^T\beta)^2$$

where we assume that α is such that $n(1-\alpha) = h$ is an integer and $\mathcal{H}$ is the class of subsets of $\{1, \ldots, n\}$ with h elements. Obviously, for fixed $\beta \in \mathbb{R}^d$

$$\min_{H\in\mathcal{H}} \sum_{i\in H} (y_i - x_i^T\beta)^2 = \sum_{i=1}^{h} (r^2)_{i:n}(\beta)$$

where $(r^2)_{1:n}(\beta) \leq \ldots \leq (r^2)_{n:n}(\beta)$ are the ordered square residuals.

The algorithm starts with a first set H_1 randomly generated so that $|H_1| = h$ and calculates

$$\hat{\beta}_1 = \arg\min_{\beta\in\mathbb{R}^d} Q(\beta)|_{H_1} = \arg\min_{\beta\in\mathbb{R}^d} \sum_{i\in H_1} (r_1(i))^2$$

and

$$Q_1 := \sum_{i\in H_1} (r_1(i))^2$$

where $r_1(i) = y_i - (x_{i1}\hat{\beta}_1^1 + \ldots + x_{id}\hat{\beta}_d^1)$. For the first iteration chooses H_2 in such a way that $\{|r_1(i)| : i \in H_2\} = \{|r_1|_{1:n}, \ldots, |r_1|_{h:n}\}$ where $|r_1|_{1:n} \leq \ldots \leq |r_1|_{h:n} \leq \ldots \leq |r_1|_{n:n}$. In the

same way we calculate $\hat{\beta}_2$ and Q_2. It can be proved (see Rousseeuw and Driessen (2006)) that $Q_2 \leq Q_1$ and, then, that the objective value of the residuals function decreases with each iteration. The steps from H_k to H_{k+1} are known as C-Steps, short for concentration steps.

Since the set H_1 is randomly chosen, to reduce computation time instead of generating it with h elements we can choose a smaller size (for instance, Alfons et al. (2013) suggests choosing a set of three elements).

Algorithm 7: C-Steps

1. Randomly generate an initial set of H_1 of l elements.
2. Calculate $\hat{\beta}_1$ and Q_1.
3. Sort the errors obtained in step 2 and get H_2 with the remaining h couples that have the smallest errors.
4. Repeat steps 2 and 3 with $\hat{\beta}_2$ to get H_3.
5. Make *n.init* repetitions of steps 1 through 4 and choose the *n.cSteps* subsets H_3 that provide the smallest generalized empirical errors.
6. For every H_3 chosen in step 5 we repeat steps 2 and 3 until the difference between the trimmed minimum square error of two successive iterations is small enough.
7. From the *n.cSteps* errors that we get in step 6 we choose the smallest and the corresponding $\hat{\beta}$ will be the optimal one.

The final output of this algorithm can be affected by the choice of the initial set. The descent property only guarantees convergence to a local optimum. This is enhanced by choosing a large number, *n.init*, of random initializations. For each one of them, we perform two iterations of the algorithm since it is usually possible to differentiate after only 2 or 3 iterations if the proposed initial solution will lead to a good solution or not. Of those *n.init* solutions we keep the best *n.cSteps* to apply the algorithm until convergence. Of these *n.cSteps* results we keep the best. The good performance of this concentration algorithm with many initializations has been tested, for example, in Rousseeuw and Driessen (2006).

Chapter 3

Statistical methods related to partial transportation problem

This chapter presents the problem of optimal partial transportation and its connections to methods based on trimming and contamination models.

The problem of partial transportation is a variant of the classic optimal transportation problem described in section 2.3. Partial transportation problem also consists in transporting a mass quantity from an origin to a destination at the lowest possible cost. The difference between these two problems is that in the problem of partial transportation there is an excess of mass offered at the origin that we are not going to transport and, possibly, it is not necessary to satisfy all the demand at the destination but only a fraction of it.

According to section 2.1 there is an equivalent way to express contaminated models in terms of trimming sets. This equivalence takes on a more convenient shape when a metric is used between probabilities for which trimming sets are closed. Wasserstein metrics, associated with the problem of optimal transportation, have that attribute. This chapter exploits this fact to properly formulate contamination models in terms of the partial optimal transportation problem.

With the exception of the one-dimensional case, the calculation of optimal transportation plans does not accept simple closed expressions. These difficulties are transmitted (even in the one-dimensional case) to partial optimal transportation problem and numerical methods are necessary for the approximate calculation of optimal solutions. One possibility is to replace origin and destination distributions with discrete versions. This possibility occurs naturally when working with sample distributions. For this reason, we

will devote the section 3.2 to the study of the discrete problem of partial transportation, with special attention to designing an efficient algorithm for calculating solutions.

Another interesting situation corresponds to the case in which we want to compare a sample distribution with a continuous model. This leads to a semi-discrete version of partial optimal transportation problem. This problem is addressed in section 3.3. With the help of Kantorovich's duality for the optimal transportation problem, we will prove that the semi-discrete problem accepts an equivalent formulation in terms of minimizing a convex and differentiable function over a convex set of constraints. This allows the application of standard convex optimization methods for the approximate estimation of the semi-discrete partial transportation cost, such as the gradient descent algorithm described in section 2.5.2. However, the numerical calculation of the gradient in this partial transportation problem is too costly, which has led to the consideration of a type of stochastic gradient algorithm for the effective calculation of the partial transportation cost in this semi-discrete case.

Section 3.4 presents an application of the partial transportation problem to the validation of contaminated models. The approach is that the random data generator will belong to a contamination neighbourhood of a given model if the contamination level is large enough. On the other hand, classic fitting contrasts to a model generally do not allow us to conclude that the model is valid. At best, when the model is not rejected, the weak conclusion is that there is not enough evidence against the model. Here we intend to evaluate how much contamination we must admit to give the model as valid. We reformulated the problem of fitting a model to the model selection point of view. We measure the suitability of a model by means of the cost of transporting the empirical measure to a contamination neighbourhood of the model. This distance decreases as contamination levels increase. By adding an appropriate penalization, it is possible to guarantee (see Theorem 3.16) a good selection of the contamination level.

Finally, section 3.5 describes implementation details of the algorithms introduced in sections 3.3 and 3.4 and presents the results of applying the essential model validation procedure to simulated data.

3.1 Trimming and Wasserstein metrics

Similarly to how Wasserstein metrics and the optimal transportation problem relate to each other, a relationship can be established between the partial optimal transportation problem and Wasserstein distances between distribution's trimming sets.

In the problem of partial optimal transportation we have an excess of offer, so that it is not necessary to serve all the mass at origin, or we have some slackness with respect to the demand and it is only necessary to satisfy a fraction of this or both situations at the same time. Given an origin distribution P, a destination distribution Q and $\alpha_1, \alpha_2 \in [0,1]$, a partial transportation plan will be a probability $\pi \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the metric spaces where distributions P and Q are defined respectively, so that

$$\pi(A \times \mathcal{Y}) \leq \frac{1}{1-\alpha_1} P(A) \ \forall A \tag{3.1}$$

$$\pi(\mathcal{X} \times B) \leq \frac{1}{1-\alpha_2} Q(B) \ \forall B. \tag{3.2}$$

We can express the cost of partial optimal transportation as

$$\inf_{\pi \in \Pi_{\alpha_1,\alpha_2}(P,Q)} \int c(x,y) d\pi(x,y), \tag{3.3}$$

where $\Pi_{\alpha_1,\alpha_2}(P,Q)$ is the set of all probabilities π in $\mathcal{X} \times \mathcal{Y}$ that satisfy conditions (3.1) and (3.2).

Consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. If we set $c(x,y) = \|x-y\|^p$, (3.3) is the same as

$$\mathcal{W}_p^p(\mathcal{R}_{\alpha_1}(P), \mathcal{R}_{\alpha_2}(Q)).$$

In Álvarez-Esteban et al. (2012) it is proved that this distance can be obtained as

$$\mathcal{W}_p(\mathcal{R}_{\alpha_1}(P), \mathcal{R}_{\alpha_2}(Q)) = \min_{P' \in \mathcal{R}_{\alpha_1}(P), Q' \in \mathcal{R}_{\alpha_2}(Q)} \mathcal{W}_p(P', Q').$$

If instead of considering an excess of mass both at origin and destination we consider that there is only excess in one of them, we obtain that, in this case, the optimal partial transportation plan is precisely $\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q)$ and, again, from Álvarez-Esteban et al. (2012) we get

$$\mathcal{W}_p(\mathcal{R}_\alpha(P), Q) = \min_{P' \in \mathcal{R}_\alpha(P)} \mathcal{W}_p(P', Q)$$

As a direct consequence of Proposition 2.1 of Álvarez-Esteban et al. (2011) and the fact that the trimming set is compact for the Wasserstein metric, we can obtain the following result that relates contamination models and trimming sets.

Proposition 3.1. *If $P, Q, P' \in \mathcal{P}_p(\mathcal{X})$ then*

$$P = (1-\alpha)Q + \alpha P' \iff \mathcal{W}_p(\mathcal{R}_\alpha(P), Q) = 0.$$

Similarly, we can use the Wasserstein distance between trimming sets to determine if two distributions are similar at a level α, this is illustrated by the following result from Proposition 2 of Álvarez-Esteban et al. (2012).

Proposition 3.2. *If* $P, Q, R, P', Q' \in \mathcal{P}_p(\mathcal{X})$ *then*

$$\begin{cases} P = (1-\alpha)R + \alpha P' \\ \\ Q = (1-\alpha)R + \alpha Q' \end{cases} \iff \mathcal{W}_p(\mathcal{R}_\alpha(P), \mathcal{R}_\alpha(Q)) = 0.$$

One objective of this thesis is to propose statistical methods based on trimming that can be used in practice. Some methods based on Wasserstein distance between trimming sets have been treated in Álvarez-Esteban et al. (2012). These methods are limited to univariate data and a main objective of this work is to extend the applicability of this type of methods to the multivariate case. A first step is to have a good method of calculating empirical versions of Wasserstein distances between trimming sets. A response to this issue will be given in section 3.2, with an algorithm for calculating the partial optimal transportation in the discrete case. In other applications it may be necessary to calculate the partial transportation cost for a continuous model. As we have already said, this will lead to a semi-discrete problem whose treatment is discussed in section 3.2.1.

3.2 The discrete partial transportation problem

In this section we focus on the problem described by the equations (3.1) to (3.3) in the particular case where P and Q have finite support. More precisely, we will assume that $P\{x_i\} = p_i,\ i = 1, \ldots, n$ and $Q\{y_j\} = q_j,\ j = 1, \ldots, m$ (so that $p_i > 0,\ q_j > 0,\ p_1 + \cdots + p_n = q_1 + \cdots + q_m = 1$). Each element of the cost matrix will be of the form $C_{ij} = \|x_i - y_j\|^p$. A probability over $\{x_1, \ldots, x_n\} \times \{y_1, \ldots, y_m\}$ can be described by the probabilities of the pairs (x_i, y_j), which we will denote by Π_{ij}. Then the conditions (3.1) and (3.2) become

$$\begin{aligned} \sum_{j=1}^{m} \Pi_{ij} &\leq \frac{p_i}{1-\alpha_1},\ i = 1, \ldots, n \\ \sum_{i=1}^{n} \Pi_{ij} &\leq \frac{q_j}{1-\alpha_2},\ j = 1, \ldots, m \\ \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{ij} &= 1,\ \Pi_{ij} \geq 0,\ i = 1, \ldots, n,\ j = 1, \ldots, m, \end{aligned}$$

so

$$\begin{aligned}
\mathcal{W}_p^p(\mathcal{R}_{\alpha_1}(P), \mathcal{R}_{\alpha_2}(Q)) = \min_{\Pi} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{ij} C_{ij} \qquad (3.4)\\
\text{s.t.} \quad & \sum_{j=1}^{m} \Pi_{ij} \leq \frac{p_i}{1-\alpha_1},\ i = 1, \ldots, n\\
& \sum_{i=1}^{n} \Pi_{ij} \leq \frac{q_j}{1-\alpha_2},\ j = 1, \ldots, m\\
& \sum_{i=1}^{n}\sum_{j=1}^{m} \Pi_{ij} = 1\\
& \Pi_{ij} \geq 0,\ i = 1, \ldots, n,\ j = 1, \ldots, m.
\end{aligned}$$

More generally we can consider the problem

$$\begin{aligned}
\text{(PTP)} \quad \min_{X=(x_{ij})_{i,j=1}^{n,m}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} C_{ij}\\
\text{s.t.} \quad & \sum_{j=1}^{m} x_{ij} \leq o_i,\ i = 1, \ldots, n \qquad (3.5)\\
& \sum_{i=1}^{n} x_{ij} \leq d_j,\ j = 1, \ldots, m \qquad (3.6)\\
& \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = 1,\ x_{ij} \geq 0,\ i = 1, \ldots, n,\ j = 1, \ldots, m. \qquad (3.7)
\end{aligned}$$

For the problem to be feasible it is necessary that $\sum_{i=1}^{n} o_i \geq 1$ and $\sum_{j=1}^{m} d_j \geq 1$.

This formulation is very similar to that of the discrete optimal transportation problem discussed in section 2.5.1. However, there is a fundamental difference. Inequalities deal symmetrically with offer and demand nodes. If it was not for restriction (3.7), which guarantees that X is a probability (the total displaced mass is 1), the minimum in (PTP) would be 0. This means that (PTP) is a problem of a different nature from that described in section 2.5.1 and that it is not possible to rewrite (PTP) directly as a balanced transportation problem.

In order to turn this problem into a balanced optimal transportation problem, and to be able to apply the usual algorithms, we must transform inequalities (3.5) and (3.6) into equalities. As with the classic unbalanced problem, the way to do this is by adding dummy nodes. In this case we must add an origin and a destination node. Since these nodes are dummy, the shipments made from the dummy source node or those that reach the dummy

destination node will also be dummy and, therefore, we do not want them to affect the solution of the problem. Thus, the cost between any source node and the destination dummy node and the cost from a source dummy node to any destination node will be 0. Besides, we also have a forbidden path. We cannot transport any goods between the two dummy nodes. Since eliminating this path would destroy the structure of the problem, the way to solve it is to assign this path a cost that is high enough to ensure that in any optimal solution this path will not be taken.

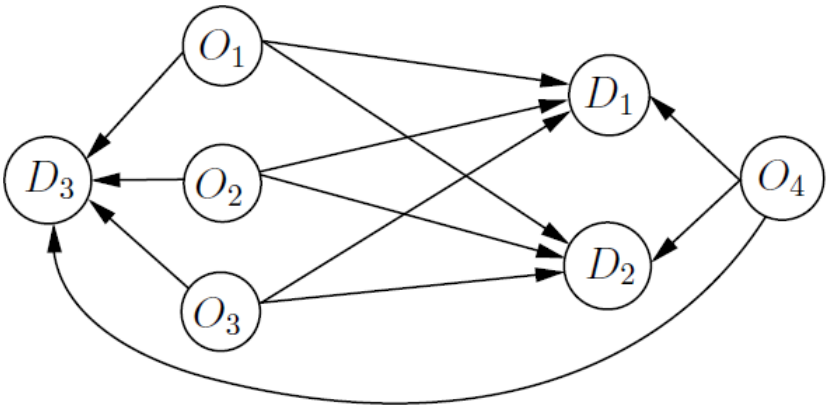


Figure 3.1: Graph of a partial transportation problem with 3 origin nodes, 2 destination nodes and 2 dummy nodes.

In Figure 3.1 we show the extended graph corresponding to a partial transportation problem between 3 source nodes and 2 demand nodes. Added dummy nodes are D_3 and O_4. Transportation costs between O_1, O_2, O_3 and D_3 are null, the same as O_4 to D_1 or D_2. Our next result provides a possible assignment to the transportation cost between dummy nodes that ensures that nothing is transported between them in the optimal solution.

Proposition 3.3. *Given a balanced transportation problem of the form*

$$
\begin{aligned}
(BTP) \quad \min_{x_{11},\ldots,x_{nm}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}C_{ij} \\
s.t. \quad & \sum_{j=1}^{m} x_{ij} = o_i,\ i = 1,\ldots,n \\
& \sum_{i=1}^{n} x_{ij} = d_j,\ j = 1,\ldots,m \\
& x_{ij} \geq 0,\ i = 1,\ldots,n,\ j = 1,\ldots,m.
\end{aligned}
$$

with $C_{ij} \in \mathbb{R}$ *known* $\forall i, j$. *If* $\exists\ k, l /$

$$C_{kl} \geq \frac{DC}{d_l} \tag{3.8}$$

where $D = \sum_{j=1}^{m} d_j$ *and* $C = (\sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}) - C_{kl}$ *then in the optimum* $\hat{x}_{kl} = 0$.

Proof: Let X and Y be two feasible solutions to the problem (not necessarily optimal), with $y_{kl} = 0$. We know that if we take C_{kl} big enough and $x_{kl} \neq 0$ then $\forall i \in \{1, \ldots, n\} \backslash \{k\}$ and $\forall j \in \{1, \ldots, m\} \backslash \{l\}$ there are y_{ij} solution to the problem so that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} C_{ij} \geq \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} C_{ij}. \tag{3.9}$$

Clearly

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} C_{ij} = x_{kl} C_{kl} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} x_{ij} C_{ij}$$

and

$$\sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} C_{ij} = y_{kl} C_{kl} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} y_{ij} C_{ij} = 0 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} y_{ij} C_{ij}.$$

From (3.9)

$$x_{kl} C_{kl} \geq \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} y_{ij} C_{ij} - \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} x_{ij} C_{ij} = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} (y_{ij} - x_{ij}) C_{ij}. \tag{3.10}$$

As $y_{ij} \leq d_j$ and $0 \leq x_{ij}$ we can bound

$$y_{ij} - x_{ij} \leq d_j < D.$$

So, we have

$$\sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} (y_{ij} - x_{ij}) C_{ij} \leq \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} d_j C_{ij} < D \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} C_{ij}.$$

and going back to (3.10)

$$x_{kl} C_{kl} \geq D \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (i,j) \neq (k,l)}}^{m} C_{ij},$$

then, as $x_{kl} \leq d_l$,

$$C_{kl} \geq \frac{D\sum_{i=1}^{n}\sum_{\substack{j=1 \\ (i,j)\neq(k,l)}}^{m} C_{ij}}{x_{kl}} \geq \frac{D\sum_{i=1}^{n}\sum_{\substack{j=1 \\ (i,j)\neq(k,l)}}^{m} C_{ij}}{d_l}.$$

□

With this we conclude that we can solve the partial transportation problem by means of a classic balanced discrete transportation problem by adding dummy nodes as described and imposing the conditions

$$\begin{aligned}
d_{m+1} &= \sum_{i=1}^{n} o_i - 1, \\
o_{n+1} &= \sum_{j=1}^{m} d_j - 1, \\
C_{(n+1)j} &= 0,\ j = 1, \ldots, m, \\
C_{i(m+1)} &= 0,\ i = 1, \ldots, n \ \text{ and} \\
C_{(n+1)(m+1)} &= \frac{\left(\sum_{j=1}^{m} d_j\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}\right)}{\sum_{i=1}^{n} o_i - 1}.
\end{aligned}$$

Therefore, taking

$$\tilde{C} = \begin{pmatrix} C & 0_{n\times 1} \\ 0_{1\times m} & \frac{(1-\alpha_1)DC}{O\alpha_1} \end{pmatrix}.$$

we can formulate the partial transportation problem as follows,

$$
\begin{aligned}
\text{(PTPB)} \quad \min_{X=(x_{ij})_{i,j=1}^{n,m}} \quad & \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} x_{ij}\tilde{C}_{ij} \\
\text{s.t} \quad & \sum_{j=1}^{m+1} x_{ij} = o_i,\ i = 1,\dots,n \\
& \sum_{j=1}^{m+1} x_{(n+1)j} = \sum_{j=1}^{m} d_j - 1 \\
& \sum_{i=1}^{n+1} x_{ij} = d_j,\ j = 1,\dots,m \\
& \sum_{i=1}^{n+1} x_{i(m+1)} = \sum_{i=1}^{n} o_i - 1, \\
& \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} x_{ij} = 1 \\
& x_{ij} \geq 0,\ i = 1,\dots,n+1;\ j = 1,\dots,m+1,
\end{aligned}
$$

This is precisely an optimal transportation problem and, therefore, we can apply algorithm 5 we saw in section 2.5.1 to solve it.

Returning to the problem of calculating the Wasserstein distance between sets of trimmings of probabilities with finite support, using the above we can rewrite problem (3.4) as a transportation problem as follows,

$$
\begin{aligned}
\mathcal{W}_p^p(\mathcal{R}_{\alpha_1}(P), \mathcal{R}_{\alpha_2}(Q)) = \min_{\Pi} \quad & \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} \Pi_{ij} C_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{m+1} \Pi_{ij} = \frac{p_i}{1-\alpha_1},\ i = 1, \ldots, n \\
& \sum_{j=1}^{m+1} \Pi_{(n+1)j} = \frac{\alpha_2}{1-\alpha_2} \\
& \sum_{i=1}^{n+1} \Pi_{ij} = \frac{q_j}{1-\alpha_2},\ j = 1, \ldots, m \\
& \sum_{i=1}^{n+1} \Pi_{i(m+1)} = \frac{\alpha_1}{1-\alpha_1} \\
& \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} \Pi_{ij} = 1 \\
& \Pi_{ij} \geq 0,\ i = 1, \ldots, n+1,\ j = 1, \ldots, m+1.
\end{aligned}
$$

3.2.1 A fast algorithm for the partial transportation problem

In this section we will obtain an algorithm for the R statistical package (R Core Team (2013)) that will allow us to efficiently solve the optimal transportation problem and the partial transportation problem seen in the previous section.

Nowadays, there is function `lp.transport` in the `lpSolve` package (see Berkelaar (2013)) which calculates the solution to a balanced optimal transportation problem, but this function is only efficient if we work with small datasets. This is why we have implemented the algorithm described below, which calculates the optimal solution to the transportation problem faster than `lp.transport`. In addition, function `lp.transport` presents problems when it has to solve big problems, while our algorithm has been successfully tested with samples os size 10000×10000 nodes. We have called our algorithm `Partial.Transport`, this name comes from the fact that, in addition to solving the transportation problem quickly and efficiently, it is also prepared to solve the partial transportation problem by automatically making all the changes on the formulation of the problem seen in the previous section.

In section 2.5.1 we introduced the simplex for the calculation of optimal transportation plans. Below we will review the modifications introduced to obtain an improvement in

the efficiency of the solution. First, we are going to compare the IVAM heuristic with the others considered to highlight its advantages. Table 3.1 presents the average number of iterations of algorithm 4 needed to reach the optimal solution when starting from the initial solution provided by each heuristic for 10 different problems. In problems up to 500×500 it is clearly seen that NC is much worse than the others. Due to this and the immense amount of time required to reach the optimum in greater problems we have skipped this calculations.

Table 3.1: Number of iterations

	Iterations		
Size	NC	VAM	IVAM
10×10	26	9.3	8
100×100	740.5	227.2	212.5
200×200	1956	525	470.3
500×500	5966.2	1767.5	1467.9
1000×1000	-	6518.4	5996.2
2000×2000	-	15256.8	12970.7
5000×5000	-	44771	39400.4

In table 3.1 we can see not only that the average of iterations required by the program to reach the optimum is lower when we use IVAM, but also that the difference with the other two algorithms becomes larger with the sample size. This translates into significantly less time for calculating the optimal solution.

The second modification comes in the resolution of the system because although it is very simple and also very fast when we are working with small problems, increasing the size of the problem also increases the size of the system and, therefore, the computation time needed to solve it. Since we are looking to implement an algorithm that resolves the transportation problem fast, we need to find a way to reduce the computation time of this solution.

If we write the system in matrix notation $Ax = b$, where b is the vector whose $n+m-1$ first components will be the costs of the basic boxes and the last component will be 0, $x = (u_1, \ldots, u_n, v_1, \ldots, v_m)^T$ and A is the $(n+m) \times (n+m)$ matrix which will have only two elements different from 0 in the first $n+m-1$ rows and whose last row will be formed by a 1 in the first column and 0 in the rest.

The time to solve this system of equations using matrix algebra becomes too long when the problem grows. To avoid this we will take advantage of the fact that matrix A is sparse and that we already know the value of the first variable. Instead of matrix A we will be working with $\tilde{A}$ of size $(n+m-1)\times 2$. Matrix $\tilde{A}$ consists of the number of the columns in which the 1's of each row were placed, i. e. if the first row was the one corresponding to cell $(1,1)$, $\tilde{A}_{1\cdot} = [1, 1+n]$. Clearly we cannot express the system in matrix form using $\tilde{A}$ instead of A, but, by means of algorithm 8 we can get all values of the vector x in a quick way. We will also need a list where we are going to write down the rows of $\tilde{A}$ we have already used.

Algorithm 8: Resolution of the system

1. Find the first unused row of $\tilde{A}$ and call it i.
2. See if we have already obtained either the value of $x_{\tilde{A}_{i1}}$ or of $x_{\tilde{A}_{i2}}$, if so, we get the other one by solving $x_{\tilde{A}_{i1}} + x_{\tilde{A}_{i2}} = b_i$.
3. If we've got all the components of x we're done, if not make $i = i+1$ and continue to step 4.
4. If i is in the list of used rows make $i = i+1$ and repeat this step, if $i > n+m-1$ return to step 1. Otherwise go to step 2.

Once the system is solved we will have to calculate $\min_{i,j} e_{i,j} = X_{ij} - x_i - x_{j+n}$ where the possible values (i,j) are the components of all the non-basic cells.

Finally, to find the loop we have implemented a recursive algorithm that starts in the non-basic cell selected in the previous step and goes from basic cell to basic cell until returning to the initial one, giving as a result the path that has been passed through.

The operation of the algorithm is simple, in each step, starting from the second, receives a cell and the direction from which we have arrived at it. First, the algorithm checks if the received cell is the initial cell and, if so, it finishes the program and returns the path with the basic cells it has passed through. If this is not the case, makes a second check to see if the cell received is a feasible cell, that is, if we have not exceeded the limits of number of offer or demand nodes. If it is not within the limits, the algorithm ends that path and returns nothing. If it is within the limits then the algorithm calls itself back with the following cells to investigate. In order to decide which are these cells, it makes yet another check: if the received cell is not basic, it makes a single call with the next cell in the same direction. If the cell is basic then there are two options, it is part of the path and therefore there is a change of direction or it is not part of the path and then the algorithm will treat it as if it was a non-basic cell. In the case where there is a change of direction the algorithm is launched twice, one for each possible direction with the cell

next to the current one in the appropriate direction.

Finally, the first step has yet to be described. In this case the algorithm receives the initial cell and starts the loop search without doing any checking since it is already known that we are in a loop cell. This implies that there has to be a change of direction in this cell when the loop is finished, i.e. it will be enough to look for the loop only in two directions up and down or left and right.

To make this clearer, assume that we are in the case of Figure 3.2 where the blue painted cells form the current feasible basic solution and the cell marked with the green dot is the initial box of the loop. The algorithm will start by searching for the loop in the two directions indicated by the arrows. When it goes down it will do all the checks and, as it is in a non-basic cell, it will continue with the cell below in the same direction. This time when it does the checks it will see that the cell is out of bounds and it will finish that search. Now, when it goes upwards, it will arrive at a basic cell, in this one the algorithm will be called three times, one with the cell to the left, one with the cell above and another with the cell to the right. For each of these three cells it will do the same thing again, ending the process or calling each other as many times as necessary until it finally finds the loop.

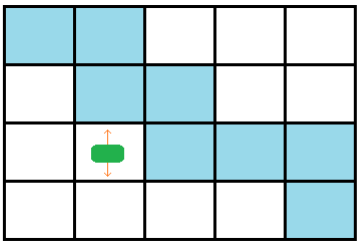


Figure 3.2: Example of how the algorithm works.

To finish this section, we have tested by simulations that our algorithm is significantly faster than `lp.transport`. In table 3.2 we have gathered the average time taken by each algorithm to solve the same problems with different dimensions. There have been 10 simulations of each size with an i7-1790 computer at 3.60 GHz and 16.0GB of RAM.

As you can see in table 3.2, our algorithm usually takes much less time than `lp.transport`. This difference is not very noticeable when the problems are not very big, but as the size increases this difference can be of several hours.

This algorithm can obviously solve the partial transportation problem once transformed into an ordinary transportation problem. It also solves the partial transport problem without having to input anything other than the original partial transport problem

Table 3.2: Comparison in seconds of average speed in the resolution of transportation problem.

Dimension	**lp.Transport**	**Partial.Transport**
100×100	0.084	0.029
200×200	0.607	0.315
300×300	2.302	0.949
500×500	12.234	4.439
1000×1000	113.436	55.5260
2000×2000	955.198	601.840
5000×5000	17404.363	13893.384

and the maximum trimming levels. In addition to making all these changes automatically, it also takes advantage of the shape of the matrix $\tilde{C}$ that has, except for the last element, a column and a row of 0 to speed up calculations of the initial solution.

3.3 A stochastic approximation algorithm for the computation of optimal transportation costs

In this section we will consider the problem of computing the 2-Wasserstein distance between probabilities P and Q in $\mathbb{R}^d$ when P has finite support. First, we will state and prove a lemma that we will need in the proof of the main result of this section.

Lemma 3.4. *Let P be a probability with finite support $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ with $P(x_i) = b_i > 0$ for $i = 1, \ldots, n$ and Q a Borel probability in $\mathbb{R}^d$ with finite second moment, then*

$$\mathcal{W}_2^2(P,Q) = \int_{\mathbb{R}^d} \|x\|^2 dP(x) + \int_{\mathbb{R}^d} \|y\|^2 dQ(x) - 2 \inf_{(z,u)\in\Phi} \left[\sum_{i=1}^{n} b_i z_i + \int u(y) dQ(y) \right], \quad (3.11)$$

where Φ is the class of pairs (z,u) such that $z \in \mathbb{R}^n$, $u \in L_1(Q)$ and $x_j y \leq z_j + u(y)$ for $1 \leq j \leq n$, $y \in \mathbb{R}^n$.

Proof: From the theory of duality for optimal transportation we can find, for example, in Villani (2003) that

$$\mathcal{W}_2^2(P,Q) = \max_{\substack{\tilde{\varphi},\tilde{\psi}\in L_1 \\ \tilde{\varphi}+\tilde{\psi}\leq\|x-y\|^2}} \left[\int \tilde{\varphi} dP + \int \tilde{\psi} dQ \right].$$

From here, taking

$$\begin{aligned}\tilde{\varphi} &= \|x\|^2 - 2\varphi(x)\\ \tilde{\psi} &= \|y\|^2 - 2\psi(y),\end{aligned}$$

and operating with them

$$\|x\|^2-2\varphi(x)+\|y\|^2-2\psi(y) = \tilde{\varphi}+\tilde{\psi} \leq \|x-y\|^2 = \|x\|^2+\|y\|^2-2x\cdot y \Leftrightarrow x\cdot y \leq \varphi(x)+\psi(y),$$

we arrive to

$$\mathcal{W}_2^2(P,Q) = \int_{\mathbb{R}^d} \|x\|^2 dP(x) + \int_{\mathbb{R}^d} \|y\|^2 dQ(y) - 2 \min_{\substack{\varphi,\psi\in L_1\\ \varphi+\psi\geq x\cdot y}} \left[\int \varphi dP + \int \psi dQ\right].$$

Since the support of P is finite, we have that

$$\int \varphi dP = \sum_{i=1}^n b_i\varphi(x_i) = \sum_{i=1}^n b_i z_i,$$

with $z_i = \varphi(x_i) \in \mathbb{R}$. With this, 3.3 will now be of the form

$$\begin{aligned}&x_i \cdot y \leq z_i + \psi(y)\ \forall i = 1,\dots,n \Leftrightarrow \psi(y) \geq x_i\cdot y - z_i\ \forall i = 1,\dots,n\\ &\Leftrightarrow \psi(y) \geq \sup_{i\in\{1,\dots,n\}} (x_i\cdot y - z_i).\end{aligned}$$

□

We can now proceed to state the main result of this section.

Theorem 3.5. *Let P be a probability with finite support $\{x_1,\dots,x_n\} \subset \mathbb{R}^d$ where $P(x_i) = b_i > 0$ for $i = 1,\dots,n$ and Q a Borel probability in $\mathbb{R}^d$ with finite second moment, then*

$$\mathcal{W}_2^2(P,Q) = \sum_{i=1}^n b_i\|x_i\|^2 + \int_{R^d} \|y\|^2 dQ(y) - 2 \min_{(z_1,\dots,z_n)\in B} V(z_1,\dots,z_n),$$

where V is the convex function given by

$$V(z_1,\dots,z_n) = \sum_{i=1}^n b_i z_i + E \max_{1\leq j\leq n} \big(x_j\cdot Y - z_j\big), \tag{3.12}$$

with B the closed ball centered in $-\frac{1}{2}(\|x_1\|^2,\dots,\|x_n\|^2)$ with radius $M/(\min_{1\leq j\leq n} b_j)$ where $M = \sum_{i=1}^n b_i\|x_i\|^2 + \int_{R^d}\|y\|^2 dQ(y)$ and Y is a random vector with distribution Q.

Proof. Starting from equality (3.11) for Wasserstein distance we saw in Lemma 3.4, we have that

$$\mathcal{W}_2^2(P,Q) = \sum_{i=1}^n b_i\|x_i\|^2 + \int_{R^d}\|y\|^2 dQ(y) - 2\inf_{(z,u)\in\Phi}\Big[\sum_{i=1}^n b_i z_i + \int u(y)dQ(y)\Big],$$

where $\Phi = \{(z,u):\ z\in\mathbb{R}^n,\ u\in L_1(Q),\ x_j\cdot y \le z_j + u(y)\ 1\le j\le n\ y\in\mathbb{R}^n\}$.
We will call $\tilde{u}(y) := \max\limits_{1\le j\le n}(x_j\cdot y - z_j)$ and given that $u(y)\ge\tilde{u}(y)$ and $(z,\tilde{u})\in\Phi$ we can see that

$$\inf_{(z,u)\in\Phi}\Big[\sum_{i=1}^n b_i z_i + \int u(y)dQ(y)\Big] = \inf_{z\in\mathbb{R}^n} V(z)$$

with V as in the statement.
Consider now $(z_1,\dots,z_n)\in\mathbb{R}^n$ and $u(y) = \max\limits_{1\le j\le n}(x_j\cdot y - z_j)$. We have

$$u(y) + \frac{\|y\|^2}{2} \ge \frac{\|x_j+y\|^2}{2} - z_j - \frac{\|x_j\|^2}{2} \ge -z_j - \frac{\|x_j\|^2}{2}.$$

Then if we set $a := \inf\limits_{y\in\mathbb{R}^d}\left(u(y) + \frac{\|y\|^2}{2}\right)$ we have that a is finite. If we replace $(z_1,\dots,z_n)$ by $(z_1+a,\dots,z_n+a)$ then $u(y)$ becomes $u(y)-a$ and V remains unchanged. As a consequence of this, for the minimization of V we only have to consider the points $(z_1,\dots,z_n)$ for which the corresponding function $u(y) = \max\limits_{1\le j\le n}(x_j\cdot y - z_j)$ satisfies

$$\inf_{y\in\mathbb{R}^d}\left(u(y) + \frac{\|y\|^2}{2}\right) = 0. \tag{3.13}$$

Assume that (3.13) holds and define $\tilde{z}_j = \sup\limits_{y\in\mathbb{R}^d}(x_j\cdot y - u(y))$. As $z_j \ge x_j\cdot y - u(y)$ for all y we have that $\tilde{z}_j \le z_j$, $j=1,\dots,n$. We are going to fix now $\tilde{u}(y) = \max\limits_{1\le j\le n}(x_j\cdot y - \tilde{z}_j)$. Then we have $\tilde{u}(y) = \max\limits_{1\le j\le n}(x_j\cdot y - \tilde{z}_j) \ge \max\limits_{1\le j\le n}(x_j\cdot y - z_j) = u(y)$. On the other hand, $\tilde{z}_j + u(y) \ge x_j\cdot y$ for all y and j and this implies that $u(y) \ge \max\limits_{1\le j\le n}(x_j\cdot y - \tilde{z}_j) = \tilde{u}(y)$. Then, $\tilde{u} = u$ and $V(\tilde{z}_1,\dots,\tilde{z}_n) \le V(z_1,\dots,z_n)$. We can observe that

$$\begin{aligned}\tilde{z}_j + \frac{\|x_j\|^2}{2} &= \sup_{y\in\mathbb{R}^d}\Big(x_j\cdot y + \frac{\|x_j\|^2}{2} - u(y)\Big)\\ &\ge \sup_{y\in\mathbb{R}^d}\Big(-\frac{\|y\|^2}{2} - u(y)\Big) = -\inf_{y\in\mathbb{R}^d}\Big(u(y) + \frac{\|y\|^2}{2}\Big) = 0.\end{aligned}$$

From where we conclude that $\tilde{z}_j \ge -\frac{\|x_j\|^2}{2}$. On the other hand,

$$\begin{aligned}&V(\tilde{z}_1,\dots,\tilde{z}_n) + \frac{1}{2}\sum_{i=1}^n b_i\|x_i\|^2 + \frac{1}{2}\int_{\mathbb{R}^d}\|y\|^2 dQ(y)\\ &= \sum_{i=1}^n b_i\Big(\tilde{z}_i + \frac{\|x_i\|^2}{2}\Big) + \int_{\mathbb{R}^d}\left(u(y) + \frac{\|y\|^2}{2}\right)dQ(y),\end{aligned}$$

this, by (3.13), implies that

$$\sum_{i=1}^{n} b_i\Big(\tilde{z}_i + \frac{\|x_i\|^2}{2}\Big) \leq V(\tilde{z}_1, \ldots, \tilde{z}_n) + \frac{1}{2}\sum_{i=1}^{n} b_i\|x_i\|^2 + \frac{1}{2}\int_{\mathbb{R}^d} \|y\|^2 dQ(y).$$

We know that V is a finite convex function in $\mathbb{R}^n$ and

$$\begin{aligned} \inf_{z\in\mathbb{R}^n} V(z) &= \frac{-1}{2}\Big(\mathcal{W}_2^2(P,Q) - \sum_{i=1}^{n} b_i\|x_i\|^2 - \int_{\mathbb{R}^d} \|y\|^2 dQ(y)\Big) \\ &\leq \frac{1}{2}\Big(\sum_{i=1}^{n} b_i\|x_i\|^2 + \int_{\mathbb{R}^d} \|y\|^2 dQ(y)\Big) =: C \end{aligned}$$

Obviously, $\inf_{z\in\mathbb{R}^n} V(z) = \inf_{z\in\mathbb{R}^n : V(z)\leq 2C} V(z)$. All this argumentation implies that we can limit the minimization to the set of points z such that

$$\begin{aligned} &z_j \geq -\frac{\|x_j\|^2}{2},\ j = 1, \ldots, n, \\ &\sum_{i=1}^{n} b_i\Big(z_i + \frac{\|x_i\|^2}{2}\Big) \leq 2C + \frac{1}{2}\sum_{i=1}^{n} b_i\|x_i\|^2 + \frac{1}{2}\int_{\mathbb{R}^d} \|y\|^2 dQ(y), \end{aligned}$$

which is a compact subset of $\mathbb{R}^n$. This proves that V reaches its minimum value, C, at some point z. Again, by the previous reasoning, we know that we can choose the optimal z such that $z_j \geq -\frac{\|x_j\|^2}{2}$, $j = 1, \ldots, n$ and

$$\sum_{i=1}^{n} b_i\Big(z_i + \frac{\|x_i\|^2}{2}\Big) \leq C + \frac{1}{2}\sum_{i=1}^{n} b_i\|x_i\|^2 + \frac{1}{2}\int_{\mathbb{R}^d} \|y\|^2 dQ(y) = \frac{1}{2}\mathcal{W}_2^2(P,Q) \leq M.$$

If we write $v_i = z_i + \frac{\|x_i\|^2}{2}$ then the set of v such that $\sum_{i=1}^{n} b_i v_i \leq M$, $v_i \geq 0$ is a closed convex set contained in the ball $\bar{B}(0, M/(\min_{1\leq j\leq n} b_j))$.

□

The following result corresponds to Proposition 4.2 in del Barrio and Loubes (2017), which proves that function V is differentiable if $Q \ll \ell^d$, the $d-$dimensional Lebesgue measure, and its gradient is obtained.

Proposition 3.6. *Let P be a probability with finite support $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ where $P(x_i) = b_i > 0$ for $i = 1, \ldots, n$ and Q a Borel probability in $\mathbb{R}^d$ such that $Q \ll \ell^d$ then the V function defined in* (3.12) *is differentiable and, furthermore, its gradient has the form*

$$\nabla V(z) = (b_1, \ldots, b_n) - (Q(A_1(z)), \ldots, Q(A_n(z))),$$

with

$$A_j(z) = \{Y \in \mathbb{R}^d :\ (x_j \cdot Y - z_j \geq \max_{i\neq j}(x_i \cdot Y - z_i))\},\ j = 1, \ldots, n.$$

We are now going to introduce an iterative algorithm that will allow us to compute effectively $\min_{z\in B} V(z)$ with V and z as in Theorem 3.5. The expression

$$V(z) = E(b \cdot z + \max_{1\leq j\leq n} (x_j \cdot -z_j))$$

suggests that we use some version of the stochastic approach algorithm that can be found in Nemirovski et al. (2008). We write

$$G(z, Y) = b - (I_{A_1(z)}, \ldots, I_{A_n(z)}) \tag{3.14}$$

and note that $EG(z, Y) = b - (Q(A_1(z)), \ldots, Q(A_n(z))) \in \partial V(z)$. Also with probability 1 (provided that Q has a density) $G(z, Y)$ will be equal to the difference between b and a vector for which one coordinate is 1 while the others are null. As a result, $\|G(z, Y)\| \leq 1 + \|b\|^2$ and, therefore,

$$E(\|G(z, Y)\|) \leq 1 + \|b\|^2.$$

We are taking now $z_1 \in B$. Set $N \in \mathbb{N}$ and define

$$\gamma = \frac{M}{(\min_{1\leq j\leq n} b_j)\sqrt{N(1 + \|b\|^2)}}.$$

For $t = 1, \ldots, N-1$ we consider $Y_1, \ldots, Y_{N-1}$ i.i.d. random variables with law Q and compute

$$z_{t+1} = \mathrm{Pr}_B(z_t - \gamma G(z_t, Y_t)),$$

where G is defined as in (3.14) and Pr_B denotes the metric projection over B, this is $\mathrm{Pr}_B(x) = \arg\min_{z\in B} \|y - x\|$. Finally we set

$$\tilde{z}_N = \frac{1}{N} \sum_{t=1}^{N} z_t.$$

With this notation we obtain the following result.

Theorem 3.7. *If* $\underline{V} = \min_{z\in B} V(z)$ *then*

$$E(V(\tilde{z}_N) - \underline{V}) \leq \frac{M\sqrt{1 + \|b\|^2}}{(\min_{1\leq j\leq n} b_j)\sqrt{N}}$$

with M *as in Theorem 3.5.*

This result follows easily from Proposition 3.6 and equation (2.21) in Nemirovski et al. (2008).

As a final note, we would like to point out that $\mathrm{Pr}_B(z)$ can be computed easily as follows. A change of location $z \mapsto z' = z + \frac{1}{2}(\|x_1\|^2, \ldots, \|x_n\|^2)$ reduces the problem to

projecting on B', the closed ball centered in the origin with radius $m = M/(\min_{1\leq j\leq n} b_j)$. Now, if $z' \in B'$ then $\mathrm{Pr}_{B'}(z') = z'$. If, on the other hand, $z' \notin B'$ then

$$\mathrm{Pr}_{B'}(z') = m\frac{z'}{\|z'\|}.$$

Finally, $\mathrm{Pr}_B(z) = \mathrm{Pr}_{B'}(z') - \frac{1}{2}(\|x_1\|^2, \ldots, \|x_n\|^2)$.

We return now to the problem which focused our interest in this chapter, that of partial transportation. We can modify Theorem 3.5, Proposition 3.6 and Theorem 3.7 to deal with the semidiscrete partial transportation problem. We start with a representation result.

Theorem 3.8. *Let P be a probability with finite support $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ where $P(x_i) = b_i > 0$ for $i = 1, \ldots, n$ and Q a Borel probability such that $Q \ll \ell^d$, then*

$$\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q) = \max_{z\geq 0}\left[\frac{-1}{n(1-\alpha)}\sum_{i=1}^n b_i z_i + \int \min_{1\leq i\leq n}\left(\|x_i - v\|^2 + z_i\right) dQ(v)\right].$$

Proof: We start from the equality

$$\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q) = \min_{P'\in\mathcal{R}_\alpha(P)} \mathcal{W}_2^2(P', Q).$$

And, given that the distribution of P is discrete, the optimization problem we are going to consider is

$$\begin{aligned} (\mathrm{P}) \ \min_{b\in C} \quad & \mathcal{W}_2^2\left(\sum_{i=1}^n b_i\delta_{x_i}, Q\right) \\ \text{s.t.} \quad & 0 \leq b_i \leq \frac{1}{n(1-\alpha)}, \ i = 1, \ldots, n \\ & \sum_{i=1}^n b_i = 1, \end{aligned}$$

where $C = \{b \in \mathbb{R}^n : \ b_i \geq 0, \sum_{i=1}^n b_i = 1\}$. The Lagrangian function associated with (P) is the function

$$L(b, z) = \mathcal{W}_2^2\left(\sum_{i=1}^n b_i\delta_{x_i}, Q\right) + \sum_{i=1}^n z_i\left(b_i - \frac{1}{n(1-\alpha)}\right).$$

So, because of the duality of the transportation problem for $b \in C$ and $z \geq 0$ we have

$$\begin{aligned} L(b,z) &= \min_{\pi \in \Pi(\sum_{i=1}^n b_i \delta_{x_i}, Q)} \int \|u-v\|^2 d\pi(u,v) + \sum_{i=1}^n z_i \left(b_i - \frac{1}{n(1-\alpha)} \right) \\ &= \sup_{\substack{\varphi \leq 0, \psi \\ \varphi(x_i)+\psi(v) \leq \|x_i - v\|^2}} \left(\sum_{i=1}^n b_i \varphi(x_i) + \int \psi dQ \right) + \sum_{i=1}^n z_i \left(b_i - \frac{1}{n(1-\alpha)} \right) \\ &= \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \sup_{\substack{\varphi \leq 0, \psi \\ \varphi(x_i)+\psi(v) \leq \|x_i - v\|^2}} \left(\sum_{i=1}^n b_i (z_i + \varphi(x_i)) + \int \psi dQ \right). \end{aligned}$$

If we take $\varphi(x_i) = -z_i$ we get

$$L(b,z) \geq \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \sup_{\substack{\psi \\ -z_i + \psi(v) \leq \|x_i - v\|^2}} \int \psi dQ.$$

Denoting $g(z) = \inf_{b \in \mathbb{R}^n} L(b,z) = \inf_{b \in C} L(b,z)$, then for $z \geq 0$,

$$\begin{aligned} g(z) &\geq \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \sup_{\substack{\psi \\ -z_i + \psi(v) \leq \|x_i - v\|^2}} \int \psi dQ \\ &= \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \int \min_{1 \leq i \leq n} \left(\|x_i - v\|^2 + z_i \right) dQ(v). \end{aligned}$$

On the other hand, for $z \geq 0$ if we denote by $\Pi_1(Q)$ to the set of probabilities in $\mathbb{R}^d$ with second marginal Q and first marginal concentrated in $\{x_1, \ldots, x_n\}$ and if we identify z with the function $z(x_i) = z_i$ then

$$\begin{aligned} g(z) &= \inf_{b \in C} L(b,z) = \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \inf_{b \in C} \left[\mathcal{W}_2^2 \left(\sum_{i=1}^n b_i \delta_{x_i}, Q \right) + \sum_{i=1}^n z_i b_i \right] \\ &= \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \inf_{\pi \in \Pi_1(Q)} \int (\|u-v\|^2 + z(u)) d\pi(u,v) \\ &= \frac{-1}{n(1-\alpha)} \sum_{i=1}^n z_i + \int \min_{1 \leq i \leq n} \left(\|x_i - v\|^2 + z_i \right) dQ(v). \end{aligned}$$

A pair $(\bar{b}, \bar{z})$ is said to be a saddle point for L if $L(\bar{b}, z) \leq L(\bar{b}, \bar{z}) \leq L(b, \bar{z})$ $\forall b, z \in \mathbb{R}^n$. Even more, $(\bar{b}, \bar{z})$ is a saddle point if and only if

$$\sup_{z \in \mathbb{R}^n} L(\bar{b}, z) = L(\bar{b}, \bar{z}) = \inf_{b \in \mathbb{R}^n} L(b, \bar{z}) = g(\bar{z}) = \max_{z \in \mathbb{R}^n} g(z). \tag{3.15}$$

Corollary 28.3 in Rockafellar (1997) implies the existence of a saddle point and by (3.15) a saddle point is a point $(\bar{b}, \bar{z})$ with $\bar{z}$ a maximizer for g, therefore we have

$$\max_{z \geq 0} g(z) = g(\bar{z}) = L(\bar{b}, \bar{z}) = \mathcal{W}_2^2 \left(\sum_{i=1}^n \bar{b}_i \delta_{x_i}, Q \right) = \mathcal{W}_2^2 \left(\mathcal{R}_\alpha(P), Q \right).$$

□

So to minimize for $P' \in \mathcal{R}_\alpha(P)$, $\mathcal{W}_2^2(P', Q)$and thus obtain $\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q)$ we must maximize the function

$$g(z) = \frac{-1}{n(1-\alpha)} \sum_{i=1}^{n} z_i + \int \min_{1\le i\le n} \left(\|x_i - v\|^2 + z_i\right) dQ(v)$$

which is a concave function. If we take $G = -g$ then

$$G(z) = \frac{1}{n(1-\alpha)} \sum_{i=1}^{n} z_i - \int \min_{1\le i\le n} \left(\|x_i - v\|^2 + z_i\right) dQ(v) \tag{3.16}$$

in a convex function and next we will see that it is also differentiable.

Proposition 3.9. *Let P be a probability with finite support $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ where $P(x_i) = b_i$ for $i = 1, \ldots, n$ and Q a Borel probability such that $Q \ll \ell^d$, then function $G(z)$ defined as in* (3.16) *is differentiable and*

$$\nabla G(z) = \frac{1}{n(1-\alpha)}(b_1, \ldots, b_n) - (Q(B_1(z)), \ldots, Q(B_n(z))), \tag{3.17}$$

where

$$B_j(z) = \{v \in \mathbb{R}^d : \|x_i - v\|^2 + z_j \le \min_{i\ne j}(\|x_i - v\|^2 + z_i)\}. \tag{3.18}$$

Proof: First let's see that the function G is differentiable. Since the first sum is linear in z and, therefore, differentiable, it is enough to focus on the second. Let

$$\tilde{G}(z) = \int \min_{1\le i\le n} \left(\|x_i - v\|^2 + z_i\right) dQ(v),$$

obviously $G(z) = \dfrac{1}{n(1-\alpha)} \sum_{i=1}^{n} z_i - \tilde{G}(z)$. Note that

$$\begin{aligned}
&\tilde{G}(z+h) - \tilde{G}(z) - \sum_{j=1}^{n} h_j Q(B_j(z)) \\
&= \sum_{j=1}^{n} \int_{B_j(z)} \left[\min_{1\le i\le n} (\|x_i - v\|^2 + (z_i + h_i)) - (\|x_j - v\|^2 + (z_j + h_j))\right] dQ(v),
\end{aligned}$$

and also

$$\begin{aligned}
0 &\ge \left[\min_{1\le i\le n} (\|x_i - v\|^2 + (z_i + h_i)) - (\|x_j - v\|^2 + (z_j + h_j))\right] I_{B_j(z)}(v) \\
&\ge \left[(\|x_j - v\|^2 + z_j) + \min_{1\le i\le n} h_i - (\|x_j - v\|^2 + (z_j + h_j))\right] I_{B_j(z)}(v) \\
&\ge \min_{1\le i\le n} h_i - h_j \ge -2 \max_{1\le i\le n} |h_j|.
\end{aligned}$$

So, when $h \to 0$,

$$\left[\min_{1\le i\le n}(\|x_i - v\|^2 + (z_i + h_i)) - (\|x_j - v\|^2 + (z_j + h_j))\right] I_{B_j(z)}(v)$$

goes to 0 (except perhaps at the border points of $B_j(z)$). Then by dominated convergence we conclude that

$$\frac{\tilde{G}(z+h) - \tilde{G}(z) - \sum_{j=1}^n h_j Q(B_j(z))}{\|h\|} \to 0$$

when $\|h\| \to 0$ this proves that $\tilde{G}$, and also G, are differentiable and that (3.17) holds. □

In a similar way to how we did in Theorem 3.5 we can write

$$\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q) = \min_{\substack{0\le b_i\le \frac{1}{n(1-\alpha)} \\ b_1+\cdots+b_n=1}} f_0(b),$$

where $f_0(b) = \sum_{i=1}^n b_i\|x_i\|^2 + \int_{\mathbb{R}^d}\|y\|^2 dQ(y) - 2\min_{z\in\mathbb{R}^n} V(z)$ with Y a random vector with distribution Q and $V(z) = \left[\sum_{i=1}^n b_i z_i + \int_{\mathbb{R}^d}\max_{1\le i\le n}(x_j\cdot y - z_j)dQ(z)\right]$.

Note that if $b_1 + \cdots + b_n \neq 1$ $f_0(b) = +\infty$ so that

$$V(z + k\mathbf{1}) = k\left(\sum_{i=1}^n b_i - 1\right) + V(z)$$

and, choosing k in a suitable way, we manage to get $V(z)$ arbitrarily small. So, if $b \ge 0$ and $b_1 + \cdots + b_n = 1$ then $f_0(b) = \mathcal{W}_2^2\left(\sum_{i=1}^n b_i\delta_{x_i}, Q\right)$. Concretely, by Theorem 3.8 we have that

$$\min_{\substack{0\le b_i\le \frac{1}{n(1-\alpha)} \\ b_1+\cdots+b_n=1}} f_0(b) = \max_{z\ge 0}\left[\frac{-1}{n(1-\alpha)}\sum_{i=1}^n b_i z_i + \int \min_{1\le i\le n}\left(\|x_i - v\|^2 + z_i\right) dQ(v)\right].$$

Let us denote

$$\begin{aligned} G(z) &:= \frac{-1}{n(1-\alpha)}\sum_{i=1}^n b_i z_i + \int \min_{1\le i\le n}\left(\|x_i - v\|^2 + z_i\right) dQ(v) \\ &= -\frac{1}{n(1-\alpha)}\sum_{i=1}^n x_i + \int_{\mathbb{R}^d}\|v\|^2 dQ(v) \\ &- 2\int_{\mathbb{R}^d}\max_{1\le i\le n}(x_i\cdot v - \frac{1}{2}(\|x_i\|^2 + z_i))dQ(v). \end{aligned} \tag{3.19}$$

As we saw in Proposition 3.9 this function is differentiable and furthermore

$$\nabla G(z) = \left[\frac{-1}{n(1-\alpha)} + Q(B_1(z)), \ldots, \frac{-1}{n(1-\alpha)} + Q(B_n(z))\right],$$

where the sets $B_j(z)$ are as in (3.18).

Theorem 3.10. *Let P be a probability with finite support $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ and Q a Borel probability such that $Q \ll \ell^d$, then $\bar{z}$ minimizes G if and only if $\bar{z}\nabla G(\bar{z}) = 0$*

Proof: Starting from equality (3.19), if $z \geq 0$, $0 \leq b_i \leq \frac{1}{n(1-\alpha)}$ for $i = 1, \ldots, n$ and $b_1 + \cdots + b_n = 1$ then

$$\begin{aligned} G(z) &\leq -\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \|v\|^2 dQ(v) - 2\int_{\mathbb{R}^d} \max_{1\leq i\leq n} (x_i \cdot v - \frac{1}{2}(\|x_i\|^2 + z_i)) dQ(v) \\ &= -\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) dQ(v). \end{aligned}$$

Let $\Pi_\alpha(P, Q)$ be the set of distributions with first marginal P' and second one Q being $P' \in \mathcal{R}_\alpha(P)$. If $\pi \in \Pi_\alpha(P, Q)$ and we define the function $z(x_j) = z_j$,

$$-\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) dQ(v) = \int_{\mathbb{R}^d \times \mathbb{R}^d} -z(u) + \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) d\pi(u, v). \tag{3.20}$$

If we take $u = x_j$ then

$$-z(u) + \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) \leq -z_j + \|x_j + v\|^2 + z_j \leq \|u - v\|^2.$$

Therefore

$$-\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) dQ(v) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|u - v\|^2 d\pi(u, v).$$

As this reasoning is valid for every π with marginals $\sum_{i=1}^{n} b_i \delta_{x_i}$ and Q we have

$$\begin{aligned} -\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \min_{1\leq i\leq n} (\|x_i - v\|^2 + z_i) dQ(v) &= \mathcal{W}_2^2 \left(\sum_{i=1}^{n} b_i \delta_{x_i}, Q \right) \\ &= \int_{\mathbb{R}^d} \|v - T(v)\|^2 dQ(v), \end{aligned}$$

where $T(v) = \nabla\varphi(v)$ and $\varphi(v) = \max_{1\leq i\leq n} (x_j \cdot v - y_j)$ with y_j is such a way that $Q(A_j(y)) = b_j$ for $A_j(y) = \{v \in \mathbb{R}^d : x_j \cdot v - y_j \geq \max_{i\neq j}(x_i \cdot v - y_i)\}$. Note that if $v \in A_j(y)$ then $\nabla\varphi(v) = x_j$.

Coming back to (3.20),

$$
\begin{aligned}
&-\sum_{i=1}^{n} b_i z_i + \int_{\mathbb{R}^d} \min_{1\le i\le n} (\|x_i - v\|^2 + z_i) dQ(v) \\
&= \int_{\mathbb{R}^d} \left[-z(\nabla\varphi(v)) + \min_{1\le i\le n} (\|x_i - v\|^2 + z_i) \right] dQ(v) \\
&\le \int_{\mathbb{R}^d} \|v - \nabla\varphi(v)\|^2 dQ(v) = \mathcal{W}_2^2 \left(\sum_{i=1}^{n} b_i \delta_{x_i}, Q \right),
\end{aligned}
$$

and equality holds if and only if

$$
\begin{aligned}
&- z(\nabla\varphi(v)) + \min_{1\le i\le n} (\|x_i - v\|^2 + z_i) = \|v - \nabla\varphi(v)\|^2 \quad Q\text{-almost sure} \\
&\Leftrightarrow -z(\nabla\varphi(v)) + \min_{1\le i\le n} (\|x_i\|^2 - 2x_i \cdot v + z_i) + \|v\|^2 = \|v\|^2 + \|\nabla\varphi(v)\|^2 - 2v \cdot \nabla\varphi(v) \\
&\Leftrightarrow -z(\nabla\varphi(v)) - 2 \max_{1\le i\le n} (x_i \cdot v - \frac{1}{2}(\|x_i\|^2 + z_i)) = \|\nabla\varphi(v)\|^2 - 2v \cdot \nabla\varphi(v) \quad Q\text{-a.s.} \\
&\Leftrightarrow Q\text{-almost sure in } A_j(z) \quad - z_j - 2 \max_{1\le i\le n} (x_i \cdot v - \frac{1}{2}(\|x_i\|^2 + z_i)) = \|x_j\|^2 - 2v \cdot x_j \\
&\Leftrightarrow v \cdot x_j - \frac{1}{2}(\|x_j\|^2 + z_j) = \max_{1\le i\le n} (x_i \cdot v - \frac{1}{2}(\|x_i\|^2 + z_i)) \\
&\Leftrightarrow \|x_j\|^2 + z_j = y_j + C, \ j = 1, \dots, n \\
&\Leftrightarrow Q(B_j(z)) = b_j.
\end{aligned}
$$

The last equivalence comes from Theorem 4.2 in del Barrio and Loubes (2017).
Hence $\bar{z} \ge 0$, $\bar{b} \ge 0$ with $\bar{b}_1 + \cdots + \bar{b}_n = 1$ and $\bar{b}_i \le \frac{1}{n(1-\alpha)}$ are optimal or, in other words, $G(\bar{z}) = f_0(\bar{b})$ if and only if

$$
\left(\bar{b}_i - \frac{1}{n(1-\alpha)} \right) \bar{z}_i = 0
$$

and besides

$$
\nabla G(\bar{z}) + \left[\frac{1}{n(1-\alpha)}, \dots, \frac{1}{n(1-\alpha)} \right] = \bar{b}.
$$

Given that by (3.3)

$$
\frac{\partial G(z)}{\partial z_j} = \left(\bar{b}_i - \frac{1}{n(1-\alpha)} \right),
$$

we get that $\bar{z} \ge 0$ is optimal if and only if $\bar{z} \nabla G(\bar{z}) = 0$.

□

From this point we could apply a gradient iteration and obtain convergence guarantees as in Theorem 3.7. We omit details.

3.4 Application to contaminated model validation

In this section we address the problem of model validation from a particular point of view: it is admitted that the proposed model is not the 'true' model, i.e. that the observations available do not come exactly from a generator within the model. It is a long-standing fact (Berkson (1938)) that a classic fit test applied on a sufficiently large sample will reject (that the generator is in the model) even in situations where the model appears to be a good description of the data. These considerations are behind the distinction between 'statistical significance' and 'practical significance' in Hodges and Lehmann (1954), where the convenience of relaxing the null hypothesis by a broader one consisting of a controlled deviation from the initial model is suggested. This deviation can be measured in many possible ways. Here we look at contamination neighbourhoods along the line in Lindsay and Liu (2009). We remember from section 2.1 that if $\mathcal{F}$ is a collection of probabilities, then the neighbourhood of α-contamination is

$$\mathcal{F}_\alpha = \{(1-\alpha)Q + \alpha R : \ Q \in \mathcal{F} \text{ and } R \text{ a probability}\}.$$

The type of validation we are considering here, which we will refer to as *essential validation*, replaces the rigid model $\mathcal{F}$ with the relaxed one $\mathcal{F}_\alpha$. We note that if α is big enough then $P \in \mathcal{F}_\alpha$ (it is enough to take $\alpha = 1$). Then we can understand that $\mathcal{F}$ is a reasonable model for the data generator if $P \in \mathcal{F}_\alpha$ for α small. Since the sets $\mathcal{F}_\alpha$ grow with α it is of interest the lower level of contamination for which $\mathcal{F}$ is an admissible model for the data, that is,

$$\alpha^* = \min\{\alpha \in [0,1] : P \in \mathcal{F}_\alpha\}.$$

In Hodges and Lehmann (1954) there are procedures to contrast the null hypothesis $\alpha \leq \alpha_0$ in a multinomial model. Here we have a similar problem with two fundamental differences. On one hand we consider simple models, $\mathcal{F} = \{Q\}$ with Q probability in $\mathbb{R}^d$ not necessarily with finite support. Another, more fundamental difference is that we will not set any threshold α_0 a priori but we will adopt a model selection viewpoint, measuring the discrepancy between P and the contaminated model $\mathcal{F}_\alpha$ by means of the distance $\mathcal{W}_2$. Since $\mathcal{W}_2(P, \mathcal{F}_\alpha)$ is a function that grows with α (just as it happens with $\mathcal{W}_2(P_n, \mathcal{F}_\alpha)$) it is not reasonable to estimate the value of the level of contamination necessary by minimizing in α, $\mathcal{W}_2(P_n, \mathcal{F}_\alpha)$. Instead, we will search appropriate penalizations to ensure that the value $\hat{\alpha}$ that minimizes $\mathcal{W}_2(P, \mathcal{F}_\alpha) + pen(\alpha)$ provides a good estimate of the level of contamination necessary to admit $\mathcal{F}$ as a model. The quality of such an estimator will be guaranteed by an oracle inequality.

The proposed criterion is applicable since, although we do not know a way to calculate $\mathcal{W}_2(P, \mathcal{F}_\alpha)$ we will dedicate the first part of this section to prove that

$$\mathcal{W}_p^p(P, \mathcal{F}_\alpha(Q)) = (1-\alpha)\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q).$$

Applying the results of section 3.3 we can calculate numerically $\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q)$.

The problem we are going to consider next is to find the best approximation to P, which we will denote for P_α in $\mathcal{F}_\alpha(Q)$. If this probability exists, it must satisfy

$$\mathcal{W}_p(P, P_\alpha) = \min_R \mathcal{W}_p(P, (1-\alpha)Q + \alpha R).$$

As

$$\inf_R \mathcal{W}_p^p(P, (1-\alpha)Q + \alpha R) = \inf_{\pi \in \Pi^{(\alpha)}(P,Q)} \int \|x-y\|^p d\pi(x,y),$$

where $\Pi^{(\alpha)}(P,Q)$ denotes the set of probabilities in $\mathbb{R}^d \times \mathbb{R}^d$ such that

$$\pi(A \times \mathbb{R}^d) = P(A) \text{ and } \pi(\mathbb{R}^d \times B) \geq (1-\alpha)Q(A)$$

for any measurable set A, B, the problem can be formulated as a problem of minimization of a linear functional over the convex set $\Pi^{(\alpha)}(P,Q)$. In the following result we will see that even though the set $\Pi^{(\alpha)}(P,Q)$ is not uniformly tight, the problem has an equivalent form in which minimization is performed on a set of probabilities in $\mathbb{R}^d \times \mathbb{R}^d$ uniformly tight.

Lemma 3.11. *For every $\pi \in \Pi^{(\alpha)}(P,Q)$ exists $\tilde{\pi} \in \Pi^{(\alpha)}(P,Q)$ such that*

$$\tilde{\pi}(\mathbb{R}^d \times B) = (1-\alpha)Q(B) + \alpha R(B) \tag{3.21}$$

for some $R \in \mathcal{R}_{1-\alpha}(P)$ and all measurable sets B and, besides,

$$\int \|x-y\|^p d\tilde{\pi}(x,y) \leq \int \|x-y\|^p d\pi(x,y).$$

Proof. We will denote by π_2 the second marginal of π and by $\pi(x|y)$ the conditional distribution of the first given the second, then

$$\pi(A \times B) = \int_B \left(\int_A 1 d\pi(x|y) \right) d\pi_2(y).$$

We will define the measures π_Q and π_P as

$$\pi_Q(A \times B) = (1-\alpha) \int_B \left(\int_A 1 d\pi(x|y) \right) dQ(y) \text{ and}$$
$$\pi_P(A \times B) = P(A \cap B) - \pi_Q((A \cap B) \times \mathbb{R}^d).$$

Note that π_P this is a positive measure since for a set C, considering that $\pi_2(B) \geq (1-\alpha)Q(B)$, we have

$$\begin{aligned}\pi_Q(A \times \mathbb{R}^d) &= (1-\alpha)\int_{\mathbb{R}^d}\left(\int_C 1 d\pi(x|y)\right) dQ(y) \\ &\leq \int_{\mathbb{R}^d}\left(\int_C 1 d\pi(x|y)\right) d\pi_2(y) = P(C).\end{aligned}$$

Moreover, π_P is concentrated in the diagonal set $\{(x,x) : x \in \mathbb{R}^d\}$.
Let us denote by $\tilde{\pi}$ the sum of these two measures

$$\tilde{\pi} = \pi_Q + \pi_P,$$

we want to see that $\tilde{\pi} \in \Pi^{(\alpha)}(P,Q)$. To this end, we have that

$$\tilde{\pi}(A \times \mathbb{R}^d) = P(A) - \pi_Q(A \times \mathbb{R}^d) + \pi_Q(A \times \mathbb{R}^d) = P(A),$$

for every measurable set A and, on the other hand,

$$\tilde{\pi}(\mathbb{R}^d \times B) = \pi_Q(\mathbb{R}^d \times B) + \pi_P(\mathbb{R}^d \times B) = (1-\alpha)Q(B) + \pi_P(\mathbb{R}^d \times B).$$

If we set $R(B) = \frac{1}{\alpha}\pi_P(\mathbb{R}^d \times B)$ we have a probability $R \in \mathcal{R}_{1-\alpha}(P)$ for which (3.21) is fulfilled.
For the second part of the lemma we see that

$$\begin{aligned}\int \|x-y\|^p d\tilde{\pi}(x,y) &= \int \|x-y\|^p d\pi_Q(x,y) \\ &= (1-\alpha)\int\left(\int \|x-y\|^p d\pi(x|y)\right) dQ(y) \\ &\leq \int\left(\int \|x-y\|^p d\pi(x|y)\right) d\pi_2(y) \\ &= \int \|x-y\|^p d\pi(x,y).\end{aligned}$$

□

As a consequence of this lemma we obtain the existence and uniqueness of the best approximation of P in $\mathcal{F}_\alpha(Q)$ as well as a characterization of it.

Corollary 3.12. *If P, Q are probabilities with finite p moment and $\alpha \in (0,1)$ then exists $P_\alpha \in \mathcal{F}_\alpha(Q)$ such that*

$$P_\alpha = (1-\alpha)Q + \alpha\tilde{P}_\alpha$$

for some $\tilde{P}_\alpha \in \mathcal{R}_{1-\alpha}(P)$ and

$$\mathcal{W}_p(P,P_\alpha) = \min_{R \in \mathcal{F}_\alpha(Q)} \mathcal{W}_p(P,R).$$

Moreover, if $p > 1$ and $P \ll \ell^d$ then that P_α (and consequently, $\tilde{P}_\alpha$) is unique.

Proof. In Lemma 3.11 we saw that

$$\inf_{R\in\mathcal{F}_\alpha(Q)} \mathcal{W}_p(P,R) = \inf_{R\in\mathcal{R}_{1-\alpha}(P)} \mathcal{W}_p(P,(1-\alpha)Q+\alpha R).$$

Let us see now that the application $R \mapsto \mathcal{W}_p(P,(1-\alpha)Q+\alpha R)$ is Lipschitz (and, therefore, continuous) with respect to $\mathcal{W}_p$. Since

$$\begin{aligned}
&|\mathcal{W}_p(P,(1-\alpha)Q+\alpha R_1) - \mathcal{W}_p(P,(1-\alpha)Q+\alpha R_2)| \\
&\leq \mathcal{W}_p((1-\alpha)Q+\alpha R_1,(1-\alpha)Q+\alpha R_2) \\
&\leq \alpha^{1/p}\mathcal{W}_p(R_1,R_2),
\end{aligned}$$

where we've used the fact that

$$\mathcal{W}_p^p((1-\alpha)Q_1+\alpha R_1,(1-\alpha)Q_2+\alpha R_2) \leq (1-\alpha)\mathcal{W}_p^p(Q_1,Q_2)+\alpha\mathcal{W}_p^p(R_1,R_2).$$

The existence of the minimizer is due to the fact that $\mathcal{R}_{1-\alpha}(P)$ is compact for the topology in $\mathcal{W}_p$ (see Proposition 2.8 in Álvarez-Esteban et al. (2011), the statement considers the case $p=2$, but the proof works the same for a general p) and the uniqueness is due to the strict convexity of $R \mapsto \mathcal{W}_p^p(P,R)$ for $p>1$ when $P \ll \ell^d$ (see, for example, the Corollary 2.10 in Álvarez-Esteban et al. (2011)). □

Let us assume now that $\hat{P}_\alpha \in \mathcal{R}_\alpha(P)$ is such that $\mathcal{W}_p(\hat{P}_\alpha,Q) = \mathcal{W}_p(\mathcal{R}_\alpha(P),Q)$ and we're going to denote by π_1 the optimal matching of $\hat{P}_\alpha$ and Q. Let us also consider the probability measure on $\mathcal{R}_{1-\alpha}(P)$, $\tilde{P}_\alpha = \frac{1}{\alpha}(P-(1-\alpha)\hat{P}_\alpha)$. We define π_2 as

$$\pi_2(A\times B) = \tilde{P}_\alpha(A\cap B),$$

which is the probability in $\mathbb{R}^d\times\mathbb{R}^d$ induced by $\tilde{P}_\alpha$ through the application $x\mapsto(x,x)$, therefore, it will be the optimal matching between $\tilde{P}_\alpha$ and itself. If we set $\pi=(1-\alpha)\pi_1+\alpha\pi_2$, then $\pi\in\Pi^{(\alpha)}(P,Q)$ and

$$\begin{aligned}
\int \|x-y\|^p d\pi(x,y) &= (1-\alpha)\int \|x-y\|^p d\pi_1(x,y) \\
&= (1-\alpha)\mathcal{W}_p^p(\hat{P}_\alpha,Q) = (1-\alpha)\mathcal{W}_p^p(\mathcal{R}_\alpha(P),Q),
\end{aligned}$$

this proves that

$$\mathcal{W}_p^p(P,\mathcal{F}_\alpha(Q)) \leq (1-\alpha)\mathcal{W}_p^p(\mathcal{R}_\alpha(P),Q). \tag{3.22}$$

Next we will see that this inequality is, in fact, an equality, but first we will introduce a technical lemma.

Lemma 3.13. *If P_1, P_2 and Q are probabilities in $\mathbb{R}^d$ with finite p moment, $p \geq 1$ then*

$$\left|\mathcal{W}_p(\mathcal{R}_\alpha(P_1), Q) - \mathcal{W}_p(\mathcal{R}_\alpha(P_2), Q)\right| \leq (1-\alpha)^{-1/p}\mathcal{W}_p(P_1, P_2).$$

Proof. Assume first that P_1 is uniformly distributed in $\{x_1, \ldots, x_n\}$ and P_2 is uniformly distributed in $\{y_1, \ldots, y_n\}$. It is well known that

$$\mathcal{W}_p^p(P_1, P_2) = \frac{1}{n}\sum_{i=1}^n \|x_i - y_{\sigma(i)}\|^p,$$

where σ is the permutation of $\{1, \ldots, n\}$ that minimizes $\sum_{i=1}^n \|x_i - y_{\sigma(i)}\|^p$. For $b = (b_1, \ldots, b_n)$ such that $0 \leq b_i \leq \frac{1}{n(1-\alpha)}$ and $b_1 + \cdots + b_n = 1$ we write $P_{1,b} = \sum_{i=1}^n b_i \delta_{x_i}$ and $P_{2,b} = \sum_{i=1}^n b_i \delta_{y_{\sigma(i)}}$. Note that $\mathcal{R}_\alpha(P_i)$ is the set of all probabilities $P_{i,b}$ of that type. Now

$$\begin{aligned} |\mathcal{W}_p(P_{1,b}, Q) - \mathcal{W}_p(P_{2,b}, Q)|^p &\leq \mathcal{W}_p^p(P_{1,b}, P_{2,b}) \leq \sum_{i=1}^n b_i \|x_i - y_{\sigma(i)}\|^p \\ &\leq \frac{1}{n(1-\alpha)}\sum_{i=1}^n \|x_i - y_{\sigma(i)}\|^p = \frac{1}{1-\alpha}\mathcal{W}_p^p(P_1, P_2). \end{aligned}$$

This implies that

$$\left|\mathcal{W}_p(\mathcal{R}_\alpha(P_1), Q) - \mathcal{W}_p(\mathcal{R}_\alpha(P_2), Q)\right| \leq (1-\alpha)^{-1/p}\mathcal{W}_p(P_1, P_2).$$

For general P_1, P_2 we consider $P_{1,n}$, $P_{2,n}$ of the above type so that $\mathcal{W}_p(P_{i,n}, P_i) \to 0$ (we can take, for example, empirical measures in i.i.d. samples of P_i). Then, $\mathcal{W}_p(\mathcal{R}_\alpha(P_{i,n}), Q) \to \mathcal{W}_p(\mathcal{R}_\alpha(P_i), Q)$, $i = 1, 2$. This is deduced, for example, from Theorem 2.13 in Álvarez-Esteban et al. (2011) (again, the article considers the case $p = 2$ but the proof works for any p). Now we have

$$\left|\mathcal{W}_p(\mathcal{R}_\alpha(P_{1,n}), Q) - \mathcal{W}_p(\mathcal{R}_\alpha(P_{2,n}), Q)\right| \leq (1-\alpha)^{-1/p}\mathcal{W}_p(P_{1,n}, P_{2,2}).$$

By taking $n \to \infty$ we get the result. □

Proposition 3.14. *With the previous notation*

$$\mathcal{W}_p^p(P, \mathcal{F}_\alpha(Q)) = (1-\alpha)\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q). \tag{3.23}$$

Furthermore, if $\hat{P}_\alpha \in \mathcal{R}_\alpha(P)$ is such that $\mathcal{W}_p(\hat{P}_\alpha, Q) = \mathcal{W}_p(\mathcal{R}_\alpha(P), Q)$ then, fixing $\tilde{P}_\alpha = \frac{1}{\alpha}(P - (1-\alpha)\hat{P}_\alpha)$ we have that $P_\alpha = (1-\alpha)Q + \alpha\tilde{P}_\alpha \in \mathcal{F}(Q)$ and

$$\mathcal{W}_p(P, P_\alpha) = \mathcal{W}_p(P, \mathcal{F}_\alpha(P)).$$

In particular, if $p > 1$ and $P, Q \ll \ell_d$ then the unique probability $P_\alpha \in \mathcal{F}_\alpha(Q)$ such that $\mathcal{W}^p(P, \mathcal{F}_\alpha(Q)) = \mathcal{W}^p(P, P_\alpha)$ is $P_\alpha = (1-\alpha)Q + \alpha\tilde{P}_\alpha$ with $\tilde{P}_\alpha = \frac{1}{\alpha}(P - (1-\alpha)\hat{P}_\alpha)$ and $\hat{P}_\alpha$ is the unique probability in $\mathcal{R}_\alpha(P)$ such that $\mathcal{W}(\hat{P}_\alpha, Q) = \mathcal{W}_p(\mathcal{R}_\alpha(P), Q)$.

Proof. We will start by proving (3.23) for the case where the common support of P and Q is the finite set $\{x_1, \ldots, x_n\}$ with $P\{x_i\} = p_i > 0$ and $Q\{x_i\} = q_i > 0$, $i = 1, \ldots, n$. Note that in this case

$$\mathcal{W}_p^p(P, \mathcal{F}_\alpha(Q)) = \min_{\pi \in C} \sum_{1 \le i,j \le n} c_{i,j}\pi_{i,j},$$

where $c_{i,j} = \|x_i - x_j\|^p$ and C is the set of points $\pi = (\pi_{i,j})_{1 \le i,j \le n}$ such that $\pi_{i,j} \ge 0$, $1 \le i, j \le n$ and $\sum_{j=1}^n \pi_{i,j} = p_i$, $\sum_{j=1}^n \pi_{j,i} \ge (1-\alpha)q_i$, $i = 1, \ldots, n$. By duality in linear programming (see, for example, Corollary 28.3.1 and Theorem 28.4 in Rockafellar (1997)) we have

$$\begin{aligned} \mathcal{W}_p^p(P, \mathcal{F}_\alpha(Q)) &= \max_{(\varphi,\psi) \in D} \Big[\sum_{i=1}^n p_i\varphi_i + (1-\alpha) \sum_{j=1} q_j\psi_j \Big] \\ &= (1-\alpha) \max_{(\varphi,\psi) \in D} \Big[\tfrac{1}{1-\alpha} \sum_{i=1}^n p_i\varphi_i + \sum_{j=1} q_j\psi_j \Big], \end{aligned} \tag{3.24}$$

where D is the set of pairs $(\varphi, \psi) = ((\varphi_i)_{1 \le i \le n}, (\psi_j)_{1 \le j \le n})$ such that $\psi_j \ge 0$, $j = 1, \ldots, n$ and $\varphi_i + \psi_j \le c_{i,j}$, $1 \le i, j \le n$. On the other hand,

$$\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q) = \min_{\pi \in \tilde{C}} \sum_{1 \le i,j \le n} c_{i,j}\pi_{i,j},$$

with $c_{i,j} = \|x_i - x_j\|^p$ as before and $\tilde{C}$ the set of points $\pi = (\pi_{i,j})_{1 \le i,j \le n}$ such that $\pi_{i,j} \ge 0$, $1 \le i, j \le n$ and $\sum_{j=1}^n \pi_{i,j} \le \frac{1}{1-\alpha} p_i$, $\sum_{j=1}^n \pi_{j,i} = (1-\alpha)q_i$, $i = 1, \ldots, n$. Duality leads us to

$$\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q) = \max_{(\varphi,\psi) \in \tilde{D}} \Big[\tfrac{1}{1-\alpha} \sum_{i=1}^n p_i\varphi_i + \sum_{j=1} q_j\psi_j \Big], \tag{3.25}$$

where $\tilde{D}$ is the set of points $((\varphi_i)_{1 \le i \le n}, (\psi_j)_{1 \le j \le n})$ such that $\varphi_i \le 0$, $1 \le i \le n$ and $\varphi_i + \psi_j \le c_{i,j}$, $1 \le i, j \le n$. Then, it is enough to show that the optimal values in the maximization problems in (3.24) and (3.25) are the same. Now, note that for every $((\varphi_i)_{1 \le i \le n}, (\psi_j)_{1 \le j \le n}) \in D$ we have $\varphi_i + \psi_i \le c_{i,i} = 0$ and, hence, $\varphi_i \le -\psi_i \le 0$. This proves that $D \subset \tilde{D}$ and, as a consequence,

$$\max_{(\varphi,\psi) \in D} \Big[\tfrac{1}{1-\alpha} \sum_{i=1}^n p_i\varphi_i + \sum_{j=1} q_j\psi_j \Big] \le \max_{(\varphi,\psi) \in \tilde{D}} \Big[\tfrac{1}{1-\alpha} \sum_{i=1}^n p_i\varphi_i + \sum_{j=1} q_j\psi_j \Big].$$

Now, if $((\varphi_i)_{1 \le i \le n}, (\psi_j)_{1 \le j \le n}) \in \tilde{D}$ we have

$$\psi_j \le \min_{1 \le i \le n} (c_{i,j} - \varphi_i) := \tilde{\psi}_j. \tag{3.26}$$

Note that $\varphi_i + \tilde{\psi}_j \leq c_{i,j}$ and also that, as $\varphi_i \leq 0$, we have $\tilde{\psi}_j \geq \min_{1\leq i\leq n} c_{i,j} = 0$, which means that $((\varphi_i)_{1\leq i\leq n}, (\tilde{\psi}_j)_{1\leq j\leq n}) \in D$. But from the inequality in (3.26) we see that

$$\frac{1}{1-\alpha}\sum_{i=1}^{n} p_i\varphi_i + \sum_{j=1} q_j\psi_j \leq \frac{1}{1-\alpha}\sum_{i=1}^{n} p_i\varphi_i + \sum_{j=1} q_j\tilde{\psi}_j.$$

This implies that

$$\max_{(\varphi,\psi)\in\tilde{D}} \left[\tfrac{1}{1-\alpha}\sum_{i=1}^{n} p_i\varphi_i + \sum_{j=1} q_j\psi_j\right] \leq \max_{(\varphi,\psi)\in D} \left[\tfrac{1}{1-\alpha}\sum_{i=1}^{n} p_i\varphi_i + \sum_{j=1} q_j\psi_j\right]$$

and proves (3.23) in this particular case.

Generalizing for any P, Q we set $\varepsilon > 0$, there is P_ε, Q_ε, with finite common support in such a way that $\mathcal{W}_p(P, P_\varepsilon) < \varepsilon$ and $\mathcal{W}_p(Q, Q_\varepsilon) < \varepsilon$. For these P_ε, Q_ε we know that

$$\mathcal{W}_p(P_\varepsilon, \mathcal{F}_\alpha(Q_\varepsilon)) = (1-\alpha)^{1/p}\mathcal{W}_p(\mathcal{R}_\alpha(P_\varepsilon), Q_\varepsilon). \tag{3.27}$$

Since $\mathcal{W}_p$ is a metric we have that

$$|\mathcal{W}_p(P_\varepsilon, \mathcal{F}_\alpha(Q_\varepsilon)) - \mathcal{W}_p(P, \mathcal{F}_\alpha(Q_\varepsilon))| \leq \mathcal{W}_p(P, P_\varepsilon) < \varepsilon$$

and, for any measure of probability R with finite p moment,

$$\mathcal{W}_p((1-\alpha)Q + \alpha R, (1-\alpha)Q_\varepsilon + \alpha R) \leq (1-\alpha)^{1/p}\mathcal{W}_p(Q, Q_\varepsilon) < (1-\alpha)^{1/p}\varepsilon.$$

Hence,

$$|\mathcal{W}_p(P, \mathcal{F}_\alpha(Q_\varepsilon)) - \mathcal{W}_p(P, \mathcal{F}_\alpha(Q))| \leq (1-\alpha)^{1/p}\varepsilon.$$

By combining these two things we obtain

$$|\mathcal{W}_p(P_\varepsilon, \mathcal{F}_\alpha(Q_\varepsilon)) - \mathcal{W}_p(P, \mathcal{F}_\alpha(Q))| \leq ((1-\alpha)^{1/p} + 1)\varepsilon.$$

Similarly, from triangular inequality for $\mathcal{W}_p$ we see that

$$|\mathcal{W}_p(\mathcal{R}_\alpha(P_\varepsilon), Q_\varepsilon) - \mathcal{W}_p(\mathcal{R}_\alpha(P_\varepsilon), Q)| \leq \varepsilon.$$

By Lemma 3.13 we have

$$|\mathcal{W}_p(\mathcal{R}_\alpha(P_\varepsilon), Q) - \mathcal{W}_p(\mathcal{R}_\alpha(P), Q)| \leq (1-\alpha)^{-1/p}\varepsilon,$$

and, as a consequence,

$$|\mathcal{W}_p(\mathcal{R}_\alpha(P_\varepsilon), Q_\varepsilon) - \mathcal{W}_p(\mathcal{R}_\alpha(P), Q)| \leq (1 + (1-\alpha)^{-1/p})\varepsilon.$$

By making ε tend to 0 we get from (3.27) that (3.23) is met for any P, Q.

Finally, if $\pi = (1-\alpha)\pi_1 + \alpha\pi_2$ is the transportation plan introduced in the discussion that led us to t(3.22), then (3.23) implies that π is optimal, or what is the same, that

$$\mathcal{W}_p^p(P, \mathcal{F}_\alpha(Q)) = (1-\alpha)\mathcal{W}_p^p(\mathcal{R}_\alpha(P), Q).$$

In particular, π is an optimal pairing of P with its second marginal which is $P_\alpha = (1-\alpha)Q + \alpha\tilde{P}_\alpha$ and P_α is the best approximation of P by probabilities in $\mathcal{F}_\alpha(Q)$. □

To conclude this section we are going to obtain an oracle inequality that guarantees the quality of the estimator of the optimal trimming level $\hat{\alpha}$ that minimizes $\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q) + pen(\alpha)$.

Firstly, we study the relationship between the expected value of the distance for the set of trimmings of the empirical distribution and the distance for the set of trimmings of theoretical distribution. After this lemma we will obtain the oracle inequality.

Lemma 3.15. *Let P, P_n and Q be probabilities with finite second moment, then*

$$\begin{aligned} 0 &\leq E\left(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q)\right) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q) \\ &\leq \frac{1}{1-\alpha} E\left(\mathcal{W}_2^2(P_n, P)\right) + 2\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q)\frac{1}{\sqrt{1-\alpha}}\sqrt{E\left(\mathcal{W}_2^2(P_n, P)\right)} \end{aligned} \quad (3.28)$$

Proof. For the first inequality we will use duality

$$\begin{aligned} E\left(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q)\right) &= E\left(\sup_{\substack{\varphi\leq 0 \\ \varphi(x)+\psi(y)\leq\|x-y\|^2}} \left[\frac{1}{1-\alpha}\int \varphi dP_n + \int \psi dQ\right]\right) \\ &\geq \sup_{\substack{\varphi\leq 0 \\ \varphi(x)+\psi(y)\leq\|x-y\|^2}} \left[E\left(\frac{1}{1-\alpha}\int \varphi dP_n + \int \psi dQ\right)\right] \\ &= \sup_{\substack{\varphi\leq 0 \\ \varphi(x)+\psi(y)\leq\|x-y\|^2}} \left[\frac{1}{1-\alpha}\int \varphi dP + \int \psi dQ\right] \\ &= \mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q). \end{aligned}$$

For the second inequality we will use Lemma 3.13,

$$
\begin{aligned}
&E\left(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)-\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q)\right)\\
&=E\left[\left(\mathcal{W}_2(\mathcal{R}_\alpha(P_n),Q)-\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\right)\left(\mathcal{W}_2(\mathcal{R}_\alpha(P_n),Q)+\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\right)\right]\\
&=E\left[\left(\mathcal{W}_2(\mathcal{R}_\alpha(P_n),Q)-\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\right)^2\right]\\
&+2\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q)E\left[\left(\mathcal{W}_2(\mathcal{R}_\alpha(P_n),Q)-\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\right)\right]\\
&\leq\frac{1}{1-\alpha}E\left(\mathcal{W}_2^2(P_n,P)\right)+2\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q)\frac{1}{\sqrt{1-\alpha}}E\left(\mathcal{W}_2(P_n,P)\right).
\end{aligned}
$$

Using that $E\left(\mathcal{W}_2(P_n,P)\right)\leq\sqrt{E\left(\mathcal{W}_2^2(P_n,P)\right)}$ we get (3.28). □

Theorem 3.16. *Let P and Q be two distributions with moment of order two finite and Q with compact support. Let $X_1,\ldots,X_n$ be i.i.d. random vectors of dimension d with distribution P and let P_n be the sample distribution in such a way that $E\|X_1\|^4<\infty$. Assume that $d>4$ and that $\exists M\in\mathbb{R}$ such that for a given $q>\frac{2d}{d-2}$, $\int_{\mathbb{R}^d}|v|^qP(dv)\leq M$. Let $A=\{0,\alpha_1,\ldots,\alpha_{max}\}$ be the set of possible trimming levels. If we consider the penalization function*

$$
pen(\alpha)=\sqrt{\frac{2C(P,Q)ln(n)}{n(1-\alpha)^2}},\tag{3.29}
$$

and define

$$
\hat{\alpha}=\arg\min_{\alpha\in A}(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)+pen(\alpha)),\tag{3.30}
$$

then for two constants $C(d,q)$ and $C(P,Q)$ the following bound holds

$$
\begin{aligned}
E\left(\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P),Q)\right)\quad\leq\quad&\inf_{\alpha\in A}\left\{\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q)+pen(\alpha)+\frac{C(d,q)M^{\frac{2}{q}}}{n^{\frac{2}{d}}(1-\alpha)}\right.\\
+\quad&\left.2\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\frac{C(d,q)^{\frac{1}{2}}M^{\frac{1}{q}}}{n^{\frac{1}{d}}\sqrt{1-\alpha}}\right\}+\sqrt{\frac{12\pi C(P,Q)}{n(1-\alpha_{max})^2}}.
\end{aligned}\tag{3.31}
$$

Proof. We start from the basic inequality

$$
\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P_n),Q)+pen(\hat{\alpha})\leq\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)+pen(\alpha)\ \ \forall\alpha\in A,
$$

which is satisfied by the definition of $\hat{\alpha}$. By adding and subtracting the distances for the theoretical distribution we arrive at

$$
\begin{aligned}
\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P),Q)\quad\leq\quad&\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q)+pen(\alpha)\\
+\quad&\left(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)-\mathcal{R}_\alpha(P),Q)\right)-pen(\hat{\alpha})\qquad(3.32)\\
+\quad&\left(\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P),Q)-\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P_n),Q)\right)\qquad(3.33)
\end{aligned}
$$

Let us bound the two parentheses of the equation above. First we will study the parentheses in (3.32) which can be written as

$$[\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q) - E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q))] \tag{3.34}$$

$$+ [E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) - \mathcal{R}_\alpha(P),Q)]. \tag{3.35}$$

Expression (3.34) is bounded by Theorem 3.9 in del Barrio and Matrán (2013) taking $t = \sqrt{\frac{z4C(P,Q)}{n(1-\alpha)^2}}$, obtaining

$$P\left(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q) - E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) \geq \sqrt{\frac{z4C(P,Q)}{n(1-\alpha)^2}}\right) \leq e^{-z}. \tag{3.36}$$

To bound (3.35) we are going to use (3.28) combined with Theorem 1 in Fournier and Guillin (2015) which leads us to

$$E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q) \leq \frac{C(d,q)M^{\frac{2}{q}}}{n^{\frac{2}{d}}(1-\alpha)} + 2\mathcal{W}_2(\mathcal{R}_\alpha(P),Q)\frac{C(d,q)^{\frac{1}{2}}M^{\frac{1}{q}}}{n^{\frac{1}{d}}\sqrt{1-\alpha}}. \tag{3.37}$$

On the other side we are going to bound (3.33). To this end we will use the first inequality from Lemma 3.15 and, again, Theorem 3.9 in del Barrio and Matrán (2013) this time with $t = \sqrt{\frac{(z+\ln(n))4C(P,Q))}{n(1-\hat{\alpha})^2}}$:

$$\begin{aligned}
&\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P),Q) - \mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P_n),Q) \\
&\leq \sup_{\alpha\in A}(\mathcal{W}_2^2(\mathcal{R}_\alpha(P),Q) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) \\
&\leq \sup_{\alpha\in A}(E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q))
\end{aligned}$$

Given that by Theorem 3.9 in del Barrio and Matrán (2013) we have that

$$P\left(E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q) \geq \sqrt{\frac{(z+\ln(n))4C(P,Q))}{n(1-\hat{\alpha})^2}}\right) \leq \frac{1}{n}e^{-z},$$

we get

$$\begin{aligned}
&P\left(\sup_{\alpha\in A}(E(\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}_\alpha(P_n),Q)) \geq \sqrt{\frac{(z+\ln(n))4C(P,Q))}{n(1-\hat{\alpha})^2}}\right) \\
&\leq \sum_{\alpha'\in A} P\left(E(\mathcal{W}_2^2(\mathcal{R}'_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}'_\alpha(P_n),Q) \geq \sqrt{\frac{(z+\ln(n))4C(P,Q))}{n(1-\hat{\alpha})^2}}, \alpha' = \hat{\alpha}\right) \\
&\leq \sum_{\alpha'\in A} P\left(E(\mathcal{W}_2^2(\mathcal{R}'_\alpha(P_n),Q)) - \mathcal{W}_2^2(\mathcal{R}'_\alpha(P_n),Q) \geq \sqrt{\frac{(z+\ln(n))4C(P,Q))}{n(1-\hat{\alpha})^2}}\right) \\
&\leq n\frac{1}{n}e^{-z}.
\end{aligned}$$

The with probability at least $1 - e^{-z}$

$$\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P), Q) - \mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P_n), Q) \leq \sqrt{\frac{2C(P,Q)ln(n)}{n(1-\hat{\alpha})^2}} + \sqrt{\frac{2C(P,Q)z}{n(1-\hat{\alpha})^2}}. \tag{3.38}$$

By joining (3.36), (3.37) and (3.38) with the basic inequality of the beginning we get that with probability at most $1 - 2e^{-z}$

$$\begin{aligned}
\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P), Q) &\leq \mathcal{W}_2^2(\mathcal{R}_{\alpha}(P), Q) + pen(\alpha) + \sqrt{\frac{z4C(P,Q)}{n(1-\alpha)^2}} + \frac{C(d,q)M^{\frac{2}{q}}}{n^{\frac{2}{d}}(1-\alpha)} \\
&+ 2\mathcal{W}_2(\mathcal{R}_{\alpha}(P), Q)\frac{C(d,q)^{\frac{1}{2}}M^{\frac{1}{q}}}{n^{\frac{1}{d}}\sqrt{1-\alpha}} - pen(\hat{\alpha}) + \sqrt{\frac{2C(P,Q)ln(n)}{n(1-\hat{\alpha})^2}} \\
&+ \sqrt{\frac{2C(P,Q)z}{n(1-\hat{\alpha})^2}}.
\end{aligned}$$

Taking as a penalization

$$pen(\alpha) = \sqrt{\frac{2C(P,Q)ln(n)}{n(1-\alpha)^2}},$$

we come to

$$\begin{aligned}
\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P), Q) &\leq \mathcal{W}_2^2(\mathcal{R}_{\alpha}(P), Q) + pen(\alpha) + \sqrt{\frac{z4C(P,Q)}{n(1-\alpha)^2}} + \frac{C(d,q)M^{\frac{2}{q}}}{n^{\frac{2}{d}}(1-\alpha)} \\
&+ 2\mathcal{W}_2(\mathcal{R}_{\alpha}(P), Q)\frac{C(d,q)^{\frac{1}{2}}M^{\frac{1}{q}}}{n^{\frac{1}{d}}\sqrt{1-\alpha}} + \sqrt{\frac{2C(P,Q)z}{n(1-\hat{\alpha})^2}}.
\end{aligned}$$

By bounding α and $\hat{\alpha}$ by α_{max} and integrating with respect to z we get that $\forall \alpha \in A$,

$$\begin{aligned}
E(\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P), Q)) &\leq \mathcal{W}_2^2(\mathcal{R}_{\alpha}(P), Q) + pen(\alpha) + \frac{C(d,q)M^{\frac{2}{q}}}{n^{\frac{2}{d}}(1-\alpha)} \\
&+ 2\mathcal{W}_2(\mathcal{R}_{\alpha}(P), Q)\frac{C(d,q)^{\frac{1}{2}}M^{\frac{1}{q}}}{n^{\frac{1}{d}}\sqrt{1-\alpha}} + \sqrt{\frac{12\pi C(P,Q)}{n(1-\alpha_{max})^2}},
\end{aligned}$$

and taking the infimum in the set A we come to (3.31). □

We see in the oracle inequality that if Q is an α-contaminated version of P, then $E(\mathcal{W}_2^2(\mathcal{R}_{\hat{\alpha}}(P), Q))$ will tend to 0 when the sample size increases. However, although this convergence is guaranteed for very large sample sizes, this convergence quickly slows down when we work with large dimensions as we have a convergence rate of the order of $\frac{1}{n^{\frac{1}{d}}}$.

With this theorem it is proven that $\hat{\alpha}$ defined as in (3.30) is a good estimator of optimal trimming when the sample is large enough. To calculate $\hat{\alpha}$ we can apply a gradient descent method that we can obtain from the results of Proposition 3.9 and Theorem 3.10. This

method has many computational limitations due to the difficulty of explicit calculation of the gradient. This gradient will be approached by a quasi-Monte Carlo method. The immense number of calculations required for gradient estimation and calculation of the optimum by a gradient descent method makes it impossible to calculate the distance for large samples (over 500 elements).

Because of all this and the fact that the function $\frac{1}{1-\alpha}$ is strictly convex for α, we believe that choosing a penalization that depends not only on α but also on the data could lead to better rates in the oracle inequality and possibly make the calculation of the optimal trimming more computationally affordable. But this escapes from the objective of this work and we leave it as a possible line to follow in the future.

Note that the restrictions over d and q on Theorem 3.16 have been imposed for simplicity of the result, but could be relaxed leading to a result that differs from the current one in a term of size at most logarithmic using Theorem 1 in Fournier and Guillin (2015) (the same result used in the proof of the theorem). Therefore $\hat{\alpha}$ will be a good estimator of the optimal trimming for any dimension.

3.5 Algorithm and simulations

To finish the chapter we will see an example in which we apply the results obtained to estimate Wasserstein distance between the set of trimmings of a distribution P and a distribution Q. In the example we are going to consider both distributions will be normal bivariate centered on the origin and with covariance matrix the identity. To apply the results obtained in sections 3.3 and 3.4 instead of working with the theoretical distribution P we will work with its empirical version P_n. In addition to estimating the distance we will also estimate the optimal trimming level, so we have introduced a 10 percent contamination to the P_n distribution by substituting 10 percent of the sample elements with others that come from a normal bivariate with matrix of co-variances identity as before, but now centered on $(4, 4)^T$.

In Theorem 3.8 we proved that calculating $\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q)$ was equivalent to minimizing the function $G(z)$ defined in (3.16). In addition, in Proposition 3.9 we had obtained an expression for the gradient of this function, this allows us to use a gradient descent algorithm for its minimization. Because minimization in (3.16) has a restriction, the usual stopping criterion (gradient equal to 0) is not valid, but in Theorem 3.10 we obtained a new criterion for our problem that consists of minimizing the product of vector z with the value of the gradient at that point.

Although we have an explicit expression for the calculation of the gradient of $G(z)$, this gradient cannot be calculated in practice, so we will have to settle for an approximation of this gradient. For this it we will use a quasi-Monte Carlo method. Likewise, we cannot calculate the value of $G(z)$ exactly and we will use a similar technique to approximate it. It should be highlighted that these calculations are very expensive computationally speaking which limits us a lot when choosing the sample size of P_n that we will consider. In this case we have taken $n = 100$. The set of possible trimming levels is $A = \{0, 0.02, \ldots, 0.14, 0.16\}$.

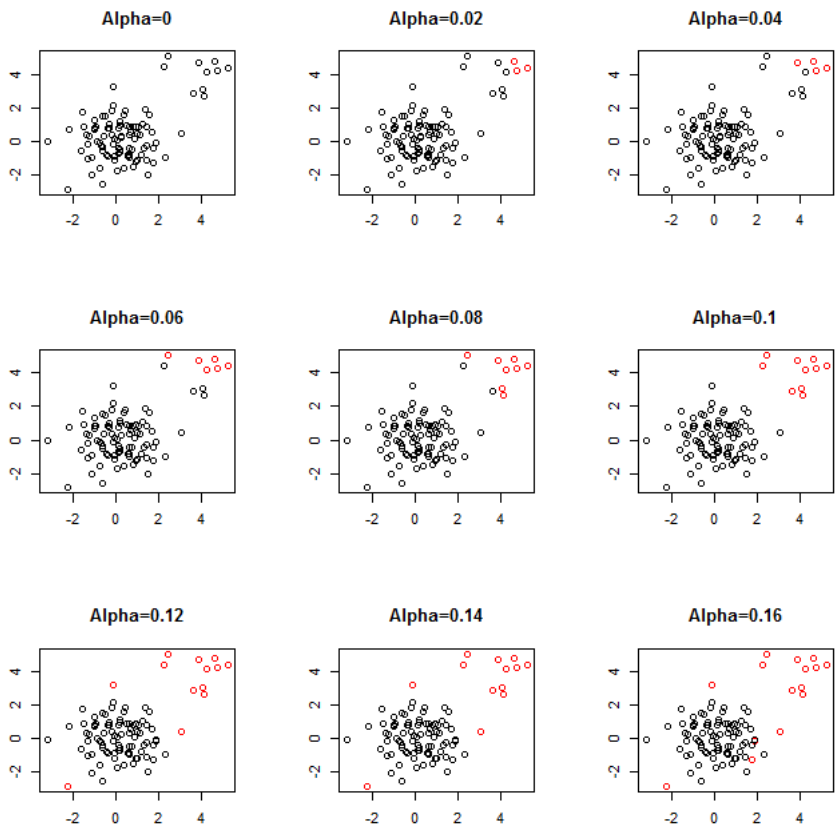


Figure 3.3: Plots of the sample's trimmed points for each trimming level in A.

In Figure 3.3 we have represented P_n and the points that the algorithm decides to trim for each of the values in A. Before viewing the graphs, it is important to note that both the gradient value and the function value calculations are approximate, so that the stop criterion cannot be expected to be met exactly. This added to the fact that convergence is slow and costly makes the value of the weights to be approximate. In addition, convergence is much slower for smaller α values so the approximation to optimal weights is worse than

for higher values. For the sake of simplicity, when we talk about weights we are referring to a scaling of the value of each of them between 0 and 1, obtained by multiplying the real value of the weight by the factor $n(1-\alpha)$.

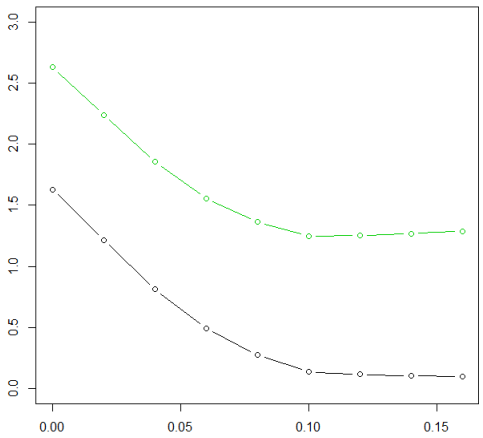


Figure 3.4: Values of $\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q)$ and of $\mathcal{W}_2^2(\mathcal{R}_\alpha(P_n), Q) + pen(\alpha)$.

In red are displayed in Figure 3.3 the weights with a value of less than 0.5, in black are the rest. To analyze the results obtained, we will separate the levels of trimming in 3 groups in which the behavior has been the same. First we will treat the result for the trimming levels 0 and 0.02, then from 0.04 to 0.1 and finally from 0.12 to 0.16. In the first case the convergence of the method has not been good in spite of having performed many more iterations of the gradient method than for the rest of the cases, so we may still be far from the optimal, this is reflected in a bad estimation of the points to be trimmed for level 0.02. In the second group we see that the algorithm perfectly estimates the number of points to be trimmed, it can be seen in the weights that have gone from a value of 0 or very close to 0 to values around 1. In the last group it seems that the estimation of the points to be trimmed is not adequate because we have represented the points with weights above 0.5 as not trimmed and the points with weights below as trimmed. Seeing the values of the weights it can be seen that the 10 contaminated points have a weight of 0 and, therefore, the algorithm considers that they should be trimmed. But there is also a group of about 10 points that have weights surrounding 0.5 and therefore they should be partially trimmed.

In Figure 3.4 we see in black the value of the distance between the trimming set of P_n and Q for each of the values in A. In green we see the value of the distance penalized

by (3.29). Although we consider that the best way to estimate the constant $C(P,Q)$ is through cross-validation, for this case it is not computationally efficient and we have taken $C(P,Q) = \frac{n}{2\ln(n)}$, which means that $pen(\alpha) = \frac{1}{1-\alpha}$. We see that although the distance values are decreasing with the level of trimming, when the penalization is added this function decreases to a level of trimming of 0.1 and grows from there on. In other words, our algorithm selects 0.1 as the optimal trimming level that coincides with the contamination level we had included in the sample. Therefore, we can conclude that our method correctly selects the optimal trimming level.

Chapter 4

Deformation models. Applications to signal alignment

This chapter deals with the problem of aligning deformed signals according to the model described in section 2.4. Alignment problems find applications in functional data analysis (for example in Bercu and Fraysse (2012), Dupuy et al. (2011), Gamboa et al. (2007) or Ramsay and Silverman (2005)), image analysis (see Amit et al. (1991) or Trouve and Younes (2005)), etc. The deformation model discussed in this chapter assumes that different observed samples $X_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, J$ (where j indicates the sample and i refers to the different observations within the sample), come, essentially, from the same signal, i. e., from the same random generator, μ, but that in each sample the signal has suffered a certain distortion, so that the observations in the jth sample actually come from μ_j, where μ_j is a distorted version of μ. The alignment problem, then, is to look for appropriate transformations that will unwarp the original deformations in order to make the transformed samples as similar as possible.

This chapter is organized as follows. Section 4.1 describes more precisely the deformation model considered. A procedure for estimating deformations and establishing asymptotic properties is proposed in section 4.2. Section 4.3 deals with computational aspects related to the criterion used. Some examples of deformation models that fit the general framework described in section 4.1 are shown in section 4.4. Finally, a simulation study illustrating the behaviour of the proposed alignment procedure is included in section 4.5.

The material included in this chapter has been published in Agulló-Antolín et al. (2015).

4.1 A model for distribution deformation

To describe the deformation model considered in this work we will begin by assuming that ε_{ij}, $1 \leq i \leq n$, $1 \leq j \leq J$, are random i.i.d. variables with unknown distribution μ defined in I_a a subset of the separable and complete metric space $(\mathcal{X}, d)$. We will assume that we observe only some deformations of these observations, more specifically, we will consider a family of invertible deformation functions indexed by parameters λ that deform a point x into another point $\varphi_\lambda(x)$. The shape of the deformation will be modeled by the known function φ while the amount of deformation is characterized by the parameter $\lambda \in \mathbb{R}^d$ with $d > 0$. Specifically

$$\begin{array}{rcl} \varphi : \Lambda \times I_a & \to & I_b \\ (\lambda, x) & \mapsto & \varphi_\lambda(x) \end{array}$$

for Λ an open subset of $\mathbb{R}^d$ and I_a, I_b subsets of $\mathcal{X}$ that can be unbounded. As we said before, we consider the model

$$X_{ij} = \varphi_{\theta_j^*}(\varepsilon_{ij}) \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant J, \tag{4.1}$$

where θ_j^* is the unknown deformation parameter in $\Lambda \in \mathbb{R}^d$ associated with the jth sample $(X_{1j}, \ldots, X_{nj})$.

Our objective is to estimate the parameter $\theta^* \in \prod_{j=1}^{J} \Lambda$. To this end, we will study a criteria based on an alignment process for the distribution μ_j of each i.i.d. sample $(X_{1j}, \ldots, X_{nj})$ for all $j = 1, \ldots, J$.

To recover the deformation parameter θ^* we will minimize the energy needed to align the distributions μ_j. A natural distance for measuring the cost of aligning two distributions is the 2-Wasserstein distance defined in (2.9).

The proposed procedure for aligning the law of observations X_j is as follows:

- For a parameter θ, we compute the image of the observation using the reverse operator to undo the deformation of the observations. In particular, for all candidates $\theta = (\theta_1, \ldots, \theta_J)$ and for every observation X_{ij}, we apply the inverse deformation with parameter θ_j, which is equivalent to computing the following variables

$$Z_{ij}(\theta) = \varphi_{\theta_j}^{-1}(X_{ij}), \ i = 1, \ldots, n \text{ and } j = 1, \ldots, J.$$

- We take the parameter among the chosen ones that minimizes the energy (measured by 2-Wasserstein distance) needed to align the distribution of the variables once the deformation has been undone with the distribution of its mean.

We specify the details below. If we denote by $\mu_j(\theta)$ the common distribution of the i.i.d. sample $(Z_{1j}(\theta), \ldots, Z_{nj}(\theta))$, then $\mu_j(\theta) = \mu_j \circ \varphi_{\theta_j}$. We are going to set $\mu(\theta) = \frac{1}{J}\sum_{j=1}^{J} \mu_j(\theta)$, which is the mean distribution of the $\mu_j(\theta)$'s.

By means of the following quantity we will quantify the cost of aligning the $\mu_j(\theta)$'s with the mean distribution:

$$M(\theta) = \sum_{j=1}^{J} \mathcal{W}_2^2(\mu_j(\theta), \mu(\theta)), \text{ with } \theta \in \prod_{j=1}^{J} \Lambda. \tag{4.2}$$

It is worth remembering that as for every j, $\mu_j(\theta^\star) = \mu$, we have $M(\theta^\star) = 0$. This function provides a characterization of our parameter of interest θ^*. Often equality $M(\theta^*) = 0$ is not enough to characterize θ^*.

For example, we can look at the case of random variables with values in $\mathbb{R}^d$ and location deformations, i.e. $\varphi_\lambda(x) = x + \lambda$. If the deformation model (4.1) is correct, then $M(\theta^*) = 0$, however, it is easy to check that if we have $\tilde{\theta} = (\theta_1^* + a, \ldots, \theta_J^* + a)$, then $M(\tilde{\theta}) = 0$. In order to avoid this type of identifiability problems, we choose the distribution of the first sample as a reference. This is equivalent to considering that θ_1^* is known, or to identifying $\varepsilon_{i1} = X_{i1}$ for every i and then $\mu(\theta^*) = \mu_1$. Therefore, from now on we will consider as parameter set $\Theta = \prod_{j=2}^{J} \Lambda$.

4.2 Estimation of the warping parameters

The idea that comes naturally is to study the empirical version of criterion (4.2) that we obtain by considering empirical laws rather than real ones,

$$M_n(\theta) = \sum_{j=1}^{J} W_2^2\left(\mu_j^{(n)}(\theta), \mu^{(n)}(\theta)\right). \tag{4.3}$$

Where $\mu_j^{(n)}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_{ij}(\theta)}$ is the empirical distribution of the sample $(Z_{ij}(\theta))_{1\leq i\leq n}$, this is, $\mu_j^{(n)}(\theta) = \mu_j^{(n)} \circ \varphi_{\theta_j}$ with $\mu_j^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{ij}}$ the empirical law of each sample. On the other hand, $\mu^{(n)}(\theta) := \frac{1}{J}\sum_{j=1}^{J} \mu_j^{(n)}(\theta)$ is the mean distribution of the $\mu_j^{(n)}(\theta)$'s (note that it corresponds to the empirical law of the sample $(Z_{ij}(\theta))_{\substack{1\leq i\leq n \\ 1\leq j\leq J}}$ formed by random variables that are independent but not identically distributed, so that it does not correspond to the empirical law associated with $\mu(\theta)$).

This leads us to consider as an estimator of the deformation parameters any estimator within the set

$$\arg\min_{\theta\in\Theta} M_n(\theta). \tag{4.4}$$

Next we will study the asymptotic behaviour of the estimator defined in (4.4). But first we need to establish some conditions. The deformations to be considered must fulfil

$$\forall\lambda\in\Lambda, x\mapsto\varphi_\lambda(x) \text{ is invertible from } I_a \text{ to } I_b. \tag{4.5}$$

$$\forall\lambda\in\Lambda,\forall j\in\{1,\dots,J\}, \int_{\mathcal{X}} d(\varphi_\lambda^{-1}\circ\varphi_{\theta_j^\star}(x),x_0)^2 d\mu<\infty \text{ with } x_0\in\mathcal{X}. \tag{4.6}$$

We note that the condition (4.6) is equivalent to

$$\int_{\mathcal{X}} d(x,x_0)^2 d\mu_j(\theta)<\infty \quad \forall\theta\in\Theta \text{ and } x_0\in\mathcal{X}.$$

This condition is necessary for the calculation of the Wasserstein distance. The following is a regularity condition on the deformation functions.

$$\varphi^{-1} \text{ is continuous in } \Lambda\times I_b. \tag{4.7}$$

The following hypothesis ensures that the mass loaded by law $\mu_j(\theta)$ goes to sets where the probability μ_j is very small if $\|\theta_j\|$ is big. We will denote by B the balls in the space $(\mathcal{X},d)$ defined as $B=B(x_0,R):=\{x\in\mathcal{X}: d(x,x_0)<R\}$.

$$\begin{aligned}&\text{For any ball } B \text{ and any number } \nu>0,\\ &\text{there exists a closed set } S \text{ and a constant } H>0 \text{ such that}\\ &\|\lambda\|>H \text{ implies that } \varphi_\lambda(B)\subset S \text{ with } \mu_j[S]<\nu \quad \forall j=1,\dots,J.\end{aligned} \tag{4.8}$$

It is important to remark that the set S may not be bounded.

$$M \text{ has a unique minimizer.} \tag{4.9}$$

This assumption is verified only by the existence of a j such that $2\le j\le J$ and $\varphi_{\theta_j}^{-1}\circ\varphi_{\theta_j^\star}\neq Id$ for $\theta\neq\theta^\star$ in a set of positive measure μ where Id denotes the identity function.

All these assumptions allow us to obtain the following result on the convergence of the estimator $\widehat{\theta}^{(n)}$ defined in (4.4), i.e., $\widehat{\theta}^{(n)}\in\arg\min_{\theta\in\Theta} M_n(\theta)$.

Theorem 4.1. *Under conditions* (4.5) *to* (4.9), $\widehat{\theta}^{(n)}\to\theta^\star$ *almost surely when* $n\to\infty$.

Proof. This proof is inspired by Cuesta and Matrán (1988). We will divide it into three parts, we will start by proving that $\forall\theta \in \Theta$, $M_n(\theta)$ converges almost surely to $M(\theta)$. Then we shall see that $P\left(\{\widehat{\theta}^{(n)}\}_{n\in\mathbb{N}} \text{ is bounded}\right) = 1$. To finish the proof we will see that with probability 1, the sequence $\{\widehat{\theta}^{(n)}\}_{n\in\mathbb{N}}$ has a single accumulation point that will be θ^*.

First part: $\forall\theta \in \Theta$, $M_n(\theta)$ converges almost surely to $M(\theta)$.

We will use the characterization for Wasserstein distance seen in Proposition 2.7. By Varadarajan's theorem (Theorem 11.4.1 in Dudley (2002)) we have for all j and θ it is satisfied almost surely that

$$\mu_j^{(n)}(\theta) \rightharpoonup \mu_j(\theta). \tag{4.10}$$

On the other hand, by the Law of Large Numbers,

$$\int_{\mathcal{X}} d(x,x_0)^2 d\mu_j^{(n)}(\theta) \xrightarrow{n\to\infty} \int_{\mathcal{X}} d(x,x_0)^2 d\mu_j(\theta). \tag{4.11}$$

So, using the characterization of convergence in distributions for continuous and bounded functions, from (4.10), we have that almost surely

$$\mu^{(n)}(\theta) = \frac{1}{J}\sum_{j=1}^{J} \mu_j^{(n)}(\theta) \rightharpoonup \frac{1}{J}\sum_{j=1}^{J} \mu_j(\theta) = \mu(\theta), \tag{4.12}$$

and from (4.11), almost surely

$$\int_{\mathcal{X}} d(x,x_0)^2 d\mu^{(n)}(\theta) \xrightarrow{n\to\infty} \int_{\mathcal{X}} d(x,x_0)^2 d\mu(\theta). \tag{4.13}$$

Using the equivalence (2.11) we get that (4.12) and (4.13) imply that

$$\mathcal{W}_2\left(\mu^{(n)}(\theta), \mu(\theta)\right) \xrightarrow{n\to\infty} 0 \text{ a.s. }.$$

Gathering now the equivalence (2.11) with (4.10) and (4.11), we conclude that for all j and θ fixed

$$\mathcal{W}_2^2\left(\mu_j^{(n)}(\theta), \mu_j(\theta)\right) \xrightarrow{n\to\infty} 0, \text{a.s.}.$$

These last two relationships finally lead us to

$$M_n(\theta) = \sum_{j=1}^{J} \mathcal{W}_2^2\left(\mu_j^{(n)}(\theta), \mu^{(n)}(\theta)\right) \xrightarrow{n\to\infty} M(\theta) = \sum_{j=1}^{J} \mathcal{W}_2^2\left(\mu_j(\theta), \mu(\theta)\right) \text{ a.s.},$$

and with this, the first part is proven.

Second part: $P\left(\{\widehat{\theta}^{(n)}\}_{n\in\mathbb{N}} \text{ is bounded}\right) = 1$.

Since we are assuming that $\mu_1 = \mu$, and that $\mu_1^{(n)}\left(\widehat{\theta}^{(n)}\right) = \mu_1^{(n)}$ because the first sample is not changed, the previous result together with the properties of empirical distributions imply that there is a set with probability 1, Ω_0, where

$$\mathcal{W}_2\left(\mu_1^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) \to 0, \tag{4.14}$$

$$M_n\left(\theta^\star\right) \to M\left(\theta^\star\right), \tag{4.15}$$

$$\mu_j^{(n)} \rightharpoonup \mu_j, \quad \text{para cada } j = 1, \dots J. \tag{4.16}$$

From now on we are going to fix an element in the set Ω_0 and we are going to prove that $\widehat{\theta}^{(n)} \in \arg\min_{\theta\in\Theta} M_n(\theta)$ is bounded by contradiction. Assume that $\limsup_{n\to\infty} \left\|\widehat{\theta}^{(n)}\right\|_{p(J-1)} = \infty$ and choose $j \in \{2, \dots, J\}$ such that $\limsup_{n\to\infty} \left\|\widehat{\theta}_j^{(n)}\right\|_p = \infty$. Consider a subsequence $\{n_h\}_{h\geq 1}$ such that $\lim_{h\to\infty} \left\|\widehat{\theta}_j^{(n_h)}\right\|_p = \infty$. With condition (4.8), for all ball B and all $\nu > 0$ there is a closed set S and an integer h_0 such that, for all $h \geq h_0$, $\varphi_{\widehat{\theta}_j^{(n_h)}}(B) \subset S$ and then $\mu_j\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leq \mu_j[S] \leq \nu$. In consequence,

$$\limsup_{h\to\infty} \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leqslant \limsup_{h\to\infty} \mu_j^{(n_h)}[S].$$

By the portmanteau theorem (see, for example, Theorem 11.1.1 in Dudley (2002)), using convergence in distribution of the measure $\mu_j^{(n_h)}$ given in (4.16), we can write

$$\limsup_{h\to\infty} \mu_j^{(n_h)}[S] \leq \mu_j[S],$$

hence

$$\limsup_{h\to\infty} \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \leq \nu.$$

This inequality holds for all $\nu > 0$, and we can conclude that, for every ball B

$$\mu_j^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)[B] = \mu_j^{(n_h)}\left[\varphi_{\widehat{\theta}_j^{(n_h)}}(B)\right] \xrightarrow{h\to\infty} 0.$$

Moreover, by the definition of $\widehat{\theta}^{(n)}$

$$M_{n_h}\left(\theta^\star\right) \geqslant M_{n_h}\left(\widehat{\theta}^{(n_h)}\right)$$

and, from (4.15) we get

$$0 = M\left(\theta^\star\right) \geq \lim_{h\to\infty} M_{n_h}\left(\widehat{\theta}^{(n_h)}\right), \tag{4.17}$$

from where we deduce that

$$0 = \lim_{h\to\infty} \mathcal{W}_2^2\left(\mu_1^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right), \mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)\right),$$

This, together with (4.14) implies that $\mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right) \rightharpoonup \mu$ when $h \to \infty$ for the Wasserstein distance.

Take $\delta > 0$ and $x_0 \in \mathcal{X}$ fixed, using (4.2), we know that for every $H \in \mathbb{N}$, we can find a r_H such that for all $h \geq r_H$

$$\mu_j^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)\left[B\left(x_0; H\right)\right] \leq \delta, \tag{4.18}$$

and we can build a new subsequence (to which we will continue to call n_h for simplicity) so that for every $H \geq 0$, (4.18) is valid and

$$\mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right) \rightharpoonup \mu.$$

As $\mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)$ is a sequence of measures in a complete and separable metric space that converge in distribution, is a tight sequence. As a result, for all $\nu > 0$ there is a compact set K such that $\mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)[K] \geq 1 - \nu$ for all $h \in \mathbb{N}$. However, since K is bounded, there is M such that $K \subset B\left(x_0, M\right)$. Then

$$\begin{aligned} \nu &\geq \mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)[K^c] \geq \mu^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)\left[B\left(x_0, M\right)^c\right] \\ &\geq \frac{1}{J}\mu_j^{(n_h)}\left(\widehat{\theta}^{(n_h)}\right)\left[B\left(x_0; M\right)^c\right] \geq \frac{1-\delta}{J}. \end{aligned}$$

Taking $\nu = \frac{1-\delta}{2J}$, we have a contradiction. Therefore, we can conclude that in Ω_0, $\limsup_{n\to\infty} \left\|\widehat{\theta}^{(n)}\right\|_{p(J-1)} < \infty$.

Third part: With probability 1, the unique accumulation point of $\{\widehat{\theta}^{(n)}\}_{n\in\mathbb{N}}$ will be θ^*. As the sequence $\left\{\widehat{\theta}^{(n)}\right\}$ is bounded in Ω_0 we just have to prove that if θ^0 is the limit of a subsequence of $\left\{\widehat{\theta}^{(n)}\right\}_{n\geq 1}$, then $\theta^\star = \theta^0$. For simplicity, we will denote the subsequence by $\widehat{\theta}^{(n)}$ and assume that $\widehat{\widehat{\theta}}^{(n)} \to \theta^0$.

Starting from inequality (4.17) we have that for each $j = 1, \ldots, J$,

$$\lim_{n\to\infty} \mathcal{W}_2\left(\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu^{(n)}\left(\widehat{\theta}^{(n)}\right)\right) = 0.$$

Then,

$$\mathcal{W}_2\left(\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu^{(n)}\left(\widehat{\theta}^{(n)}\right)\right) \xrightarrow{n\to\infty} 0, \text{ for each } j = 1, \ldots, J. \tag{4.19}$$

Applying this to $j = 1$ and (4.14), we get

$$\mathcal{W}_2\left(\mu^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) \xrightarrow{n\to\infty} 0.$$

Now, if $j = 1, \ldots, J$, with this and (4.19) we have that

$$\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu. \tag{4.20}$$

On the other hand, (4.16) is satisfied, which allows us to apply Skorohod's Theorem (see, for example, Theorem 11.7.2 in Dudley (2002)) to obtain random vectors Z_n, $n = 0, 1, \ldots$ such that, sequence $\{Z_n\}$ converges almost surely to Z_0 and the distribution of Z_0 is μ_j and the distribution of Z_n is $\mu_j^{(n)}$. As $\widehat{\theta}^{(n)} \to \theta^0$, the assumption (4.7) gives us that

$$\varphi^{-1}_{\hat{\theta}_j^{(n)}}(Z_n) \xrightarrow{n\to\infty} \varphi^{-1}_{\theta_j^0}(Z_0), \text{ a.s..}$$

And, then, we have

$$\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu_j\left(\theta^0\right).$$

This, together with (4.20), means that $\mu_j\left(\theta^0\right) = \mu$ for all j. Finally, given that $M\left(\theta^0\right) = 0 = M\left(\theta^\star\right)$, from assumption (4.9) it follows that $\theta_j^0 = \theta_j^*$ and that concludes the proof of this third part and of the theorem. □

To conclude this section we prove that the average aligned empirical distribution, i.e.,

$$\mu^{(n)}\left(\widehat{\theta}^{(n)}\right) = \frac{1}{J}\sum_{j=1}^{J}\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right)$$

is a good estimator of the original distribution before deformations, μ. To do this we need to establish two new assumptions, for simplicity we assume that the metric space $(\mathcal{X}, d)$ is a closed subset of $\mathbb{R}^d$.

$$\forall x \in I_b,\ \varphi_\lambda^{-1}: \begin{array}{ccc} \Lambda & \to & I_a \\ \lambda & \mapsto & \varphi_\lambda^{-1}(x) \end{array} \text{ is continuously differentiable,} \tag{4.21}$$

and

$$\forall j, \text{ the family } \left(\partial\varphi_\lambda^{-1}(\cdot)\right)_{\lambda\in\Lambda} \text{ has an envelope in } L_2\left(\mu_j\right). \tag{4.22}$$

This means $\sup_{\lambda\in\Lambda}\left\|\partial\varphi_\lambda^{-1}(x)\right\| \leqslant H(x), H \in L^2\left(\mu_j\right)$. With these assumptions we can establish the following proposition.

Proposition 4.2. *Under assumptions* (4.5) *to* (4.9), (4.21) *and* (4.22),

$$\mathcal{W}_2\left(\mu^{(n)}\left(\widehat{\theta}^{(n)}\right), \mu\right) \xrightarrow{n\to+\infty} 0 \text{ almost surely.}$$

Proof. From the proof of Theorem 4.1, it follows that $\mu_j^{(n)}\left(\widehat{\theta}^{(n)}\right) \rightharpoonup \mu$ almost surely. Let ε be a random variable with the same distribution as the variables ε_{ij} introduced in (4.1). We will prove that,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi^{-1}_{\widehat{\theta}_j^{(n)}}(X_{ij})\right\|^2 \xrightarrow{n\to\infty} E\left[\|\varepsilon\|^2\right] \text{ a.s.}$$

Using the Strong Law of Large Numbers, we know that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\theta_j^\star}^{-1}\left(X_{ij}\right)\right\|^2=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varepsilon_{ij}\right\|^2=E\left[\left\|\varepsilon\right\|^2\right]\text{ a.s.},$$

it will therefore suffice to prove that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\theta_j^\star}^{-1}\left(X_{ij}\right)\right\|^2=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\widehat{\theta}_j^{(n)}}^{-1}\left(X_{ij}\right)\right\|^2$. We have

$$\begin{aligned}&\left|\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\widehat{\theta}_j^{(n)}}^{-1}\left(X_{ij}\right)\right\|^2}-\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\theta_j^\star}^{-1}\left(X_{ij}\right)\right\|^2}\right|\\&\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\widehat{\theta}_j^{(n)}}^{-1}\left(X_{ij}\right)-\varphi_{\theta_j^\star}^{-1}\left(X_{ij}\right)\right\|^2}\\&\leq\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|H\left(X_{ij}\right)\right\|^2\left\|\widehat{\theta}_j^{(n)}-\theta_j^\star\right\|}.\end{aligned}$$

which converges, almost surely, to 0 using the Strong Law of Large Numbers and Theorem 4.1. Hence, almost surely,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left\|\varphi_{\widehat{\theta}_j^{(n)}}^{-1}\left(X_{ij}\right)\right\|^2=E\left[\left\|\varepsilon\right\|^2\right].$$

Using the characterization of convergence in Wasserstein distance given in (2.11) the proposition is proven. □

4.3 Computational aspects

When the template measure, μ, is defined in some subset of $\mathbb{R}$, the value of the criterion M_n defined in (4.3) is easily calculable by the expression of 2-Wasserstein distance as distance L_2 between quantum functions and the order statistics of the deformed observations. Remember that if F_n is the empirical distribution associated with the sample $(Y_1,\ldots,Y_n)$, then

$$F_n^{-1}\left(t\right)=Y_{(i)},\text{ for }\frac{i-1}{n}<t\leq\frac{i}{n},$$

where $Y_{(1)},\ldots,Y_{(n)}$ denote the associated order statistics. In the same way we will denote by $Z_{(k)}\left(\theta\right)$ the kth order statistic of the whole sample $\left(Z_{ij}\left(\theta\right)\right)_{\substack{1\leq i\leq n\\1\leq j\leq J}}$, and $Z_{(k)j}$ the corresponding statistic in the sample $\left(Z_{ij}\left(\theta\right)\right)_{1\leq i\leq n}$, which brings us to

$$M_n\left(\theta\right)=\frac{1}{J}\sum_{j=1}^{J}\frac{1}{Jn}\sum_{i=1}^{n}\sum_{k=1}^{J}\left[Z_{(i)j}\left(\theta\right)-Z_{(J(i-1)+k)}\left(\theta\right)\right]^2.$$

An equivalent closed formula for the calculation of M_n is not available in higher dimension. Instead, we resort to solving linear programming problems of the type discussed in section 3.2 with $\alpha_1 = \alpha_2 = 0$. More specifically, if P and Q are uniform laws in $A = \{x_1, \ldots, x_{nJ}\}$ and $B = \{y_1, \ldots, y_n\}$ respectively, then

$$W_2^2(P, Q) = \min_{c \in \mathcal{C}} L(A, B, c),$$

where

$$\mathcal{C} = \left\{ c = (c_{ik})_{\substack{1 \le i \le nJ \\ 1 \le k \le n}} : c_{ik} \ge 0;\ \sum_{i=1}^{nJ} c_{ik} = \frac{1}{n};\ \sum_{k=1}^{n} c_{ik} = \frac{1}{nJ} \quad \forall i, k \right\}$$

and

$$L(A, B, c) := \sum_{(i,k)=(1,1)}^{(nJ,n)} d(x_i, y_k)^2 c_{ik}.$$

Then for $\theta \in \Theta$ we will follow this procedure to calculate, for all j, a sequence $c^j(\theta)$ such that

$$W_2^2\left(\mu_j^{(n)}(\theta), \mu^{(n)}(\theta)\right) = L\left((Z_{ij}(\theta))_{\substack{1 \le i \le n \\ 1 \le j \le J}}, (Z_{ij}(\theta))_{1 \le i \le n}, c^j(\theta)\right),$$

and we get $M_n(\theta) = \frac{1}{J} \sum_{j=1}^{J} L\left((Z_{ij}(\theta))_{\substack{1 \le i \le n \\ 1 \le j \le J}}, (Z_{ij}(\theta))_{1 \le i \le n}, c^j(\theta)\right)$.

4.4 Examples

In this section we show examples of admissible deformations that satisfy the hypotheses (4.5) to (4.9), (4.21) and (4.22).

Example 4.1. Location/scale model,

$$\varphi_\lambda(x) = \frac{x}{\lambda_2} + \lambda_1.$$

In this case we are considering that $\mathcal{X} = \mathbb{R}^d$. While $\lambda_2 \neq 0$, φ_λ is invertible in $\mathbb{R}^d$ hence $\Lambda \subset \mathbb{R}^d \times \mathbb{R} - \{0\}$ and

$$\varphi_\lambda^{-1}(x) = \lambda_2 x - \lambda_1 \lambda_2 = \varphi_{(-\lambda_1\lambda_2, \frac{1}{\lambda_2})}(x).$$

Then

$$\varphi_\lambda^{-1}(\varphi_\beta(x)) = \frac{\lambda_2}{\beta_2} x + \beta_1 - \lambda_1 \lambda_2$$

is in $L^2(\mu)$ if $\mu \in \mathcal{W}_2\left(\mathbb{R}^d\right)$. Therefore the assumptions (4.5), (4.6), (4.7) and (4.9) will be satisfied as long as μ is in $\mathcal{W}_2\left(\mathbb{R}^d\right)$.

We now study the condition (4.8). Assume μ has a bounded density with respect to Lebesgue's measure. So for all $\nu > 0$, we can find η such that for all x, $\mu_j\left(\bar{B}(x,\eta)\right) < \nu$ and M such that $\mu_j\left(B(0,M)^c\right) < \nu$.

We have $\varphi_\lambda\left(B(y_0,r)\right) = B(\frac{1}{\lambda_2}y_0 + \lambda_1, \frac{1}{|\lambda_2|}r)$. Then, with $|\lambda_2| \to \infty$, if $\|\lambda\|$ is sufficiently large, $\varphi_\lambda\left(B(y_0,r)\right) \subset \bar{B}(x_\lambda,\eta)$. Now, if $\|\lambda_1\| \to \infty$, then $\varphi_\lambda\left(B(y_0,r)\right) \subset B(0,M)^c$ if $|\lambda_2| \geq \alpha > 0$. Therefore, (4.8) shall be valid if, in addition, $\Lambda \subset \mathbb{R}^d \times (-\infty,-\alpha] \cup [\alpha,+\infty)$ with $\alpha > 0$.

Example 4.2. Logarithmic transformation, here $\mathcal{X} \subset (0,+\infty)$ and

$$\varphi_\lambda(x) = \frac{1}{\lambda}\ln(x).$$

Function φ_λ is invertible from $(0,+\infty)$ to $\mathbb{R}$ for all $\lambda \neq 0$. Then $\Lambda \in (0,+\infty)$ and the support of μ must be contained in $(0,+\infty)$. We have that

$$\varphi_\lambda^{-1}(x) = \exp(x\lambda)$$

and $\varphi_\lambda^{-1}(\varphi_\beta(x)) = \exp\left(\frac{\lambda\ln(x)}{\beta}\right) = x^{\frac{\lambda}{\beta}}$. Then $\varphi_\lambda^{-1} \in L^2\left(X_{1j}\right)$ if $E\left[e^{\frac{2\lambda}{\theta_j^\star}}\right] < \infty$ for all $\lambda \in \Lambda$. To sum up, the conditions (4.5), (4.6), (4.7) and (4.9) will be satisfied provided that the support of μ is contained in $(0,+\infty)$ and that $E\left[e^{\frac{2\lambda}{\theta_j^\star}}\right] < \infty$ for all $\lambda \in \Lambda$.

In this case we have more restrictive conditions on the law of μ than in the previous example. However, the exponential distribution, for example, satisfies them.

Again, hypothesis (4.8) has yet to be verified. For y_0 and r such that $B(y_0,r) \subset (0,+\infty)$, we have that $\ln(B(y_0,r)) \subset \ln\left(\bar{B}(y_0,r)\right) \subset \bar{B}(z_0,R)$ for some z_0 as the image of a compact set by a continuous function remains compact. Therefore, $\varphi_\lambda\left(B(y_0,r)\right) \subset \bar{B}\left(\frac{z_0}{\lambda}, \frac{R}{|\lambda|}\right)$ and the condition is satisfied if $\mu(0) = 0$.

Example 4.3. Affine transformation,

$$\varphi_\lambda(x) = A^{-1}x + b.$$

As in Example 4.1 We will consider $\mathcal{X} = \mathbb{R}^d$ and $\lambda = (A,b) \in GL\left(\mathbb{R}^d\right) \times \mathbb{R}^d$, where $GL(\mathbb{R}^d)$ denotes the space of invertible matrices $d \times d$ with real coefficients. We have got

$$\varphi_\lambda^{-1}(x) = A\left(x - b\right).$$

Then $\varphi_{\lambda_1}^{-1} \circ \varphi_{\lambda_2}(x) = A_1\left(A_2^{-1}x + b_2 - b_1\right)$, hence if μ is in $\mathcal{W}_2\left(\mathbb{R}^d\right)$ hypothesis (4.5), (4.6), (4.7) and (4.9) will be satisfied.

For $y_0 \in \mathbb{R}^d$ and $r > 0$ we have $\varphi_\lambda(B(y_0,r)) \subset B\left(A^{-1}y_0 + b, r\left\|A^{-1}\right\|\right)$. Then, as in Example 4.1, condition (4.8) it will be verified if we take μ with bounded density with

respect to Lebesgue measure and if we choose the matrix A in a subset of $GL\left(\mathbb{R}^d\right)$ with $\|A\| \geq \alpha > 0$.

Example 4.4. Composition,

$$\varphi_\lambda(x) = f \circ \tilde{\varphi}_\lambda(x).$$

Consider a function $\tilde{\varphi}_\lambda(x)$, a law μ and a parameter $\theta^\star$ that satisfies assumptions (4.5) to (4.9). Then, if f is a homeomorphism from I_b to I_c the deformation function $\varphi_\lambda(x) = f \circ \tilde{\varphi}_\lambda(x)$ with the same law μ and a parameter $\theta^\star$ also verifies these assumptions by replacing I_b with I_c, let us see this. Assumptions (4.5) and (4.7) are easy to check. We have

$$\varphi_\lambda^{-1} \circ \varphi_\beta = \tilde{\varphi}_\lambda^{-1} \circ f^{-1} \circ f \circ \tilde{\varphi}_\beta = \tilde{\varphi}_\lambda^{-1} \circ \tilde{\varphi}_\beta.$$

Then $\tilde{\mu}_j(\lambda) = \mu \circ \tilde{\varphi}_{\theta_j^\star}^{-1} \circ \tilde{\varphi}_\lambda = \mu \circ \varphi_{\theta_j^\star}^{-1} \circ \varphi_\lambda = \mu_j(\lambda)$ and hypothesis (4.6) and (4.9) will also be met. In addition, the criterion $\tilde{M}(\theta)$ corresponding to $\tilde{\varphi}$ is exactly the same as the criterion $M(\theta)$ corresponding to φ.

Again, it remains to be checked that assumption (4.8) is also met. We choose $\nu > 0$, a ball B and a closed set S such that $\tilde{\varphi}_\lambda(B) \subset S$ with $\tilde{\mu}_j[S] < \nu$ for all j with λ sufficiently large. We have that $\varphi_\lambda(B) \subset f(S)$ is closed because f is a homeomorphism. Even more, $\mu_j[f(S)] = \tilde{\mu}_j[S] < \nu$.

This allows us to consider new deformations. For example, in the same framework of the scale model we can consider the logit model with the deformation function $\varphi_\lambda(x) = (1 + \exp(x/\lambda))^{-1}$.

4.5 Simulations

To conclude, we present a simulation study to test the operation of the described method. Samples of independent random variables ε_{ij} of size n are generated with a standard normal distribution where $j = 1, \ldots, J$ and $i = 1, \ldots, n$. For given parameters $\theta_j^\star$ that we are going to estimate, we simulate the observations $X_{ij} = \varphi_{\theta_j^\star}(\varepsilon_{ij})$. In order to obtain the estimation of the deformation parameters, the criterion defined in (4.4) is minimized. The main difficulty comes from minimizing the Wasserstein distance. To do this we are going to use the software `R`, namely the function `optim` of the package `stats` with the L-BFGS-B option that minimizes a function with a quasi-Newton method and the function `Partial.Transport` described in section 3.2.1. Simulations are very fast at first, but the computational complexity increases with sample and parameter sizes. We have considered two different types of deformation functions:

1. A localization and scale model as in the Example 4.1:

$$\varphi_{\theta_j^*}(x) = \frac{x}{\theta_1^*} + \theta_2^*$$

The simulations have been made for 5 pairs of parameters $(\theta_1^\star, \theta_2^\star)$: (2,1), (7,2), (2.5,2), (5,5) and (10,8).

2. A matrix as a scale parameter, corresponding to the affine transformation of Example 4.3:

$$\varphi_{\theta_j^*}(x) = (\theta_1^*)^{-1}x + \theta_2^*\mathbf{1}.$$

where $x = (x_1, x_2)^T \in \mathbb{R}^2$, θ_1^* is $\begin{pmatrix} \theta_{11}^* & 0 \\ 0 & \theta_{12}^* \end{pmatrix} \in GL(\mathbb{R}^2)$ and $\theta_2^* \in \mathbb{R}$. Here $\lambda = (\theta_{11}, \theta_{12}, \theta_2)$. The simulations have been made for 4 sets of parameters which are (2,2,1), (4,3,2), (2.5,3,1.5) and (5,2,2).

All simulations have been made for $J = 2$ and sample sizes $n = 20, 50, 100, 200, 300$. To analyze the estimation error, we can see in figures 4.1 to 4.5 the boxplots of these errors.

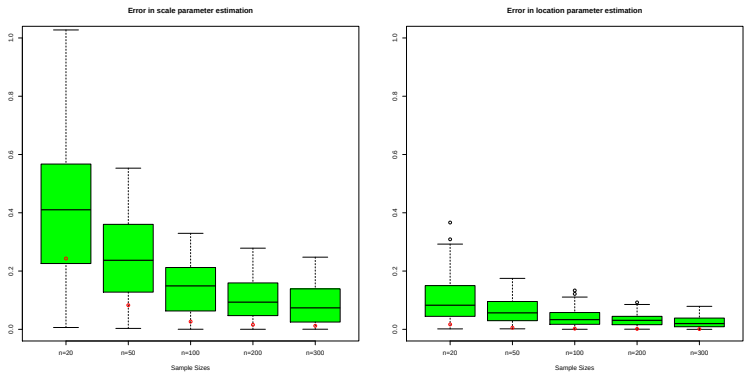


Figure 4.1: Model 1(location/scale). Estimation error. $\lambda = (2, 1)$.

As expected, the error decreases with sample size leading to good estimators for sufficiently large samples. We would like to draw attention to the fact that the estimation is better for the localization parameter than for the scale parameter. This can be seen in tables 4.1 and 4.2 showing the mean and standard deviation of these parameters for the different sample sizes.

Note that in the case of the location parameter, the mean of the estimator in simulations is very close to the actual value of the deformation parameter and the standard

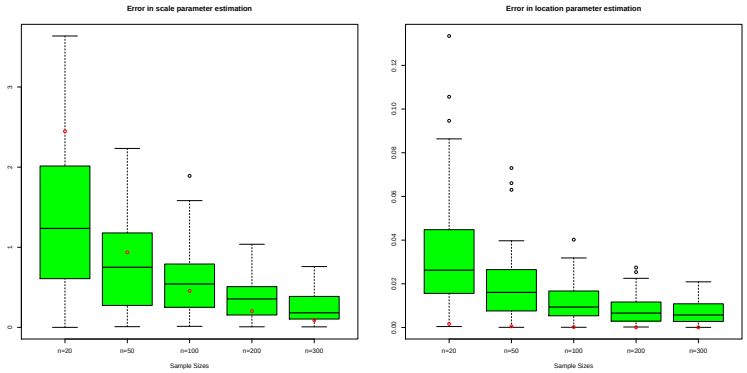


Figure 4.2: Model 1(location/scale). Estimation error. $\lambda = (7, 2)$.

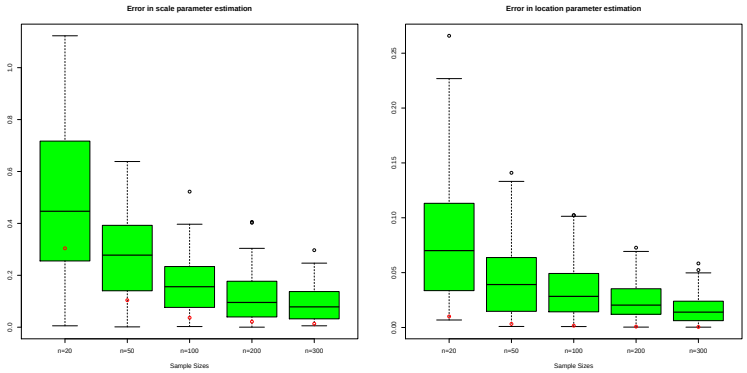


Figure 4.3: Model 1(location/scale). Estimation error. $\lambda = (2.5, 7)$.

deviation is small even for small-sized samples. For the scale parameter the results are not as good for smaller samples, but improve significantly with the increase of sample size.

We now consider model 2, affine transformation. As before, figures 4.6, 4.7, 4.8 and 4.9 present the boxplots with the error in the estimation of each parameter.

As expected, something similar to the case of localization and scale transformation happens, the error is fastly reduced by increasing sample size and is much smaller for the localization parameter than for the scale parameter. This can be seen more clearly in Tables 4.3, 4.4 and 4.5.

In conclusion, the simulations show a good behavior of the alignment procedure described in this chapter and, in particular, the consistency of the proposed estimators. The procedure works for single or multivariate data. In the latter case, the algorithm introduced in chapter 3 is essential for the practical implementation of the method.

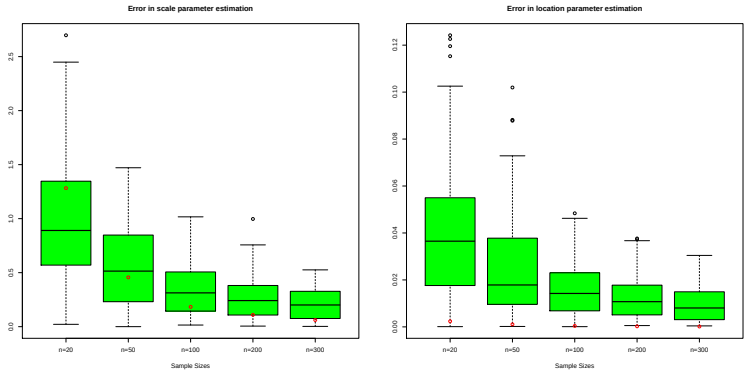


Figure 4.4: Model 1(location/scale). Estimation error. $\lambda = (5, 5)$.

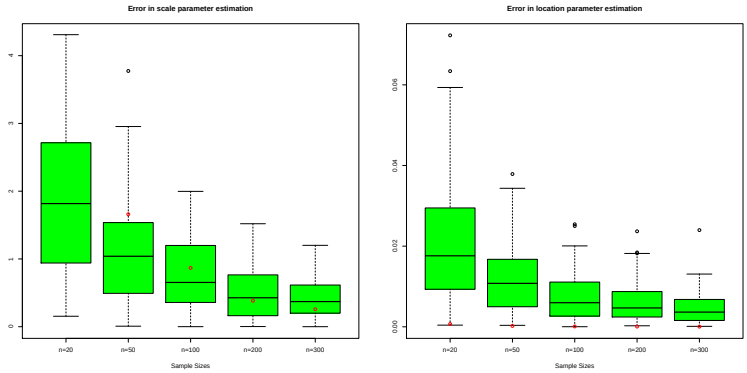


Figure 4.5: Model 1(location/scale). Estimation error. $\lambda = (10, 8)$.

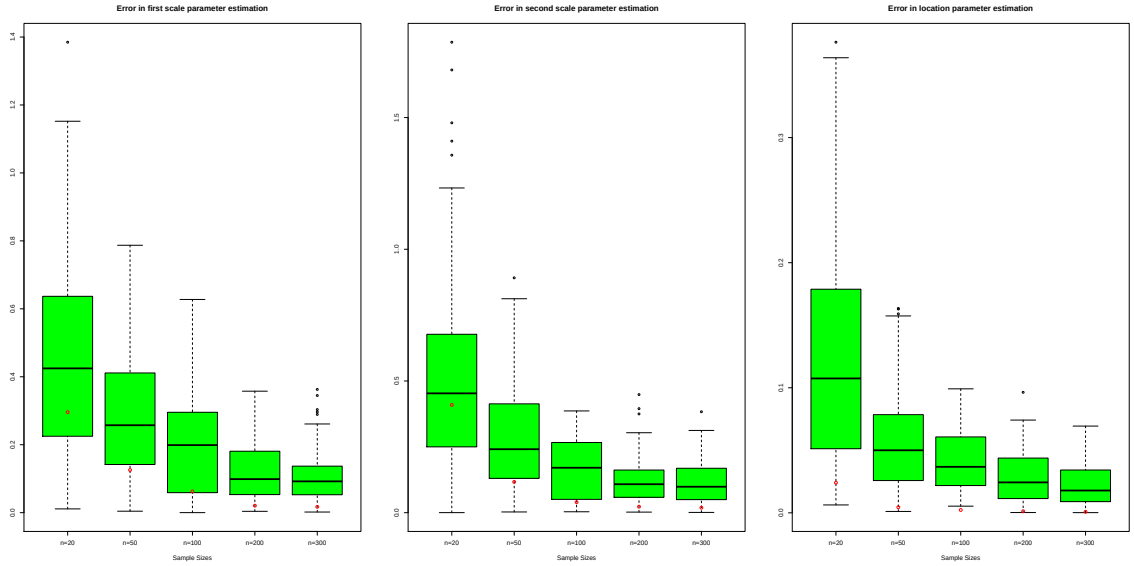


Figure 4.6: Model 2 (affine transformation). Estimation error. $\lambda = (2, 2, 1)$.

Table 4.1: Model 1(location/scale). Scale parameter estimation.

Mean

Sample Size\Parameters	**2**,1	**2.5**,7	**5**,5	**7**,2	**10**,8
20	1,627	2,084	4,167	5,968	8,389
50	1,778	2,262	4,470	6,430	9,211
100	1,892	2,378	4,708	6,536	9,379
200	1,922	2,398	4,791	6,725	9,669
300	1,950	2,435	4,875	6,810	9,719

Standard Deviation

Sample Size\Parameters	**2**,1	**2.5**,7	**5**,5	**7**,2	**10**,8
20	0,324	0,364	0,772	1,182	1,549
50	0,184	0,219	0,421	0,785	1,008
100	0,122	0,148	0,315	0,491	0,698
200	0,098	0,105	0,257	0,359	0,525
300	0,098	0,096	0,215	0,233	0,425

Table 4.2: Model 1(location/scale). Location parameter estimation.

Mean

Sample Size\Parameters	2,**1**	2.5,**7**	5,**5**	7,**2**	10,**8**
20	1,002	6,994	5,003	1,998	7,998
50	0,994	7,001	5,004	2,004	7,999
100	0,996	7,003	4,997	2,000	7,999
200	1,001	6,999	4,999	2,001	8,000
300	1,001	6,997	5,001	1,999	8,000

Standard Deviation

Sample Size\Parameters	2,**1**	2.5,**7**	5,**5**	7,**2**	10,**8**
20	0,130	0,101	0,049	0,041	0,026
50	0,074	0,057	0,033	0,023	0,014
100	0,048	0,041	0,020	0,014	0,009
200	0,039	0,028	0,015	0,010	0,008
300	0,031	0,022	0,012	0,008	0,006

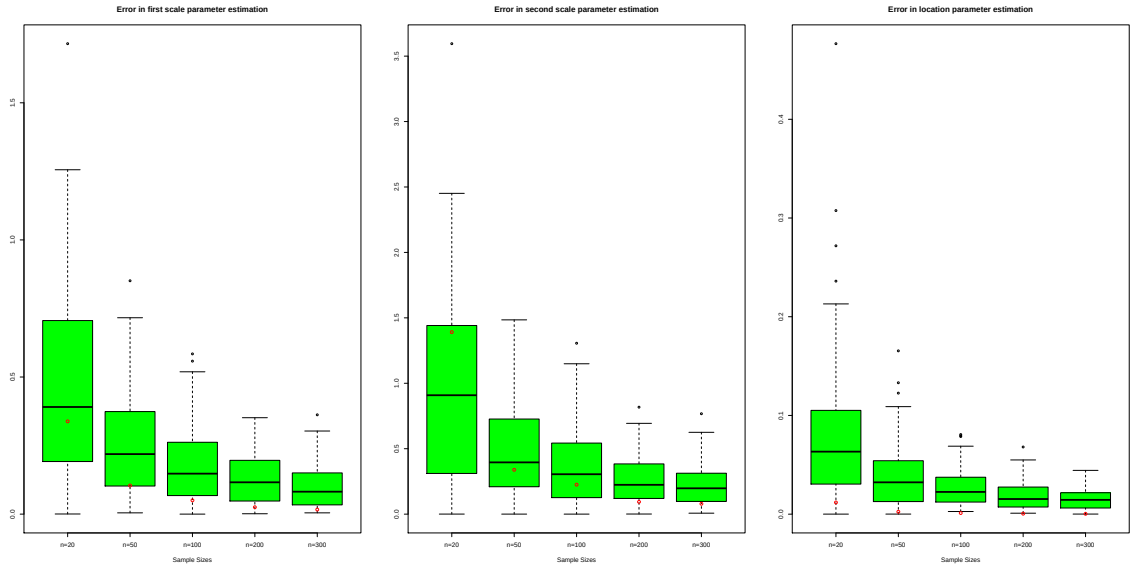


Figure 4.7: Model 2 (affine transformation). Estimation error. $\lambda = (2, 4, 3)$.

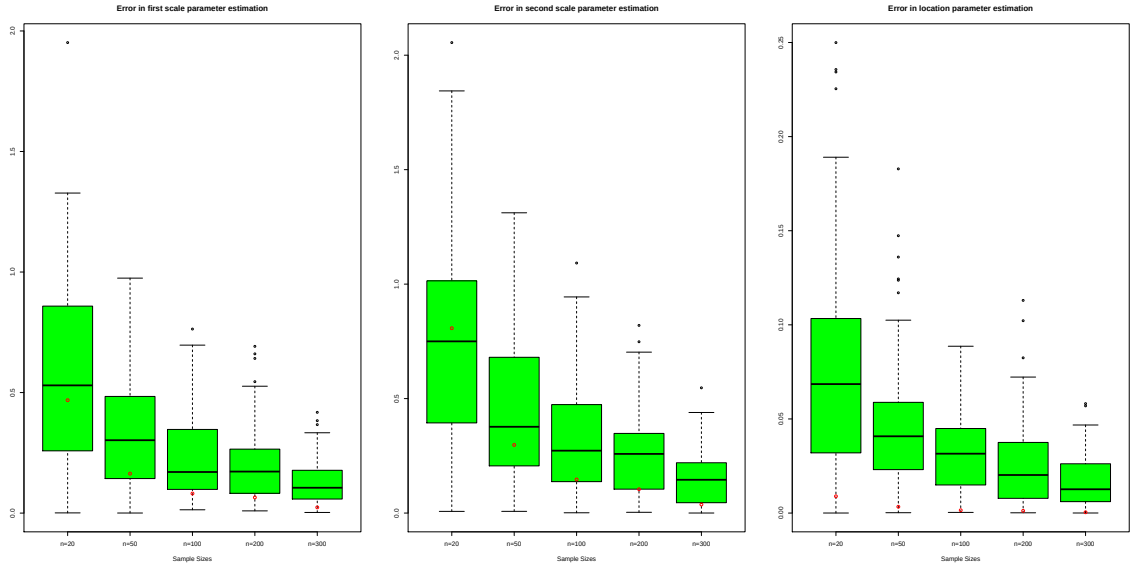


Figure 4.8: Model 2 (affine transformation). Estimation error. $\lambda = (2.5, 3, 1.5)$.

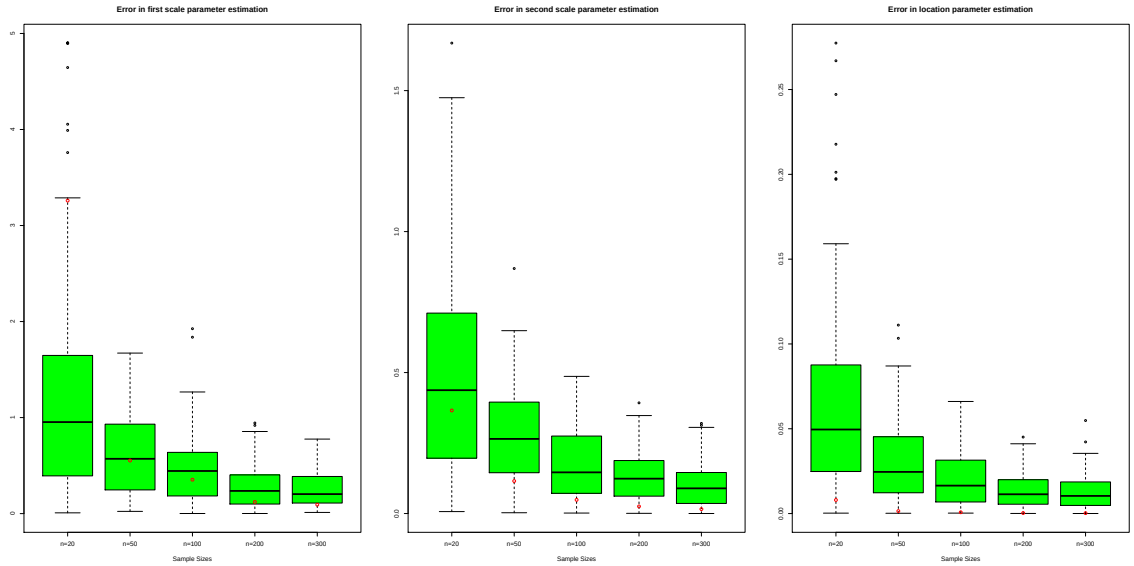


Figure 4.9: Model 2 (affine transformation). Estimation error. $\lambda = (5, 2, 2)$.

Table 4.3: Model 2 (affine transformation). First scale parameter estimation.

Mean

Sample Size\Parameters	**2**,2,1	**2.5**,3,1.5	**5**,2,2	**2**,4,3
20	1,743	2,128	4,261	1,721
50	1,803	2,315	4,629	1,839
100	1,888	2,383	4,697	1,903
200	1,956	2,392	4,851	1,933
300	1,940	2,418	4,884	1,931

Standard Deviation

Sample Size\Parameters	**2**,2,1	**2.5**,3,1.5	**5**,2,2	**2**,4,3
20	0,482	0,577	1,656	0,513
50	0,296	0,361	0,648	0,281
100	0,225	0,261	0,514	0,203
200	0,137	0,232	0,316	0,145
300	0,117	0,131	0,283	0,109

Table 4.4: Model 2 (affine transformation). Second scale parameter estimation.

Mean

Sample Size\Parameters	2,**2**,1	2.5,**3**,1.5	5,**2**,2	2,**4**,3
20	1,762	2,449	1,734	3,351
50	1,850	2,775	1,836	3,845
100	1,885	2,846	1,876	3,830
200	1,939	2,826	1,940	3,905
300	1,958	2,919	1,942	3,918

Standard Deviation

Sample Size\Parameters	2,**2**,1	2.5,**3**,1.5	5,**2**,2	2,**4**,3
20	0,596	0,713	0,546	0,990
50	0,309	0,498	0,297	0,564
100	0,163	0,352	0,182	0,445
200	0,139	0,273	0,148	0,293
300	0,132	0,177	0,112	0,271

Table 4.5: Model 2 (affine transformation). Location parameter estimation.

Mean

Sample Size\Parameters	2,2,**1**	2.5,3,**1.5**	5,2,**2**	2,4,**3**
20	0,977	1,491	1,992	3,013
50	0,995	1,495	2,005	2,997
100	1,001	1,498	2,001	2,997
200	0,999	1,503	1,997	2,998
300	0,998	1,499	1,998	2,998

Standard Deviation

Sample Size\Parameters	2,2,**1**	2.5,3,**1.5**	5,2,**2**	2,4,**3**
20	0,154	0,094	0,090	0,109
50	0,067	0,058	0,038	0,049
100	0,048	0,039	0,028	0,034
200	0,035	0,034	0,017	0,022
300	0,029	0,021	0,016	0,018

Chapter 5

Partial classification problems

This chapter introduces a new robust approach to the classic classification problem described in section 2.2. The main motivation is that the usual methods of obtaining classification rules can be seriously affected by the presence of atypical observations in the training sample. Searching for a classification rule for samples of this type can lead us to find unnecessarily complex rules. In this chapter we will address the problem of finding simpler classification rules by relaxing the objective of classifying the entire sample to classifying only a fraction of it. This problem will be referred to as partial classification problem. The first section of this chapter is devoted to formally addressing this problem.

Later, section 5.2 discusses the problem in detail from the point of view of the 0/1 loss. A penalized classification method is proposed that allows the proportion of atypical observations to be adequately detected and an appropriate rule to be selected. The quality of the rules obtained with this method is guaranteed by oracle inequalities that will be presented in this section. The effective calculation of these rules is, however, computationally problematic. This leads to the consideration of a convex loss function in the adaptation of previous ideas to a computationally feasible environment developed in section 5.3. Finally, in section 5.4, new strategies are explored for the resolution of the partial classification problem, which were either discarded due to the poor results they provided, or have been left as future work.

The methods proposed in this thesis consider only the problem of binary classification, although most methods and results could be extended to multiclass classification problems. This extension is beyond the scope of this work.

5.1 Problem statement

When the number of observations is very large, or we have samples in a high dimension, some of the observations may contain errors and should be considered as contaminating observations. The presence of such observations, if not eliminated, hampers the efficiency of the classifiers, because many of the existing classification methods are very sensitive to atypical observations. Moreover, if the training set is too contaminated, training a classifier in this set will lead us to classifiers with a high classification error. Hence there is a growing need for robust classification methods to tackle such issue.

One solution to cope with this issue is to allow the classifier not to classify all points and to reject some observations that seem too difficult to classify. This point of view is addressed in Herbei and Wegkamp (2006) and Bartlett and Wegkamp (2008). Another idea is to eliminate a proportion of the contaminated data in such a way that we guarantee the robustness of the obtained rule. These observations are known as outliers in the sense that they are far from the model used to generate the data. However, the automatic detection of outliers is a complicated task, its definition is not unique and specific criteria are often used depending on the applications of interest. For example, in the case of SVM classification, it is proposed to remove atypical ones using an 'outlier map'. In Debruyne (2009) the authors propose a function that measures the impact of distribution contamination on the classification error obtained by minimizing empirical risk. In the regression framework, LASSO estimators also suffer in the presence of outliers. Several modifications have been proposed to strengthen the estimators, for example, see Chen et al. (2010), Maronna (2011) or Alfons et al. (2013) where points with the highest residuals are discarded.

Moving from data analysis to a probabilistic framework, removing observations that achieve bad classification error, corresponds to trimming the initial distribution of the observations and replacing it by a trimmed distribution Q. Although research on trimming methods has been carried out for many years, there are still very few theoretical results that allow us to choose a clear limit between an acceptable observation and an outlier. Moreover, in this case little is known about whether or how this choice modifies the classification error. Both for a practical and theoretical purpose, this choice must be guided to take into account the variability generated by the contaminated data.

In this chapter we provide some theoretical guarantee to choose the level of data to be removed. For this we consider the set of trimmed distributions obtained from the initial distribution of the data and look for an automatic rule that reduces the classification error of a collection of classifiers by removing some properly selected observations. The

more data is removed, the easier it becomes to classify the data, leading to a perfect classification if the classification error is small enough. Yet, removing too much data reduces the interest of the classification procedure. If too many observations are left aside then the chosen classification rule may be good for a distribution which is possibly very far from the true distribution of the data. In this section we provide an empirical rule that automatically selects the minimum level of trimming to reduce the classification error for a class of classifiers. Simultaneously, the best classifier for the trimmed set of observations is chosen among a collection of classification rules.

When quantifying the generalization error, we use a loss function. In this work we are going to study two different functions. First of all, we will study the case in which the loss is 0/1, meaning that we only consider two possibilities, whether an observation is well classified or not. This is possibly the most natural loss function and the easiest to interpret. We will show that the proposed rules related to this loss function have interesting theoretical properties. However, the 0/1 loss leads to minimization problems with a non-convex (not even continuous) target function. This means that in practice the calculation of the minimum is very costly, making it impossible to calculate the optimal rule even for small samples. To fix this problem, we are going to consider a convex loss function, namely the *hinge* function. This function is related to the SVM method we saw in section 2.2.2 and is not limited to considering whether an observation is well or poorly classified. In case of misclassification also gives a measure of how bad this classification is.

5.2 Partial Classification with 0/1 loss

From now on we will focus on a single loss function, the 0/1 function. The results of this section can be found in Agulló-Antolín et al. (2017).

First of all, we will study how the generalization error (minimum) affects the modification of the previous problem in such a way that we do not have as our objective the classification based on P, but we will settle for a good classification for a fraction of P, that is, for a trimming of P. In other words, we are faced with the problem of calculating (or characterizing) the *Bayes trimmed error*:

$$\mathrm{Err}_\alpha(P) := \inf_{Q \in \mathcal{R}_\alpha(P)} \mathrm{Err}(Q).$$

In this chapter we will keep the notation introduced in section 2.2 in which we identified the probability of a measurable set $A \subset \{0,1\} \times \mathbb{R}^d$ with three probabilities p_0, P_0 and P_1

so that

$$P(A) = p_0 P_0(A_0) + (1 - p_0)P_1(A_1),$$

where $p_0 = P(\{0\} \times \mathbb{R}^d)$, $P_0(A_0) = P(\{0\} \times A_0)/p_0$ and $P_1(A_1) = P(\{1\} \times A_1)/(1 - p_0)$ with $A_i = \{x \in \mathbb{R}^d : (i, x) \in A\}$, $i = 0, 1$. Similarly, we are going to describe the set $\mathcal{R}_\alpha(P)$ in terms of the identification $P \leftrightarrow (p_0, P_0, P_1)$. This identification is independent of the loss function we are using and will be kept throughout the entire chapter.

Lemma 5.1. *With the previous notation, if $Q \equiv (q_0, Q_0, Q_1)$ with $q_0 \in (0, 1)$, then $Q \in \mathcal{R}_\alpha(P)$ if and only if*

$$q_0 \leq \frac{p_0}{1-\alpha}, \quad 1 - q_0 \leq \frac{1-p_0}{1-\alpha}, \quad Q_0 \in \mathcal{R}_{1-\frac{q_0}{p_0}(1-\alpha)}(P_0) \quad \text{and} \quad Q_1 \in \mathcal{R}_{1-\frac{1-q_0}{1-p_0}(1-\alpha)}(P_1). \tag{5.1}$$

Proof. Note first that $q_0 = Q(\{0\} \times \mathbb{R}^d)$, $Q \in \mathcal{R}_\alpha(P)$ implies $q_0 \leq \frac{1}{1-\alpha} P(\{0\} \times \mathbb{R}^d) = \frac{p_0}{1-\alpha}$. The same argument shows that $1 - q_0 \leq \frac{1-p_0}{1-\alpha}$ if $Q \in \mathcal{R}_\alpha(P)$. Observe that the conditions $q_0 \leq \frac{p_0}{1-\alpha}$ and $1-q_0 \leq \frac{1-p_0}{1-\alpha}$ guarantee that $0 \leq 1-\frac{q_0}{p_0}(1-\alpha) \leq 1$ and $0 \leq 1-\frac{1-q_0}{1-p_0}(1-\alpha) \leq 1$, hence the trimming sets in the statement are well defined. Moreover, if $Q \in \mathcal{R}_\alpha(P)$ then

$$Q_0(A_0) = \frac{Q(\{0\} \times A_0)}{q_0} \leq \frac{1}{(1-\alpha)q_0} P(\{0\} \times A_0) = \frac{1}{(1-\alpha)\frac{q_0}{p_0}} P_0(A_0),$$

which proves that $Q_0 \in \mathcal{R}_{1-\frac{q_0}{p_0}(1-\alpha)}(P_0)$. In a similar way it can be proven that $Q_1 \in \mathcal{R}_{1-\frac{1-q_0}{1-p_0}(1-\alpha)}(P_1)$, which proves that the assumptions (5.1) are necessary. To prove the sufficiency note that if we have (5.1) then $q_0 Q_0(A_0) \leq \frac{1}{1-\alpha} P_0(A_0)$, $(1 - q_0)Q_1(A_1) \leq \frac{1}{1-\alpha} P_1(A_1)$ and hence

$$\begin{aligned} Q(A) &= q_0 Q_0(A_0) + (1 - q_0)Q_1(A_1) \\ &\leq \frac{1}{1-\alpha}(p_0 P_0(A_0) + (1 - p_0)P_1(A_1)) = \frac{1}{1-\alpha} P(A). \end{aligned}$$

which completes the proof. □

Assume now that P_0, P_1 are absolutely continuous with respect to the measure μ, with densities f_0, f_1. If $Q \in \mathcal{R}_\alpha(P)$ then Q_i is also absolutely continuous with respect to μ. We denote the corresponding densities by g_i, $i = 0, 1$. As a consequence of Lemma 5.1 $q_0 g_0 \leq \frac{p_0 f_0}{1-\alpha}$ and $(1 - q_0)g_1 \leq \frac{(1-p_0)f_1}{1-\alpha}$. If we denote $u_0 = q_0 g_0$, $u_1 = (1 - q_0)g_1$ then the Bayes error associated with Q is $\int_{\mathbb{R}^d} \min(u_0, u_1) d\mu$. With this notation we have that $\int_{\mathbb{R}^d} (u_0 + u_1) d\mu = 1$, $u_0 \leq \frac{p_0 f_0}{1-\alpha}$ and $u_1 \leq \frac{(1-p_0)f_1}{1-\alpha}$. Of course, the formulation in terms of g_i is trivially recovered from u_i. This proves the next result.

Lemma 5.2. *With the previous notation*

$$\begin{aligned}\mathrm{Err}_\alpha(P) &= \inf\left\{\int_{\mathbb{R}^d}\min(u_0,u_1)d\mu : 0\le u_0\le\frac{p_0f_0}{1-\alpha},\ 0\le u_1\le\frac{(1-p_0)f_1}{1-\alpha},\right. \quad (5.2)\\ &\left.\int_{\mathbb{R}^d}(u_0+u_1)d\mu=1\right\}. \quad (5.3)\end{aligned}$$

Later, we will check that the minimum on the right-hand side of (5.2) is attained, that is, there is a probability $Q\in\mathcal{R}_\alpha(P)$ that minimizes Bayes error. This trimming will be the best possible for the classification problem when all rules are admissible.

Next we will study the minimum value of α needed to achieve a perfect separation, i.e., $Err_\alpha(P)=0$. We know that if, for example, $\alpha>\min(p_0,1-p_0)$ then $\mathrm{Err}_\alpha(P)=0$, but that bound can be improved as we will see below. Assume now that we have a given classification rule g, its generalization error in terms of identification $P\leftrightarrow(p_0,P_0,P_1)$ is

$$R(g)=p_0P_0(\{x\in\mathbb{R}^d:g(x)=1\})+(1-p_0)P_1(\{x\in\mathbb{R}^d:g(x)=0\}).$$

In analogy to how we defined the trimmed Bayes error we are now going to define the trimmed generalization error for a given rule.

Definition 5.1. *We define the* trimmed generalization error *of a rule g as,*

$$R_\alpha(g):=\inf_{Q\in\mathcal{R}_\alpha(P)}Q(g(X)\neq Y).$$

Specifically, we are interested in studying the relationship between the trimmed classification error and the classification error of the same classifier g for the whole sample. As we will see below, we can determine the trimmed error from the generalization error and the trimming level. To simplify the proof, we have divided the following result into Lemma 5.3 and Proposition 5.4.

Lemma 5.3. *With the previous notation*

$$\begin{aligned}R_\alpha(g) &= \min_{1-\frac{1-p_0}{1-\alpha}\le q_0\le\frac{p_0}{1-\alpha}}\left[\left(q_0-\frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d:g(x)=0\})\right)_+\right.\\ &+ \left.\left(1-q_0-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d:g(x)=1\})\right)_+\right]. \quad (5.4)\end{aligned}$$

Proof. The first step consists in writing the probability Q in terms of (q_0,Q_0,Q_1)

$$\begin{aligned}&Q(g(x)\neq y)=q_0\int_{\{x\in\mathbb{R}^d:g(x)=1\}}\frac{dQ_0}{d\mu}d\mu+(1-q_0)\int_{\{x\in\mathbb{R}^d:g(x)=0\}}\frac{dQ_1}{d\mu}d\mu\\ &=\int\left(q_0I_{(\{x\in\mathbb{R}^d:g(x)=1\})}\frac{dQ_0}{d\mu}+(1-q_0)I_{(\{x\in\mathbb{R}^d:g(x)=0\})}\frac{dQ_1}{d\mu}\right)d\mu. \quad (5.5)\end{aligned}$$

We are looking for the probability that minimizes the probability of error among all the probabilities $Q \in \mathcal{R}_\alpha(P)$, this means we are looking for Q_0 and Q_1 that minimize (5.5). We are going to make the calculations for Q_0, Q_1 can be treated similarly.
As we are minimizing, we are going to concentrate the probability Q_0 in the set $(g(x) = 0)$. By Lemma 5.1 we know that $Q_0 \leq \frac{p_0}{q_0(1-\alpha)} P_0$, so the value of Q_0 depends on the value of P_0. There are two possibilities,

1. $P_0(g(x) = 0) \geq \frac{q_0}{p_0}(1-\alpha)$: As $\frac{p_0}{q_0(1-\alpha)} P_0 \geq 1$ we can concentrate all the probability Q_0 in the set $\{x \in \mathbb{R}^d / g(x) = 0\}$ and hence $Q_0(g(x) = 0) = 1$.

2. $P_0(g(x) = 0) < \frac{q_0}{p_0}(1-\alpha)$: Now we cannot give probability 1 to $Q_0(g(x) = 0)$ because we would be violating the condition in Lemma 5.1, hence $Q_0(g(x) = 0) = \frac{P_0(g(x)=0)}{\frac{q_0}{p_0}(1-\alpha)}$.

And in the optimum we will have

$$Q_0(\{x \in \mathbb{R}^d : g(x) = 0\}) = \min\left(\frac{P_0(\{x \in \mathbb{R}^d : g(x) = 0\})}{\frac{q_0}{p_0}(1-\alpha)}, 1\right).$$

We are focusing in $Q_0(g(x) = 1)$, as Q_0 is a distribution,

$$Q_0(g(x) = 1) = \left(1 - \frac{p_0}{q_0(1-\alpha)} P_0(g(x) = 0)\right)_+,$$

analogously

$$Q_1(g(x) = 0) = \left(1 - \frac{1-p_0}{(1-q_0)(1-\alpha)} P_1(g(x) = 1)\right)_+.$$

So for a fixed q_0 and Q_0, Q_1 as in Lemma 5.1

$$\begin{aligned} \min_{Q_0, Q_1} Q(g(x) \neq y) &= q_0 \left(1 - \frac{p_0}{q_0(1-\alpha)} P_0(g(x) = 0)\right)_+ \\ &+ (1-q_0)\left(1 - \frac{1-p_0}{(1-q_0)(1-\alpha)} P_1(g(x) = 1)\right)_+. \end{aligned}$$

Using that q_0 and $1 - q_0$ are positive lead to (5.4). □

To determine the relationship between $R(g)$ and $R_\alpha(g)$, we have only to determine the value of q_0 that minimizes (5.4).

Proposition 5.4. *For a given classification rule g the trimmed error attains its minimum value in*

$$q_0 \in \begin{cases} \left[1 - \frac{1-p_0}{1-\alpha} P_1(\{x : g(x) = 1\}), \frac{p_0}{1-\alpha} P_0(\{x : g(x) = 0\})\right] \bigcap \left[1 - \frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}\right] & \text{if } R(g) \leq \alpha \\ \left[\frac{p_0}{1-\alpha} P_0(\{x : g(x) = 0\}), 1 - \frac{1-p_0}{1-\alpha} P_1(\{x : g(x) = 1\})\right] \bigcap \left[1 - \frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}\right] & \text{if } R(g) > \alpha \end{cases}, \tag{5.6}$$

$$Q_0(\{x \in \mathbb{R}^d : g(x) = 1\}) = \left(1 - \frac{p_0}{q_0(1-\alpha)} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})\right)_+ \quad \text{and}$$
$$Q_1(\{x \in \mathbb{R}^d : g(x) = 0\}) = \left(1 - \frac{1-p_0}{(1-q_0)(1-\alpha)} P_1(\{x \in \mathbb{R}^d : g(x) = 1\})\right)_+ .$$

This minimum value is

$$R_\alpha(g) = \frac{1}{1-\alpha}(R(g) - \alpha)_+. \tag{5.7}$$

Proof. In Lemma 5.3 we have obtained the expression (5.4) of $R_\alpha(g)$ in which we only need to minimize for q_0.
First see that $R_\alpha(g) = 0$ if and only if

$$1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) \leq \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}).$$

As we are adding two positive terms, the sum is equal to 0 only if both terms are equal to 0, leading to

$$\left(q_0 - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})\right)_+ \leq 0 \;\Leftrightarrow\; q_0 \leq \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}),$$

in a similar way we obtain

$$\left(1 - q_0 - \frac{1-p_0}{1-\alpha} P_1(\{x : g(x) = 1\})\right)_+ \leq 0 \;\Leftrightarrow\; q_0 \geq 1 - \frac{1-p_0}{1-\alpha} P_1(\{x : g(x) = 1\}).$$

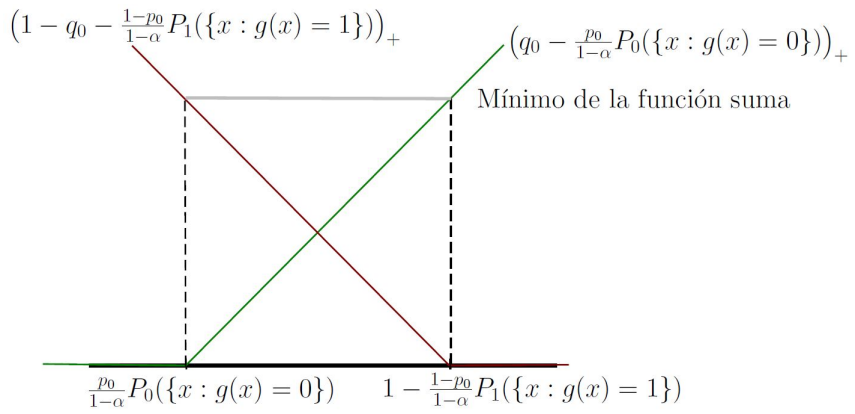
Hence $R_\alpha(g) = 0$ if and only if

$$1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) \leq \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$$

.
Now consider the case where this inequality does not hold, this means,

$$1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) > \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}).$$

The first term of (5.4) is a stepwise lineal function with value 0 up to $\frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$ and increasing with slope 1 from then on. The second term is also stepwise



linear, in this case it decreases with slope -1 until it reaches 0 in $1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\})$ with value 0 from that point on.
In the image you can see that in this case the interval

$$\left[\frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\}), 1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\})\right]$$

provides us with the minimum value of (5.4), in the case that $1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\}) < 1-\frac{1-p_0}{1-\alpha}$ or $\frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\}) > \frac{p_0}{1-\alpha}$ we would eliminate from the optimal set the unfeasible values of q_0, hence the set of values of q_0 that minimizes $R_\alpha(g)$ is (5.6). Let us see this. We are going to suppose, for simplicity, that we are in the case

$$\begin{aligned}1-\frac{1-p_0}{1-\alpha} &\leq \frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\})\\ &< 1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\}) \leq \frac{p_0}{1-\alpha}.\end{aligned}$$

Take

$$\begin{aligned}I_1 &= \left[1-\frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\})\right],\\ I_2 &= \left[\frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\}), 1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\})\right],\\ I_3 &= \left[1-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\}), \frac{p_0}{1-\alpha}\right],\end{aligned}$$

denote

$$\begin{aligned}R^i = \min_{q_0\in I_i}\Bigg[&\left(q_0-\frac{p_0}{1-\alpha}P_0(\{x\in\mathbb{R}^d : g(x)=0\})\right)_+\\ &+\left(1-q_0-\frac{1-p_0}{1-\alpha}P_1(\{x\in\mathbb{R}^d : g(x)=1\})\right)_+\Bigg],\end{aligned}$$

for $i=1,2,3$. Obviously $R_\alpha(g)=\min R^i$.

In I_1 the first term is 0 because $q_0 \leq \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$ and the second term is $1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) - q_0$. As we are looking for a minimization of this value and q_0 has a negative sign, we will give to it the biggest value it can take, that is, the upper bound of the interval. Hence,

$$\begin{aligned} R^1 &= 1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}) \\ &= 1 - \frac{(1-p_0)(1 - P_1(\{x \in \mathbb{R}^d : g(x) = 0\})) + p_0(1 - P_0(\{x \in \mathbb{R}^d : g(x) = 1\}))}{1-\alpha} \\ &= 1 - \frac{1 - R(g)}{1-\alpha}. \end{aligned}$$

If we are in I_2 none of the terms is going to be 0. The first one is $q_0 - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$ and the second one is $1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) - q_0$, when we add them, the q_0 in both terms cancels out and we obtain

$$\begin{aligned} R^2 &= 1 - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}) - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) \\ &= 1 - \frac{1 - R(g)}{1-\alpha}. \end{aligned}$$

Last, in I_3 it is the second term which becomes 0, while the first one is $q_0 - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$. In this case q_0 is positive and we want to give it the minimum value possible so

$$\begin{aligned} R^3 &= 1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) - \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\}) \\ &= 1 - \frac{1 - R(g)}{1-\alpha}. \end{aligned}$$

And, as we have already said, the minimum is attained at

$$\left[1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}), \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})\right].$$

Moreover, since $R^1 = R^2 = R^3$, the value of this minimum will be

$$R_\alpha(g) = 1 - \frac{1 - R(g)}{1-\alpha}.$$

Putting together both cases we have that $R_\alpha(g)$ attains its minimum in (5.6) and, since condition $1 - \frac{1-p_0}{1-\alpha} P_1(\{x \in \mathbb{R}^d : g(x) = 1\}) > \frac{p_0}{1-\alpha} P_0(\{x \in \mathbb{R}^d : g(x) = 0\})$ holds if and only if $R(g) > \alpha$, we have that $R_\alpha(g) = 0 \Leftrightarrow R(g) \leq \alpha$ and hence,

$$R_\alpha(g) = \frac{1}{1-\alpha}(R(g) - \alpha)_+.$$

□

As a result, we will see that the relationship established in Proposition 5.4 is maintained for $\mathrm{Err}_\alpha(P)$ and $\mathrm{Err}(P)$. In the proof we will use that, if we denote the Bayes classifier by g_B, we can write $\mathrm{Err}(P) = R(g_B)$.

Corollary 5.5. *With the previous notation, if* $\alpha \in [0,1)$

$$Err_\alpha(P) = \frac{(Err(P)-\alpha)_+}{1-\alpha}.$$

Proof: By definition

$$\begin{aligned}\mathrm{Err}_\alpha(P) &:= \inf_{Q\in\mathcal{R}_\alpha(P)} \mathrm{Err}(Q) = \inf_{Q\in\mathcal{R}_\alpha(P)} \inf_g Q(\{(y,x): g(x)\neq y\}) \\ &= \inf_g \inf_{Q\in\mathcal{R}_\alpha(P)} Q(\{(y,x): g(x)\neq y\}) = \inf_g R_\alpha(g) \\ &= \min_g \frac{(R(g)-\alpha)_+}{1-\alpha}.\end{aligned}$$

The minimum is attained (recall Proposition 5.4). We also know that this error is minimal for the Bayes classifier, hence

$$\mathrm{Err}_\alpha(P) = \frac{(R(g_B)-\alpha)_+}{1-\alpha} = \frac{(\mathrm{Err}(P)-\alpha)_+}{1-\alpha}.$$

□

From the equality in the previous corollary we can deduce that if $\mathrm{Err}(P) \leq \alpha$ then $\mathrm{Err}_\alpha(P) = 0$, but if $\mathrm{Err}(P) > \alpha$ then $\mathrm{Err}_\alpha(P) = \frac{(\mathrm{Err}(P)-\alpha)_+}{1-\alpha} > 0$, which indicates that

$$\mathrm{Err}_\alpha(P) = 0 \text{ if and only if } \mathrm{Err}(P) \leq \alpha,$$

That is, the minimum α that gives us a perfect separation is Bayes error.

Theorem 5.6. *There is* $Q_\alpha \in \mathcal{R}_\alpha(P)$ *such that* $Err_\alpha(P) = Err(Q_\alpha)$. *Besides* Q_α *is the probability from Proposition 5.4 with* $g = g_B$ (g_B *the original Bayes rule, which is also the trimmed Bayes rule).*

Proof: By definition

$$\mathrm{Err}_\alpha(P) = \inf_{Q\in\mathcal{R}_\alpha(P)} \mathrm{Err}(Q)$$

and

$$\mathrm{Err}(P) = \inf_g R(g) = \inf_g P(\{(y,x): g(x)\neq y\}).$$

Then

$$\mathrm{Err}_\alpha(P) = \inf_{Q\in\mathcal{R}_\alpha(P)} \inf_g Q(\{(y,x): g(x)\neq y\}) = \inf_g \inf_{Q\in\mathcal{R}_\alpha(P)} Q(\{(y,x): g(x)\neq y\}),$$

by definition again,

$$\inf_{Q\in\mathcal{R}_\alpha(P)} Q(\{(y,x) : g(x) \neq y\}) = \frac{(R(g)-\alpha)_+}{1-\alpha},$$

where the last equality comes from (5.7). Then

$$\mathrm{Err}_\alpha(P) = \inf_g \frac{(R(g)-\alpha)_+}{1-\alpha}$$

and, given that $f(x) = \frac{(x-\alpha)_+}{1-\alpha}$ is increasing for x, $\frac{(R(g)-\alpha)_+}{1-\alpha}$ will attain its minimum when $R(.)$ attains it, which happens when g is Bayes rule.
The value of Q_α is obtained from applying directly Proposition 5.4. □

As stated at the beginning, the interest in Bayes estimator is merely theoretical due to the impossibility of its calculation in practice. Usually we don't look for the optimal classifier among all possible classifiers but restrict ourselves to a smaller class. We will consider $\mathcal{G}$ a class of classifiers and $\hat{g} \in \mathcal{G}$ the classifier that gives us the minimum classification error in that class. In analogy to (2.3), we define the *trimmed generalization error of a class* $\mathcal{G}$ and denote it by $R_\alpha(\mathcal{G})$, this is, $R_\alpha(\mathcal{G}) := \min_{g\in\mathcal{G}} R_\alpha(g)$.

Corollary 5.7. *The relationship between the trimmed error and the generalization error is also maintained for the generalization error of a class:*

$$R_\alpha(\mathcal{G}) = \min_{g\in\mathcal{G}} \frac{(R(g)-\alpha)_+}{1-\alpha} = \frac{(R(\mathcal{G})-\alpha)_+}{1-\alpha},$$

This result is a direct consequence of Proposition 5.4. We can see that the classifier, g, which minimizes $R_\alpha(\mathcal{G})$ is the same that minimizes $R(\mathcal{G})$.

To calculate the generalization error we need to know the distribution of the training sample, which usually does not occur. Therefore, we have to estimate the value of R_α. The most natural choice for such estimation is given by the empirical trimmed error.

Definition 5.2. *We define the* empirical trimmed error *of a rule g as*

$$R_{n,\alpha}(g) := \min_{Q\in\mathcal{R}_\alpha(P_n)} Q(g(X) \neq Y) \tag{5.8}$$

where P_n is the empirical distribution associated with the training sample.

Observe that

$$R_{n,\alpha}(g) = \min_{W_\alpha} \sum_{j=1}^{n} w_i I_{(g(x_j)\neq y_j)} \tag{5.9}$$

with

$$W_\alpha = \{(w_1, \ldots, w_n)/\ 0 \leq w_i \leq \frac{1}{n(1-\alpha)};\ i = 1, \ldots, n\ \wedge\ \sum_{i=1}^{n} w_i = 1\}. \tag{5.10}$$

Proposition 5.4 can be trivially extended to the empirical error by using Definition 5.2.

Corollary 5.8. *Let g be a classifier, α a preset trimming level and $n \in \mathbb{N}$ the sample size,*

$$R_{n,\alpha}(g) = \frac{1}{1-\alpha}(R_n(g) - \alpha)_+. \tag{5.11}$$

As in Proposition 2.4, there is a bias/variance decomposition for the trimmed error. If we denote by $\hat{g}_{n,\alpha} = \arg\min_{g \in \mathcal{G}} R_{n,\alpha}(g)$, $g_B = \arg\min_g R_\alpha(g)$ and $\hat{g}_0 = \arg\min_{g \in \mathcal{G}} R_\alpha(g)$, the optimal empirical classifier of the trimmed sample, Bayes Classifier and the best classifier in the class for the trimmed sample (which we saw before are the same as for the complete sample) respectively (we assume for simplicity that $R_\alpha(g)$ has a unique minimizer in $\mathcal{G}$; the general case can be treated with minimal changes). Then the trimmed excess risk defined by $\mathcal{E}_\alpha(\mathcal{G}) := \min_{g \in \mathcal{G}} R_\alpha(g) - R_\alpha(g_B)$ can be bounded by the excess risk defined in Proposition 2.4,

$$\begin{aligned} \mathcal{E}_\alpha(\mathcal{G}) &= \frac{(R(\hat{g}_0) - \alpha)_+}{1-\alpha} - \frac{(R(g_B) - \alpha)_+}{1-\alpha} \\ &\leq \frac{R(\hat{g}_0) - R(g_B)}{1-\alpha} = \frac{\mathcal{E}(\mathcal{G})}{1-\alpha}, \end{aligned} \tag{5.12}$$

as a consequence we obtain the bias/variance decomposition.

Proposition 5.9. *Let α be a given trimming level, then*

$$\begin{aligned} R_\alpha(\hat{g}_{n,\alpha}) - R_\alpha(g_B) &\leq \frac{2}{1-\alpha} \sup_{g \in \mathcal{G}} |R_n(g) - R(g)| + \mathcal{E}_\alpha(\mathcal{G}) \\ &\leq \frac{1}{1-\alpha} \left[2 \sup_{g \in \mathcal{G}} |R_n(g) - R(g)| + \mathcal{E}(\mathcal{G}) \right]. \end{aligned}$$

Proof.

$$\begin{aligned} R_\alpha(\hat{g}_{n,\alpha}) - R_\alpha(g_B) &= (R_\alpha(\hat{g}_{n,\alpha}) - R_{n,\alpha}(\hat{g}_{n,\alpha})) + (R_{n,\alpha}(\hat{g}_{n,\alpha}) - R_\alpha(\hat{g}_\alpha)) \\ &+ (R_\alpha(\hat{g}_\alpha) - R_\alpha(g_B)) \\ &\leq \sup_{g \in \mathcal{G}} |R_{n,\alpha}(g) - R_\alpha(g)| + (R_{n,\alpha}(\hat{g}_\alpha) - R_\alpha(\hat{g}_\alpha)) + \mathcal{E}_\alpha(\mathcal{G}) \\ &\leq \sup_{g \in \mathcal{G}} |R_{n,\alpha}(g) - R_\alpha(g)| + \sup_{g \in \mathcal{G}} |R_{n,\alpha}(g) - R_\alpha(g)| + \mathcal{E}_\alpha(\mathcal{G}) \\ &\leq 2 \sup_{g \in \mathcal{G}} |R_{n,\alpha}(g) - R_\alpha(g)| + \mathcal{E}_\alpha(\mathcal{G}). \end{aligned} \tag{5.13}$$

The first term in (5.13), by Proposition 5.4 and Corollary 5.8 will be

$$\begin{aligned}\sup_{g\in\mathcal{G}}|R_{n,\alpha}(g)-R_\alpha(g)| &= \sup_{g\in\mathcal{G}}\left|\frac{1}{1-\alpha}\left(R_n(g)-\alpha\right)_+-\frac{1}{1-\alpha}\left(R(g)-\alpha\right)_+\right| \\ &\leq \sup_{g\in\mathcal{G}}\frac{1}{1-\alpha}\left|\left(R_n(g)-\alpha-\left(R(g)-\alpha\right)\right)\right| \\ &= \frac{1}{1-\alpha}\sup_{g\in\mathcal{G}}\left|R_n(g)-R(g)\right|.\end{aligned}$$

Replacing (5.12) in inequality (5.13) we obtain the desired result. □

There are analogous versions of Proposition 5.9 for the case in which we compare the risk of the best empirical rule with the optimal risk within a restricted class. In the classic case (without trimmings)

$$R(\hat{g}_n)-R(\hat{g}_0)\leq 2\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|,$$

see Lugosi (2002). By means of the following proposition we adapt this inequality to the trimmed errors.

Proposition 5.10. *For a given trimming level* α

$$R_\alpha(\hat{g}_{n,\alpha})-R_\alpha(\hat{g}_0)\leq\frac{2}{1-\alpha}\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|.$$

Proof.

$$\begin{aligned}R_\alpha(\hat{g}_{n,\alpha})-R_\alpha(\hat{g}_0) &= R_\alpha(\hat{g}_{n,\alpha})-R_{n,\alpha}(\hat{g}_{n,\alpha})+R_{n,\alpha}(\hat{g}_{n,\alpha})-R_\alpha(\hat{g}_0) \\ &\leq \sup_{g\in\mathcal{G}}|R_{n,\alpha}(g)-R_\alpha(g)|+R_{n,\alpha}(\hat{g}_0)-R_\alpha(\hat{g}_0) \\ &\leq 2\sup_{g\in\mathcal{G}}|R_{n,\alpha}(g)-R_\alpha(g)| \\ &= 2\sup_{g\in\mathcal{G}}\left|\frac{(R_n(g)-\alpha)_+}{1-\alpha}-\frac{(R(g)-\alpha)_+}{1-\alpha}\right| \\ &\leq \frac{2}{1-\alpha}\sup_{g\in\mathcal{G}}|R_n(g)-\alpha-(R(g)-\alpha)| \\ &= \frac{2}{1-\alpha}\sup_{g\in\mathcal{G}}|R_n(g)-R(g)|.\end{aligned}$$

□

It is well known (see, for example, Lugosi (2002)) that $E(R_n(g))=R(g)$, but this equality does not generally hold when working with trimmed errors, although we can find a relationship between both quantities. In the following proposition we will see that the empirical trimmed error is an estimator with small positive bias that tends to 0 when the sample size increases.

Proposition 5.11. *For a given trimming level α and a given classifier g,*

$$0 \leq E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)}.$$

Proof. The first inequality can be proved by

$$\begin{aligned} E(R_{n,\alpha}(g)) &= E\left(\frac{(R_n(g)-\alpha)_+}{1-\alpha}\right) = \frac{1}{1-\alpha}E\left((R_n(g)-\alpha)_+\right) \\ &\geq \frac{1}{1-\alpha}\left(E(R_n(g))-\alpha\right)_+ = \frac{1}{1-\alpha}\left(R(g)-\alpha\right)_+ = R_\alpha(g), \end{aligned}$$

where we have used (5.11) for the first equality and the property $E(R_n(g)) = R(g)$ and (5.7) for the two last ones. The inequality comes from applying Jensen inequality, and this is possible due to the fact that $(.)_+$ is a convex function.
For the second inequality we need Proposition 5.4 and by Corollary 5.8,

$$E(R_{n,\alpha}(g)) - R_\alpha(g) = \frac{E((R_n(g)-\alpha)_+) - (R(g)-\alpha)_+}{(1-\alpha)} \tag{5.14}$$

Denoting $X = R_n(g)$, then $E(X) = R(g)$. Take $\varphi(x) = (x-\alpha)_+$. φ is a convex function, so Jensen's inequality can be applied to get $\varphi(E(X)) \leq E(\varphi(X))$. This function also is 1-Lipschitz and increasing, so it satisfies the property $\varphi(y) - \varphi(x) \leq (y-x)_+$. As consequence we get

$$\begin{aligned} E((R_n(g)-\alpha)_+) - (R(g)-\alpha)_+ &= E(\varphi(X)) - \varphi(E(X)) \\ &= E(\varphi(X) - \varphi(E(X))) \leq E((X-E(X))_+). \end{aligned}$$

Let now Y be an independent copy of X. If we denote by E_X the expected value conditioned by X, then by independence we will have $E(Y) = E_X(Y)$. This implies,

$$\begin{aligned} E((X-E(X))_+) &= E((X-E(Y))_+) = E((X-E_X(Y))_+) \\ &= E((E_X(X-Y))_+) \leq E(E_X((X-Y)_+)) = E((X-Y)_+). \end{aligned}$$

Now, as $X - Y$ is a symmetric and centered random variable,

$$\begin{aligned} E((X-Y)_+) &= \frac{1}{2}E(|X-Y|) \leq \frac{1}{2}(Var(X-Y))^{1/2} \\ &= \frac{1}{2}(Var(X)+Var(Y))^{1/2} = \frac{1}{\sqrt{2}}(Var(X))^{1/2}. \end{aligned}$$

Finally, we observe that $nX \sim b(n, R(g))$ and, hence,

$$\begin{aligned} \frac{1}{\sqrt{2}}(Var(X))^{1/2} &= \frac{1}{\sqrt{2}}(\frac{1}{n^2}Var(nX))^{1/2} = \frac{1}{\sqrt{2}}(\frac{1}{n^2}nR(g)(1-R(g)))^{1/2} \\ &= \frac{1}{\sqrt{2n}}\sqrt{R(g)}. \end{aligned}$$

This, together with (5.14) gives

$$E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)}.$$

□

5.2.1 Optimal trimming level selection

Trimmed models allow us to reduce classification error so that the loss of information because we do not use the entire sample can be quantified and controlled. As in any robust procedure, our aim is to select the amount of information we are going to remove, which in our problem corresponds to selecting the optimal trimming level. In fact, our goal is to find a data driven $\hat{\alpha}$ so that the classification error is minimized without having to remove too much information from the initial distribution.

It is obvious that the more we trim, the smaller the generalization error will be, but less information our model will have. To find a balance we are going to introduce a penalization. In order to facilitate the understanding of the role of this penalization, we will initially consider a simple case in which the class of admissible rules has only one element. Then the penalization will only be related to the level of trimming. We will then deal with the most realistic case in which the class of admissible rules is greater and even the case in which there are different types of admissible rules with different levels of complexity. In the simple class case our main result is the following.

Theorem 5.12. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P which take values in $\{0,1\} \times \mathbb{R}^d$. Let g be a given classifier. Take $\alpha_{max} \in [0,1)$. If we consider the penalization function*

$$pen(\alpha) = \frac{1}{(1-\alpha)}\sqrt{\frac{\ln(n)}{2n}}$$

and define

$$\hat{\alpha} = \underset{\alpha \in [0,\alpha_{max}]}{\arg\min}\ R_{n,\alpha}(g) + pen(\alpha),$$

the following bound holds,

$$\begin{aligned} E(R_{\hat{\alpha}}(g)) \quad &\leq \quad \inf_{\alpha \in [0,\alpha_{max}]} \left(R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1-\alpha)} \right) \\ &+ \quad \frac{1}{(1-\alpha_{max})}\sqrt{\frac{2\pi}{n}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned} \tag{5.15}$$

This theorem allows us to understand the effect of trimming on the classification error. For a given classifier g we set a maximum trimming level of α_{max} that we do not want to exceed. This level would have to be set by the investigator and would represent the maximum permissible level of deviation in terms of contamination from the sample, or the original generator of the data. Then the automatic penalized rule for choosing the trimming level leads to an oracle inequality that guarantees that the best classification error is achieved in essence. Similarly to model selection rules (see Massart (2007)) the price to pay is a term of order $1/\sqrt{n}$ which does not hamper the classification error. In particular, if classifier g has a small classification error, in the sense that $R(g)$ is smaller than $\alpha < \alpha_{max}$, we can eliminate the incorrectly classified elements which leads to a null trimmed classification error. In this case, if $R(g) < \alpha_{max}$, we could bound (5.15) by

$$\frac{1}{1-\alpha_{max}}\sqrt{\frac{\ln(n)}{2n}} + \frac{\sqrt{\alpha_{max}}}{\sqrt{n}(1-\alpha_{max})} + \frac{1}{(1-\alpha_{max})}\sqrt{\frac{2\pi}{n}} + \frac{1}{n(1-\alpha_{max})^2},$$

and then, for a constant C,

$$E(R_{\hat{\alpha}}(g)) \leq \frac{C}{\sqrt{n}}.$$

The proof of the above theorem is based on the following technical result.

Proposition 5.13. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P that take values in $\{0,1\}\times\mathbb{R}^d$. Let g be a given classifier. Let $k_0 < n$ be a natural number and A the set of possible trimmings, $A = \{0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{k_0}{n}\}$. If we consider the penalization function*

$$pen(\alpha) = \frac{1}{(1-\alpha)}\sqrt{\frac{\ln(n)}{2n}}$$

and define

$$\hat{\alpha} = \arg\min_{\alpha\in A} R_{n,\alpha}(g) + pen(\alpha),$$

the following bound holds,

$$E(R_{\hat{\alpha}}(g)) \leq \inf_{\alpha\in A}\left(R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1-\alpha)}\right) + \frac{1}{(1-\frac{k_0}{n})}\sqrt{\frac{2\pi}{n}}.$$

Proof. By definition $\hat{\alpha}$ satisfies that $\forall\alpha \in A$

$$R_{n,\hat{\alpha}}(g) + pen(\hat{\alpha}) \leq R_{n,\alpha}(g) + pen(\alpha).$$

Starting from this inequality, as we want to look for bounds for the theoretical trimmed error and this does not appear in the basic equation, we will introduce it adding and subtracting in both terms. Hence,

$$R_{\hat{\alpha}}(g) - R_{\hat{\alpha}}(g) + R_{n,\hat{\alpha}}(g) + pen(\hat{\alpha}) \leq R_\alpha(g) - R_\alpha(g) + R_{n,\alpha}(g) + pen(\alpha).$$

or equivalently

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + (R_{n,\alpha}(g) - R_{\alpha}(g)) - pen(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)). \quad (5.16)$$

Let us focus now on the expression inside the parenthesis,

$$R_{n,\alpha}(g) - R_{\alpha}(g) = [R_{n,\alpha}(g) - E(R_{n,\alpha}(g))] + [E(R_{n,\alpha}(g)) - R_{\alpha}(g)]$$

by Proposition 5.11 the second bracket can be bounded by $\frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)}$. For the first bracket we will apply McDiarmid's inequality (see Massart (2007)) taking $R_{n,\alpha}(g) = F(\xi_1, \ldots, \xi_n)$ where $\xi_i = (Y_i, X_i)$. As

$$|F(\xi_1, \ldots, \xi_i, \ldots, \xi_n) - F(\xi_1, \ldots, \xi_i', \ldots, \xi_n)| \leq \frac{1}{n(1-\alpha)}$$

we can apply the inequality and hence

$$P(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq t) \leq e^{-2t^2 n(1-\alpha)^2}.$$

Let $z > 0$ be given, taking $t = \sqrt{\frac{z}{2n(1-\alpha)^2}}$, we get

$$P\left(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq \sqrt{\frac{z}{2n(1-\alpha)^2}}\right) \leq e^{-z}.$$

Combining this with (5.16), we get that except in a set of probability at most e^{-z}

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - pen(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)). \quad (5.17)$$

Let us consider now $R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)$. Applying again McDiarmid's inequality but taking this time $t = \sqrt{\frac{\ln(n)+z}{2n(1-\alpha)^2}}$ we have $\forall \alpha' \in A$

$$P\left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}}\right) \leq \frac{1}{n}e^{-z}. \quad (5.18)$$

On the other side,

$$\begin{aligned}
&P\left(R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\hat{\alpha})^2}}\right) \\
&= P\left(R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) - \sqrt{\frac{\ln(n)+z}{2n(1-\hat{\alpha})^2}} \geq 0\right) \\
&\leq P\left(\sup_{\alpha' \in A}\left(R_{\alpha'}(g) - R_{n,\alpha'}(g) - \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}}\right) \geq 0\right) \\
&\leq \sum_{\alpha' \in A} P\left(R_{\alpha'}(g) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}}\right) \\
&\leq \sum_{\alpha' \in A} P\left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}}\right) \leq n\frac{1}{n}e^{-z} \leq e^{-z}.
\end{aligned}$$

So with probability at least $1 - e^{-z}$

$$R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \leq \sqrt{\frac{\ln(n) + z}{2n(1-\hat{\alpha})^2}} \leq \sqrt{\frac{\ln(n)}{2n(1-\hat{\alpha})^2}} + \sqrt{\frac{z}{2n(1-\hat{\alpha})^2}}.$$

If we take as penalization

$$pen(\alpha) = \sqrt{\frac{\ln(n)}{2n(1-\alpha)^2}}$$

and replace it in (5.17), we get that except in a set of probability at most $2e^{-z}$

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} + \sqrt{\frac{z}{2n(1-\hat{\alpha})^2}}.$$

Moving to the left side the terms that do not depend on z and bounding $\frac{1}{1-\alpha}$ and $\frac{1}{1-\hat{\alpha}}$ by $\frac{1}{1-\frac{k_0}{n}}$ we get

$$R_{\hat{\alpha}}(g) - R_{\alpha}(g) - pen(\alpha) - \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} \leq \sqrt{\frac{2z}{n(1-\frac{k_0}{n})^2}},$$

integrating this with respect to z leads us to

$$E\left(R_{\hat{\alpha}}(g) - R_{\alpha}(g) - pen(\alpha) - \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)}\right) \leq \frac{1}{1-\frac{k_0}{n}}\sqrt{\frac{2\pi}{n}}.$$

Or equivalently, $\forall \alpha \in A$

$$E(R_{\hat{\alpha}}(g)) \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + \frac{1}{1-\frac{k_0}{n}}\sqrt{\frac{2\pi}{n}},$$

by taking the infimum on the right side of the inequality we get (5.23). □

The last ingredient necessary for the proof of the theorem is the following proposition which states that when two trimming levels are close, the difference between the error of generalization associated with each of them (either theoretical or empirical) is small.

Proposition 5.14. *Let α_1, α_2 be two trimming levels such that $\alpha_2 \in [\alpha_1, \alpha_1 + \frac{1}{n}]$, let α_{max} be such that $\alpha_1 \leq \alpha_2 \leq \alpha_{max} < 1$ and let g be a given classifier, then*

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) \leq \frac{1}{n(1-\alpha_{max})^2} \text{ and } R_{n,\alpha_1}(g) - R_{n,\alpha_2}(g) \leq \frac{1}{n(1-\alpha_{max})^2}.$$

Proof. We will prove this only for the theoretical loss function, the proof is identical for the empirical function.

$$\begin{aligned}
R_{\alpha_1}(g) - R_{\alpha_2}(g) &= \frac{(R(g)-\alpha_1)_+}{(1-\alpha_1)} - \frac{(R(g)-\alpha_2)_+}{(1-\alpha_2)} \\
&= \frac{((1-\alpha_2)(R(g)-\alpha_1))_+ - ((1-\alpha_1)(R(g)-\alpha_2))_+}{(1-\alpha_1)(1-\alpha_2)} \\
&\leq \frac{1}{(1-\alpha_1)(1-\alpha_2)}|R(g) - \alpha_1 - \alpha_2 R(g) + \alpha_1\alpha_2 \\
&- (R(g) - \alpha_2 - \alpha_1 R(g) + \alpha_1\alpha_2)| \\
&= \frac{1}{(1-\alpha_1)(1-\alpha_2)}| - \alpha_1 - \alpha_2 R(g) + \alpha_2 + \alpha_1 R(g)| \\
&= \frac{1}{(1-\alpha_1)(1-\alpha_2)}|(R(g)-1)(\alpha_1-\alpha_2)| \\
&= \frac{1}{(1-\alpha_1)(1-\alpha_2)}|R(g)-1||\alpha_1-\alpha_2|
\end{aligned}$$

Given that $|\alpha_1 - \alpha_2| \leq \frac{1}{n}$ and for each value of α, $\frac{1}{1-\alpha} \leq \frac{1}{1-\alpha_{max}}$, we can conclude that

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) \leq \frac{1}{n(1-\alpha_{max})^2}.$$

□

With this proposition we are already in a position to prove Theorem 5.12.

Proof of Theorem 5.12. Take $A = \{0, \frac{1}{n}, \ldots, \frac{k_0}{n}\}$ with $k_0 = [n\alpha_{max}]$. We will reason as in the proof of Proposition 5.13. Set $\alpha \in [0, \alpha_{max}]$, let $z > 0$ be given. From the basic inequality

$$R_{n,\hat{\alpha}}(g) + pen(\hat{\alpha}) \leq R_{n,\alpha}(g) + pen(\alpha).$$

one can deduce, as in (5.17), that, with probability at least $1 - e^{-z}$

$$R_{\hat{\alpha}}(g) \leq R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - pen(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g))$$

and likewise for each α' in A

$$P\left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}}\right) \leq \frac{1}{n}e^{-z},$$

which implies that in a set of probability at least $1 - e^{-z}$

$$E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \leq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} \quad \text{for all } \alpha' \in A. \tag{5.19}$$

If $\alpha' \in [0, \alpha_{max}]$ then there is $\alpha'' \in A$ such that $\alpha'' \leq \alpha' \leq \alpha'' + \frac{1}{n}$. Then by Proposition 5.14, in the set where it is satisfied (5.19) we have

$$\begin{aligned} E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) &= E(R_{n,\alpha''}(g)) - R_{n,\alpha''}(g) + E(R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) \\ &- (R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) \\ &\leq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha'')^2}} + \frac{1}{n(1-\alpha_{max})^2} \\ &\leq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \\ &\leq \sqrt{\frac{\ln(n)}{2n(1-\alpha')^2}} + \sqrt{\frac{z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \end{aligned}$$

for all $\alpha' \in [0, \alpha_{max}]$. Since, by Proposition 5.11, $R_{\alpha'}(g) \leq E(R_{n,\alpha'}(g))$, we conclude that, with probability at least $1 - 2e^{-z}$

$$\begin{aligned} R_{\hat{\alpha}}(g) &\leq R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - pen(\hat{\alpha}) \\ &+ \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \\ &\leq R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n}(1-\alpha)} + 2\sqrt{\frac{z}{2n(1-\alpha_{max})^2}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

Integrating with respect to z and taking the infimum for $\alpha \in [0, \alpha_{max}]$ we conclude that

$$\begin{aligned} E(R_{\hat{\alpha}}(g)) &\leq \inf_{\alpha \in [0,\alpha_{max}]} \left(R_\alpha(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1-\alpha)} \right) \\ &+ \frac{1}{(1-\alpha_{max})} \sqrt{\frac{2\pi}{n}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

□

A more realistic case is the situation in which we consider a class of classification rules from which we will choose the optimal rule. Moreover, it is common for a rule to be searched within a class or model and different candidate models to be considered. A complex class will usually lead us to rules with a small bias in the sense that they classify the elements in the training sample well at the expense of a large variance, which usually leads to an over-adjustment of the classification model (see bound (2.5)). In this new partial classification context, to control the complexity of the model, penalizations will now depend not only on the trimming level but also on the complexity of the classifier class. To measure the complexity of the model we will use the Vapnik-Chervonenkis

dimension introduced in section 2.2.1. Next we will prove an extension of Theorem 5.12 in this more realistic situation. In this theorem, penalizations are considered to be a function of the trimming level, the complexity of the class and a collection of weights that can be chosen in a very general way. Subsequently, in Example 5.1 we show a possible choice of these weights for the case of linear classifiers.

Theorem 5.15. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P which take values in $\{0,1\} \times \mathbb{R}^d$. Let $\{\mathcal{G}_m\}_{m\in\mathbb{N}}$ be a family of classifier classes with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$. Take $\alpha_{max} \in [0,1)$. Let Σ be a non-negative constant and consider $\{x_m\}_{m\in\mathbb{N}}$ a family of non-negative weights such that*

$$\sum_{m\in\mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

If we consider the penalization function

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n) + x_m}{2n(1-\alpha)^2}} + \frac{2}{(1-\alpha)}\sqrt{\frac{V_{\mathcal{G}_m}\ln(n+1) + \ln(2)}{n}}$$

and define

$$(\hat{\alpha}, \hat{m}) = \underset{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}}{\arg\min} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m),$$

the following bound holds:

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \quad &\leq \quad \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) \\ &+ \quad \frac{1+\Sigma}{2(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

The proof of this theorem is based on that of Proposition 5.17 below. In turn, in the proof of this proposition we will use the following elementary technical lemma.

Lemma 5.16. *Given two functions f and g and a real number $k > 0$,*

$$|f(x) - g(x)| \leq k \;\Rightarrow\; |\sup_x f(x) - \sup_x g(x)| \leq k,$$

$$|f(x) - g(x)| \leq k \;\Rightarrow\; |\min_x f(x) - \min_x g(x)| \leq k.$$

Proposition 5.17. *Let $\xi_1, \dots, \xi_n$ be n independent and identically distributed observations with distribution P which take values in $\{0,1\} \times \mathbb{R}^d$. Let $\{\mathcal{G}_m\}_{m\in\mathbb{N}}$ be a family of classifier classes with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$. Let $k_0 < n$ be a natural number and A the set of possible trimming values, $A = \{0, \frac{1}{n}, \frac{2}{n}, \dots \frac{k_0}{n}\}$. Let Σ be a non-negative constant and consider $\{x_m\}_{m\in\mathbb{N}}$ a family of non-negative weights such that*

$$\sum_{m\in\mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

If we consider the penalization function

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n) + x_m}{2n(1-\alpha)^2}} + \frac{2}{(1-\alpha)}\sqrt{\frac{V_{\mathcal{G}_m}\ln(n+1) + \ln(2)}{n}}$$

and define

$$(\hat{\alpha}, \hat{m}) = \underset{(\alpha,m)\in A\times\mathbb{N}}{\arg\min}\, R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) \quad \textit{and} \quad g' := \underset{g\in\mathcal{G}_m}{\arg\min}\, R_\alpha(g),$$

the following bound holds:

$$E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \leq \min_{(\alpha,m)\in A\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(g')}}{\sqrt{2n}(1-\alpha)} \right) + \frac{1+\Sigma}{2(1-\frac{k_0}{n})}\sqrt{\frac{\pi}{2n}}.$$

Proof. By definition $\hat{\alpha}$ and $\hat{m}$ satisfy that $\forall\alpha \in A$ and $\forall m \in \mathbb{N}$

$$R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) + pen(\hat{\alpha}, \mathcal{G}_{\hat{m}}) \leq R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m).$$

Adding and subtracting $R_{\hat{\alpha}}(\mathcal{G}_m)$ and $R_\alpha(\mathcal{G}_m)$ and by rearranging the terms we obtain the following inequality:

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + (R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m)) \\ &- pen(\hat{\alpha}, \mathcal{G}_{\hat{m}}) + (R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})). \end{aligned}$$

First we want to bound

$$R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m) = \min_{g\in\mathcal{G}_m} R_{n,\alpha}(\mathcal{G}_m) - \min_{g\in\mathcal{G}_m} R_\alpha(\mathcal{G}_m) \leq R_{n,\alpha}(g') - R_\alpha(g')$$

with $g' := \underset{g\in\mathcal{G}_m}{\arg\min}\, R_\alpha(g)$. We are now in the same condition as in Proposition 5.13 and we can bound these quantities except in a set of probability at most e^{-z}, with $z > 0$ given, by

$$R_{n,\alpha}(g') - R_\alpha(g') \leq \frac{R(g')}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}},$$

which leads us to

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(g')}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} \\ &- pen(\hat{\alpha}, \mathcal{G}_{\hat{m}}) + (R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})). \end{aligned} \tag{5.20}$$

Now we want to bound

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq \sup_{(\alpha',m')\in A\times\mathbb{N}} (R_{\alpha'}(\mathcal{G}_{m'}) - R_{n,\alpha'}(\mathcal{G}_{m'})) \\ &\leq \sup_{(\alpha',m')\in A\times\mathbb{N}} \sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)). \end{aligned}$$

Let us focus on

$$\sup_{g\in\mathcal{G}_m} (R_{\alpha}(g) - R_{n,\alpha}(g)) = E\left(\sup_{g\in\mathcal{G}_m} (R_{\alpha}(g) - R_{n,\alpha}(g))\right) \tag{5.21}$$

$$+ \left[\sup_{g\in\mathcal{G}_m} (R_{\alpha}(g) - R_{n,\alpha}(g)) - E\left(\sup_{g\in\mathcal{G}_m} (R_{\alpha}(g) - R_{n,\alpha}(g))\right)\right]. \tag{5.22}$$

In order to bound (5.22) we will use McDiarmid's inequality, first we need to see that it fulfills the condition of the bounded differences.

We define $Z := f(\xi_1, \ldots, \xi_n) = \sup_{g\in\mathcal{G}_m} (R_{\alpha}(g) - R_{n,\alpha}(g))$ and $Z^{(i)} := f(\xi_1, \ldots, \xi_i', \ldots, \xi_n)$, we want to prove that

$$|Z - Z^{(i)}| \leq c_i, \tag{5.23}$$

for certain constants c_i. The empirical error $R_{n,\alpha}(g)$ is defined as in (5.9) and we also define $R^{(i)}_{n,\alpha}(g)$ as the empirical error associated with the sample $\xi_1, \ldots, \xi_i', \ldots, \xi_n$. We start from

$$|(R_{\alpha}(g) - R_{n,\alpha}(g)) - (R_{\alpha}(g) - R^{(i)}_{n,\alpha}(g))|$$

to come, by applying Lemma 5.16, to (5.23).

$$|R_{n,\alpha}(g) - R^{(i)}_{n,\alpha}(g)| = \left| \min_{(w_1,\ldots,w_n)} \sum_j w_j I_{(g(X_j)\neq Y_j)} - \min_{(w_1,\ldots,w_n)} \sum_j w_j I_{(g(X'_j)\neq Y'_j)} \right|,$$

where (Y', X') refers to the sample $\xi_1, \ldots, \xi_i', \ldots, \xi_n$. For a vector satisfying the conditions of (5.9), $(w_1, \ldots, w_n)$,

$$\left| \sum_j w_j I_{(g(X_j)\neq Y_j)} - \sum_j w_j I_{(g(X'_j)\neq Y'_j)} \right| = w_j \left| (I_{(g(X_i)\neq Y_i)} - I_{g(X'_i)\neq Y'_i)} \right| \leq \frac{1}{n(1-\alpha)}.$$

And by using the second implication in Lemma 5.16 we come to

$$|R_{n,\alpha}(g) - R^{(i)}_{n,\alpha}(g)| \leq \frac{1}{n(1-\alpha)},$$

or, equivalently,

$$|(R_\alpha(g) - R_{n,\alpha}(g)) - (R_\alpha(g) - R^{(i)}_{n,\alpha}(g))| \leq \frac{1}{n(1-\alpha)}.$$

Applying Lemma 5.16 again, we get (5.23) with $c_i = \frac{1}{n(1-\alpha)}$. Now we can use McDiarmid's inequality to prove that

$$\begin{aligned} &P\left(\sup_{g\in\mathcal{G}_m}(R_\alpha(g) - R_{n,\alpha}(g)) - E(\sup_{g\in\mathcal{G}_m}(R_\alpha(g) - R_{n,\alpha}(g))) \geq \sqrt{\frac{\ln(n) + z + x_m}{2n(1-\alpha)^2}}\right) \\ &\leq \frac{1}{n}e^{-z-x_m}. \end{aligned} \tag{5.24}$$

To bound (5.21) we will use Vapnik-Chervonenkis' theory. Before we can apply this theory we need to use equalities (5.7) and (5.11) and that the positive part function is Lipschitz to transform our functions into proper functions.

$$\begin{aligned} E\left(\sup_{g\in\mathcal{G}_m}(R_\alpha(g) - R_{n,\alpha}(g))\right) &= \frac{1}{1-\alpha}E\left(\sup_{g\in\mathcal{G}_m}((R(g)-\alpha)_+ - (R_n(g)-\alpha)_+)\right) \\ &\leq \frac{1}{1-\alpha}E\left(\sup_{g\in\mathcal{G}_m}|R(g) - R_n(g)|\right) \\ &\leq \frac{2}{1-\alpha}\sqrt{\frac{V_{\mathcal{G}_m}\ln(n+1) + \ln(2)}{n}}. \end{aligned} \tag{5.25}$$

The latest inequality comes from (2.8). By putting together (5.24) and (5.25) we obtain $\forall\alpha' \in A$ and $\forall m' \in \mathbb{N}$

$$\begin{aligned} &P\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g) - R_{n,\alpha'}(g)) \geq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1) + \ln(2)}{n}}\right) \\ &\leq \frac{1}{n}e^{-z-x_{m'}}. \end{aligned} \tag{5.26}$$

Therefore,

$$
\begin{aligned}
&P\left(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - \sqrt{\frac{\ln(n)+z+x_{\hat{m}}}{2n(1-\hat{\alpha})^2}} - \frac{2}{1-\hat{\alpha}}\sqrt{\frac{V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)}{n}} \geq 0\right) \\
&\leq P\left(\sup_{(\alpha',m')\in A\times\mathbb{N}}\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g) - R_{n,\alpha'}(g)) - \sqrt{\frac{\ln(n)+z+x_{m'}}{2n(1-\alpha')^2}}\right.\right. \\
&\left.\left. -\frac{2}{1-\alpha'}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1)+\ln(2)}{n}}\right) \geq 0\right) \\
&\leq \sum_{\alpha'\in A}\sum_{m'\in\mathbb{N}} P\left(R_{\alpha'}(g) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z+x_{m'}}{2n(1-\alpha')^2}}\right. \\
&\left. +\frac{2}{1-\alpha'}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1)+\ln(2)}{n}}\right) \\
&\leq \sum_{\alpha'\in A}\sum_{m'\in\mathbb{N}}\frac{1}{n}e^{-z-x_{m'}} \leq \sum_{m'\in\mathbb{N}} e^{-z-x_{m'}} \leq \Sigma e^{-z}.
\end{aligned}
$$

And we can conclude that with probability at least $1-\Sigma e^{-z}$

$$
\begin{aligned}
R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq \sqrt{\frac{\ln(n)+z+x_{\hat{m}}}{2n(1-\hat{\alpha})^2}} + \frac{2}{1-\hat{\alpha}}\sqrt{\frac{V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)}{n}} \\
&\leq \sqrt{\frac{\ln(n)+x_{\hat{m}}}{2n(1-\hat{\alpha})^2}} + \sqrt{\frac{z}{2n(1-\hat{\alpha})^2}} \\
&+ \frac{2}{1-\hat{\alpha}}\sqrt{\frac{V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)}{n}}.
\end{aligned}
$$

Returning to (5.20), except in a set of probability not greater than $(\Sigma+1)e^{-z}$

$$
\begin{aligned}
R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha,\mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} \\
&- pen(\hat{\alpha},\mathcal{G}_{\hat{m}}) + \sqrt{\frac{\ln(n)+x_{\hat{m}}}{2n(1-\hat{\alpha})^2}} + \sqrt{\frac{z}{2n(1-\hat{\alpha})^2}} \\
&+ \frac{2}{1-\hat{\alpha}}\sqrt{\frac{V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)}{n}}.
\end{aligned}
$$

Considering

$$
pen(\alpha,\mathcal{G}_m) = \sqrt{\frac{\ln(n)+x_m}{2n(1-\alpha)^2}} + \frac{2}{(1-\alpha)}\sqrt{\frac{V_{\mathcal{G}_m}\ln(n+1)+ln(2)}{n}},
$$

we have

$$
\begin{aligned}
R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha,\mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} + \sqrt{\frac{z}{2n(1-\hat{\alpha})^2}} \\
&\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha,\mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{2z}{n(1-\frac{k_0}{n})^2}}.
\end{aligned}
$$

Now proceeding as we did in Proposition 5.13, grouping and integrating with respect to z,

$$E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \leq \min_{(\alpha,m)\in A\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) + \frac{1+\Sigma}{2(1-\frac{k_0}{n})}\sqrt{\frac{\pi}{2n}}.$$

□

Proof of Theorem 5.15. This proof is identical to that of Proposition 5.17 until (5.26). To bound $R_{\hat{\alpha}(\mathcal{G}_{\hat{m}})} - R_{n,\hat{\alpha}(\mathcal{G}_{\hat{m}})}$, we define the set $A = \{0, \frac{1}{n}, \frac{2}{n}, \dots \frac{k_0}{n}\}$ with $k_0 = [n\alpha_{max}]$. If $\alpha' \in [0, \alpha_{max}]$, then $\exists \alpha'' \in A$ such that $\alpha'' \leq \alpha' \leq \alpha'' + \frac{1}{n}$. By Theorem 5.17 we know that with probability at least $1 - \frac{1}{n}e^{-z-x_m}$ $\forall \alpha'' \in A$

$$\sup_{g\in\mathcal{G}_{m'}} \left(R_{\alpha''}(g) - R_{n,\alpha''}(g)\right) \leq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha'')^2}} + \frac{2}{1-\alpha''}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1) + \ln(2)}{n}},$$

then for all $\alpha' \in [0, \alpha_{max}]$

$$\begin{aligned}
&\sup_{g\in\mathcal{G}_{m'}} \left(R_{\alpha'}(g) - R_{n,\alpha'}(g)\right) \\
&= \sup_{g\in\mathcal{G}_{m'}} \left(R_{\alpha'}(g) - R_{n,\alpha'}(g) + R_{\alpha''}(g) - R_{\alpha''}(g) + R_{n,\alpha''}(g) - R_{n,\alpha''}(g)\right) \\
&= \sup_{g\in\mathcal{G}_{m'}} \left([R_{\alpha''}(g) - R_{n,\alpha''}(g)] + [R_{\alpha'}(g) - R_{\alpha''}(g)] + [R_{n,\alpha''}(g) - R_{n,\alpha'}(g)]\right) \\
&\leq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha'')^2}} + \frac{2}{1-\alpha''}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1-\alpha_{max})^2} \\
&\leq \sqrt{\frac{\ln(n) + x_{m'}}{2n(1-\alpha')^2}} + \sqrt{\frac{z}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'}\sqrt{\frac{V_{\mathcal{G}_{m'}}\ln(n+1) + \ln(2)}{n}} \\
&+ \frac{1}{n(1-\alpha_{max})^2}.
\end{aligned}$$

Where the penultimate inequality comes from applying Proposition 5.14 and from $R_{n,\alpha'}(g) \leq R_{n,\alpha''}(g)$ and hence $R_{n,\alpha'}(g) - R_{n,\alpha''}(g) \leq 0$ and the last one from $\alpha'' \leq \alpha'$. We can conclude that

$$R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq \sqrt{\frac{\ln(n) + z + x_{\hat{m}}}{2n(1-\hat{\alpha})^2}} + \frac{2}{1-\hat{\alpha}}\sqrt{\frac{V_{\mathcal{G}_{\hat{m}}}\ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1-\alpha_{max})^2}.$$

In the same way as in Proposition 5.17 we get the bound

$$\begin{aligned}
E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \quad &\leq \quad \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) \\
&+ \quad \frac{1+\Sigma}{2(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{max})^2}.
\end{aligned}$$

□

Example 5.1. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. observations where $X_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$. We are going to consider $\{\mathcal{G}_m\}_{m\in\{1,\ldots,p\}}$ the model collection where for each m, $\mathcal{G}_m$ is the class of linear classifiers built using only a selection of variables formed by the first m components of X_i. We call $X^{(i)} = [X_1^{(i)} \ldots X_i^{(i)}]^T$ if $X_j^{(i)} = [X_{j1} \ldots X_{ji}]^T$ and we define the set of possible classifiers as

$$\mathcal{G}_m = \left\{ g : g(x) = I_{[a^T X^{(m)} + b \geq 0]}; a \in \mathbb{R}^m; b \in \mathbb{R} \right\}.$$

Let $\mathcal{A}_m$ be the collection of all sets of the form

$$\{\{0\} \times \{x : g_m(x) = 1\}\} \bigcup \{\{1\} \times \{x : g_m(x) = 0\}\}$$

and $\mathcal{B}_m$ the collection of sets

$$\left\{ x \in \mathbb{R}^d : g_m(x) = 1 \right\}$$

where g_m is in $\mathcal{G}_m$.

By Theorem 13.1 and Corollary 13.1 in Devroye et al. (1996), $V_{\mathcal{A}_m} = V_{\mathcal{B}_m} = m + 1$. Then the bound (2.8) is now

$$E\left(\sup_{g\in\mathcal{G}} |R(g) - R_n(g)| \right) \leq 2\sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}}.$$

This bound is only good when the dimension m is not close to the number of observations n. This bound directly influences the size of the penalization we will choose in Proposition 5.17 and in Theorem 5.15.

We are going to choose the family of non-negative weights $x_m = \ln(p)$ $m = 1, \ldots, p$ and the universal constant $\Sigma = 1$, if we define

$$\begin{aligned} (\hat{\alpha}, \hat{m}) &= \underset{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\ldots,p\}}{\arg\min} R_{n,\alpha}(\mathcal{G}_m) + \sqrt{\frac{\ln(np)}{2n(1-\alpha)^2}} \\ &+ \frac{2}{(1-\alpha)}\sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}}, \end{aligned}$$

leads us to the following oracle inequality

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\ldots,p\}} \left(R_\alpha(\mathcal{G}_m) + \sqrt{\frac{\ln(np)}{2n(1-\alpha)^2}} \right. \\ &+ \left. \frac{1}{(1-\alpha)}\sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}} + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n}(1-\alpha)} \right) \\ &+ \frac{1}{(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

First of all we would like to point out that trimming reduces the classification error that can even disappear whenever α_{max} is sufficiently large. As in model selection techniques the last three terms on the right side of the previous inequality are of order $1/\sqrt{n}$ while for a fixed m the third term will be of order $\sqrt{m\ln(n)/n}$. Finally the second term is of order $\sqrt{\ln(n)/n + \ln(p)/n}$. Thus, as long as $\ln(p)$ is smaller than n, the expected value for the best trimmed classification error for the best class will decrease as the number of observations increases.

On the other hand, it is important to consider the case that although we have many variables (d of the order of n^k with $k > 1$), only m_0 of them contribute information to the problem, that is, $[Y, X_1, \ldots, X_{m_0}]$ is independent of $[X_{m_0+1}, \ldots, X_p]$. We consider the simple case of the Bayes rule which is defined as (see Lugosi (2002)),

$$g_B(x) = \begin{cases} 1 & \text{if } \eta(x) \leq 1/2 \\ \\ 0 & \text{other case} \end{cases}$$

where

$$\begin{aligned} \eta(x) &= E(Y|X = x) = P(Y = 1|X_1 = x_1, \ldots, X_{m_0} = x_{m_0}, \ldots, X_p = x_p) \\ &= P(Y = 1|X_1 = x_1, \ldots, X_{m_0} = x_{m_0}). \end{aligned}$$

Similarly, it can be extended for the best classifier in a class that, as we saw before, it is the best trimmed classifier in the class. This implies that $\forall m \geq m_0$, $R(\mathcal{G}_m) = R(\mathcal{G}_{m_0})$ and $R_\alpha(\mathcal{G}_m) = R_\alpha(\mathcal{G}_{m_0})$ so taking $m \in \{m_0 + 1, \ldots, p\}$ will get a worse bound without any gain in the classification error. Therefore, solving this problem for $m \in \{1, \ldots, p\}$ is equivalent to solving it for $m \in \{1, \ldots, m_0\}$.

5.3 Partial SVM classification

The minimization of $R_{n,\alpha}(\mathcal{G})$ is computationally very expensive as soon as the class $\mathcal{G}$ is slightly complex. If $\mathcal{G}$ is a linear classifiers class as in Example 5.1, then the objective function

$$R_{n,\alpha}(\mathcal{G}) = \min_{w \in W_\alpha, \beta \in \mathbb{R}^{d+1}} \sum_i w_i I[(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id})(y_i - \frac{1}{2}) < 0],$$

with W_α as in (5.10), is not convex in β (not even continuous) and this generates computational problems.

To solve this problems we propose to use a positive convex loss function. Concretely, we will use the hinge loss function, this is, $\gamma(x) = (1-x)_+$ (until we get to oracle inequalities we will consider the more general case of a positive convex function). This function is associated with support vector methods. For convenience, in this section we will change the labels to $\{-1, 1\}$. We will describe the changes observed when switching from loss function 0/1 to this new loss function in the partial classification problem and we will propose a similar penalized criteria for which we will prove oracle inequalities in a model selection context. We also study the numerical aspects for the implementation of the penalized criteria.

We will consider in this section that (Y, X) (and also (Y_i, X_i)) are random vectors with values in $\{-1, 1\} \times \mathbb{R}^d$. In most of the section, $\gamma : \mathbb{R} \mapsto \mathbb{R}_+$ will be any positive convex function. Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ with associated classifier $\tilde{g}(x) = \text{sgn}(g(x))$ (with the convention that when $g(x) = 0$, $\tilde{g}(x) = 1$). For simplicity we will refer to g as a classification rule, and we will measure the classification cost by $\gamma(Yg(X))$.

The change of labels in $\{0, 1\}$ to $\{-1, 1\}$ motivates a change in notation. Notation p_1 and P_1 will continue as before, but now we will substitute p_0 and P_0 by p_{-1} and P_{-1} respectively. Now we have

$$p_{-1} = P(Y = -1),\ p_1 = P(Y = 1) = 1 - p_{-1},$$
$$P_{-1}(A) = P(A|Y = -1) \ \text{ and } \ P_1(A) = P(A|Y = 1).$$

Lemma 5.1 can be adapted to this situation with obvious changes.

The concepts associated with 0/1 loss have equivalent concepts in this new setting. For example, the *risk* associated to rule g will be given by

$$R(g) = E(\gamma(Yg(X))).$$

Our first result from this section is a reformulation of the risk associated to rule g in terms of the parametrization (p_{-1}, P_{-1}, P_1).

Lemma 5.18. *Given $g : \mathbb{R}^d \mapsto \mathbb{R}_+$,*

$$\begin{aligned} R(g) &= p_{-1} \int_0^{+\infty} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \geq t\})dt \\ &+ p_1 \int_0^{+\infty} P_1(\{x \in \mathbb{R}^d : \gamma(-g(x)) \geq t\})dt. \end{aligned}$$

Proof. Since γ is a positive function, conditioned by the value of Y we have

$$\begin{aligned} R(g) &= \int_0^{+\infty} P(\{(y,x) : \gamma(yg(x)) \geq t\})dt \\ &= \int_0^{+\infty} [P(\{(y,x) : \gamma(yg(x)) \geq t\}|Y=-1)P(Y=-1) \\ &+ P(\{(y,x) : \gamma(yg(x)) \geq t\}|Y=1)P(Y=1)]\, dt \\ &= \int_0^{+\infty} \Big[p_{-1}P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \geq t\}) \\ &+ p_1 P_1(\{x \in \mathbb{R}^d : \gamma(-g(x)) \geq t\})\Big]\, dt. \end{aligned}$$

□

Definition 5.3. *The* trimmed risk *associated to classifier g is*

$$R_\alpha(g) = \inf_{Q \in \mathcal{R}_\alpha(P)} E_Q(\gamma(Yg(X))). \tag{5.27}$$

In the following results we will see that the infimum in (5.27) is attained. Furthermore, we will use the expression of this minimum to obtain a relationship between the risk and the trimmed risk. For sake of simplicity of the proof we have divided the result into Lemma 5.19 and Proposition 5.20.

Lemma 5.19. *Given a classifier g and a trimming level $\alpha \in [0,1)$ we have*

$$\begin{aligned} R_\alpha(g) &= \min_{1-\frac{1-p_{-1}}{1-\alpha} \leq q_{-1} \leq \frac{p_{-1}}{1-\alpha}} \left[\int_0^{+\infty} \left(q_{-1} - \frac{p_{-1}}{1-\alpha} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\})\right)_+ dt \right. \\ &+ \left. \int_0^{+\infty} \left(1 - q_{-1} - \frac{1-p_{-1}}{1-\alpha} P_1(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\})\right)_+ dt\right]. \end{aligned}$$

Proof. By Lemma 5.1, $(q_{-1}, Q_{-1}, Q_1) \in \mathcal{R}_\alpha(p_{-1}, P_{-1}, P_1)$ if and only if

$$\begin{aligned} q_{-1} &\in \left[1 - \frac{1-p_{-1}}{1-\alpha}, \frac{p_{-1}}{1-\alpha}\right] \\ Q_{-1}(A) &\leq \frac{p_{-1}}{q_{-1}(1-\alpha)} P_{-1}(A)\ \forall A \\ Q_1(A) &\leq \frac{1-p_{-1}}{(1-q_{-1})(1-\alpha)} P_1(A)\ \forall A. \end{aligned} \tag{5.28}$$

Set $q_{-1} \in \left[1 - \frac{1-p_{-1}}{1-\alpha}, \frac{p_{-1}}{1-\alpha}\right]$. If Q_{-1} satisfies (5.28) then

$$Q_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\}) \leq \frac{p_{-1}}{q_{-1}(1-\alpha)} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\}).$$

From here we deduce that

$$Q_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) > t\}) \geq \left(1 - \frac{p_{-1}}{q_{-1}(1-\alpha)} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\})\right)_+$$

for each t. On the other hand, take

$$t_\alpha = \inf\left\{t : P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\}) \geq \frac{q_{-1}(1-\alpha)}{p_{-1}}\right\}.$$

Then

$$P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t_\alpha\}) \leq \frac{q_{-1}(1-\alpha)}{p_{-1}} \leq P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t_\alpha\}).$$

Let $A = \{x \in \mathbb{R}^d : \gamma(-g(x)) < t_\alpha\}$ and $B \subset \{x \in \mathbb{R}^d : \gamma(-g(x)) = t_\alpha\}$ be such that, if $C = A \cup B$ then $P_{-1}(C) = \frac{q_{-1}(1-\alpha)}{p_{-1}}$. Finally, we define $\hat{Q}_{-1}$ with the inequality

$$\hat{Q}_{-1}(D) = \frac{P_{-1}(D \cap C)}{P_{-1}(C)} \ \forall D.$$

Then, clearly, $\hat{Q}_{-1}$ is a probability in $\mathbb{R}^d$ that satisfies (5.28). On the other hand, it is easy to check that if $t \geq t_\alpha$ then

$$\hat{Q}_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) > t\}) = 0,$$

while if $t < t_\alpha$ then

$$\begin{aligned}
&\hat{Q}_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) > t\}) \\
&= \left(1 - \frac{p_{-1}}{q_{-1}(1-\alpha)} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\})\right) > 0.
\end{aligned}$$

As a consequence we conclude that, for a fixed q_{-1},

$$\begin{aligned}
&\min_{Q_{-1} \in \mathcal{R}_{1-\frac{q_{-1}}{p_{-1}(1-\alpha)}}(P_{-1})} \int_0^{+\infty} Q_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) > t\}) dt \\
&= \int_0^{+\infty} \hat{Q}_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) > t\}) dt \\
&= \int_0^{+\infty} \left(1 - \frac{p_{-1}}{q_{-1}(1-\alpha)} P_{-1}(\{x \in \mathbb{R}^d : \gamma(-g(x)) \leq t\})\right)_+ dt.
\end{aligned}$$

With a similar argument for Q_1 the proof is completed.

□

Proposition 5.20. *For a given classification rule g and a trimming level $\alpha \in [0, 1)$, then*

$$R_\alpha(g) = \int_0^\infty \frac{(P(\{(y, x) : \gamma(yg(x)) > t\}) - \alpha)_+}{1-\alpha} dt. \tag{5.29}$$

Proof. This proof will follow the same scheme of Lemma 5.19. If $Q \in \mathcal{R}_\alpha(P)$, then

$$Q(\{(y,x) : \gamma(yg(x)) \leq t\}) \leq \frac{1}{1-\alpha} P(\{(y,x) : \gamma(yg(x)) \leq t\}),$$

or, equivalently,

$$\begin{aligned} Q(\{(y,x) : \gamma(yg(x)) > t\}) &\geq 1 - \frac{1}{1-\alpha} P(\{(y,x) : \gamma(yg(x)) \leq t\}) \\ &= \frac{P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha}{1-\alpha}, \end{aligned}$$

then

$$Q(\{(y,x) : \gamma(yg(x)) > t\}) \geq \frac{(P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+}{1-\alpha}.$$

And, by (5.27), we get

$$R_\alpha(g) \geq \frac{1}{1-\alpha} \int_0^{+\infty} (P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+ dt.$$

If $F(t) = P(\{(y,x) : \gamma(yg(x)) \leq t\})$ and we take $t_\alpha = F^{-1}(1-\alpha)$ in such a way that

$$P(\{(y,x) : \gamma(yg(x)) < t_\alpha\}) \leq 1 - \alpha \leq P(\{(y,x) : \gamma(yg(x)) \leq t_\alpha\}).$$

Let $A = \{(y,x) : \gamma(yg(x)) < t_\alpha\}$ and $B \subset \{(y,x) : \gamma(yg(x)) = t_\alpha\}$ be such that, if $C = A \cup B$, $P(C) = 1 - \alpha$. Take

$$\hat{Q}(D) = \frac{P(C \cap D)}{P(C)}.$$

Clearly $\hat{Q} \in \mathcal{R}_\alpha(P)$. Furthermore

$$\int_{\{-1,1\} \times \mathbb{R}^d} \gamma(yg(x)) d\hat{Q}(y,x) = \int_0^{+\infty} \hat{Q}(\{(y,x) : \gamma(yg(x)) > t\}) dt.$$

Take now $D = \{(y,x) : \gamma(yg(x)) \leq t\}$. If $t < t_\alpha$ then $D \cap C = \{(y,x) : \gamma(yg(x)) \leq t\}$ which implies

$$\hat{Q}(\{(y,x) : \gamma(yg(x)) \leq t\}) = \frac{1}{1-\alpha}(1 - P(\{(y,x) : \gamma(yg(x)) > t\}))$$

and hence

$$\begin{aligned} \hat{Q}(\{(y,x) : \gamma(yg(x)) > t\}) &= 1 - \frac{1}{1-\alpha}(1 - P(\{(y,x) : \gamma(yg(x)) > t\})) \\ &= \frac{(P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+}{1-\alpha}. \end{aligned}$$

On the other hand, if $t > t_\alpha$ then, $D \cap C = C$ so $\hat{Q}(\{(y,x) : \gamma(yg(x)) \leq t\}) = 1$ and, consequently,

$$\hat{Q}(\{(y,x) : \gamma(yg(x)) > t\}) = 0.$$

But as $t > t_\alpha$, $P(\{(y,x) : \gamma(yg(x)) > t\}) \leq \alpha$ and

$$(P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+ = 0.$$

Hence

$$\hat{Q}(\{(y,x) : \gamma(yg(x)) > t\}) = \frac{(P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+}{1-\alpha} \ \forall t \neq t_\alpha \ \text{ (at least)}.$$

This implies that

$$\int_0^{+\infty} \hat{Q}(\{(y,x) : \gamma(yg(x)) > t\})dt = \frac{1}{1-\alpha} \int_0^{+\infty} (P(\{(y,x) : \gamma(yg(x)) > t\}) - \alpha)_+ dt.$$

□

For some convex loss functions, the optimal classifier coincides with the optimal classifier for 0/1 loss (see Lemma 2.1 in Lin (2002)), specifically, this occurs for hinge loss (Lemma 3.1 in Lin (2002)). In this sense, it is said that a loss function whose optimal classifier coincides with the one from 0/1 loss is consistent.

Rarely we will know the distribution of P in order to obtain optimal classifiers and we will have to estimate them from a training sample. We will chose as estimators those that minimize the *empirical risk* which in this case is defined by

$$R_n(g) := \frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i g(X_i)).$$

As happened before with (5.8) and (5.9), we define the *trimmed empirical risk* as

$$R_{n,\alpha}(g) := \inf_{Q \in \mathcal{R}_\alpha(P_n)} E_Q(\gamma(Yg(X)))$$

and

$$R_{n,\alpha}(g) := \min_{W_\alpha} \sum_{i=1}^{n} w_i \gamma(y_i g(x_i)),$$

where W_α is as in (5.10).

As a consequence of Proposition 5.20, the trimmed empirical risk can be calculated by means of the following expression

$$R_{n,\alpha}(g) = \int_0^{+\infty} \frac{(P_n(\{(y,x) : \gamma(yg(x)) \geq t\}) - \alpha)_+}{1-\alpha} dt. \tag{5.30}$$

The empirical risk is an unbiased estimator for the theoretical risk. Even though, as we saw in Proposition 5.11, we know this does not occur when we work with trimmed risks, we can find a bound for the difference between both quantities.

Proposition 5.21. *Let α be a trimming level and g a given classifier,*

$$\begin{aligned} 0 &\leq E(R_{n,\alpha}(g)) - R_\alpha(g) \\ &\leq \frac{1}{1-\alpha}\int_0^{+\infty} \sqrt{\frac{P(\{(y,x) : \gamma(yg(x)) \geq t\})(1 - P(\{(y,x) : \gamma(yg(x)) \geq t\}))}{n}} dt. \end{aligned}$$

Proof.

$$\begin{aligned} &E(R_{n,\alpha}(g)) - R_\alpha(g) \\ &= \frac{1}{1-\alpha}\int_0^{+\infty} [E(P_n(\{(y,x) : \gamma(yg(x)) \geq t\}) - \alpha)_+ \\ &- (P(\{(y,x) : \gamma(yg(x)) \geq t\}) \geq t) - \alpha)_+]dt \end{aligned}$$

Let $h(z) = z_+$ which is a convex and 1-Lipschitz function. By Jensen's inequality we know that $E(h(z)) \geq h(E(z))$, if we take $Z(t) = P_n(\{(y,x) : \gamma(yg(x)) \geq t\}) - \alpha$ we know that $E(Z(t)) = P(\{(y,x) : \gamma(yg(x)) \geq t\}) - \alpha$. Recovering the last equality and applying Jensen's inequality we get

$$E(R_{n,\alpha}(g)) - R_\alpha(g) = \frac{1}{1-\alpha}\int_0^{+\infty} [E(h(Z(t)) - h(E(Z(t))))]dt \geq 0.$$

Let us now consider the second inequality, starting from the previous equality we have

$$\begin{aligned} &\frac{1}{1-\alpha}\int_0^{+\infty} [E(h(Z(t)) - h(E(Z(t))))]dt \\ &= \frac{1}{1-\alpha}\int_0^{+\infty} E(h(Z(t)) - h(E(Z(t))))dt \\ &= \frac{1}{1-\alpha}\left|\int_0^{+\infty} E(h(Z(t)) - h(E(Z(t))))dt\right| \\ &\leq \frac{1}{1-\alpha}\int_0^{+\infty} E|h(Z(t)) - h(E(Z(t)))|dt. \end{aligned}$$

As h is 1-Lispchitz we have that $|h(Z(t)) - h(E(Z(t)))| \leq |Z(t) - E(Z(t))|$ and following a reasoning as in the proof of Proposition 5.11 we come to

$$\begin{aligned} E|Z(t) - E(Z(t))| &\leq \frac{1}{\sqrt{2n}}\sqrt{Var(Z(t))} \\ &= \sqrt{\frac{P(\{(y,x) : \gamma(yg(x)) \geq t\})(1 - P(\{(y,x) : \gamma(yg(x)) \geq t\}))}{2n}}. \end{aligned}$$

□

5.3.1 Optimal trimming selection

We have repeatedly pointed out in this work the importance of getting oracle inequalities for the expected value of trimmed risk. As we said at the beginning of this section, from now on we are going to focus on the hinge loss function.

Again, for the sake of simplicity, we are going to consider first the case in which the class of classifiers in formed by a unique element, so the penalization we are considering will only depend on the trimming level. Subsequently we will extend this result to the more realistic case in which many classes of classifiers are considered each one with a different complexity level. Now we are enunciating our principal result for this simple case.

Theorem 5.22. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P where $\xi_i = (Y_i, X_i)$ such that $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. Let g be a given classifier so that $|g(X_i)| \leq K$ with $K \in \mathbb{R}_+$ a positive constant. Take $\alpha_{max} \in (0, 1]$. If we consider the penalization function*

$$pen(\alpha) = \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}}$$

and define

$$\hat{\alpha} = \underset{\alpha \in [0, \alpha_{max}]}{\arg\min}\, R_{n,\alpha}(g) + pen(\alpha),$$

then the following bound holds,

$$\begin{aligned} E(R_{\hat{\alpha}}(g)) \quad &\leq \quad \min_{\alpha \in [0, \alpha_{max}]} \left(R_\alpha(g) + pen(\alpha) + \frac{1+K}{\sqrt{n}(1-\alpha)} \right) \\ &+ \quad \frac{1+K}{n(1-\alpha_{max})} + \frac{2K}{1-\alpha_{max}}\sqrt{\frac{2\pi}{n}}. \end{aligned} \tag{5.31}$$

For technical reasons, to prove Theorem 5.22 is necessary to begin by proving the following variant in which the set of admissible trimming levels is finite.

Proposition 5.23. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P with $\xi_i = (Y_i, X_i)$ where $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. Let g be a given classifier such that $|g(X_i)| \leq K$ with $K \in \mathbb{R}_+$ a positive constant. Let $k_0 < n$ be a natural number and $A = \{0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{k_0}{n}\}$ the set of admissible trimming levels. If we consider the penalization function*

$$pen(\alpha) = \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}}$$

and define

$$\hat{\alpha} = \underset{\alpha \in A}{\arg\min}\, R_{n,\alpha}(g) + pen(\alpha),$$

the following bound holds,

$$E(R_{\hat{\alpha}}(g)) \leq \min_{\alpha \in A} \left(R_\alpha(g) + pen(\alpha) + \frac{1+K}{\sqrt{n}(1-\alpha)} \right) + \frac{2K}{1-\alpha_{max}} \sqrt{\frac{2\pi}{n}}.$$

Proof. We will start from the basic inequality

$$R_{n,\hat{\alpha}}(g) + pen(\hat{\alpha}) \leq R_{n,\alpha}(g) + pen(\alpha), \tag{5.32}$$

from there, adding and subtracting terms we come to

$$\begin{aligned} R_{\hat{\alpha}}(g) &\leq R_\alpha(g) + pen(\alpha) + [(R_{n,\alpha}(g) - E(R_{n,\alpha}(g))) + (E(R_{n,\alpha}(g)) - R_\alpha(g))] \\ &+ (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)) - pen(\hat{\alpha}). \end{aligned} \tag{5.33}$$

We will study the behaviour of each term separately, we begin with $E(R_{n,\alpha}(g)) - R_\alpha(g)$. To bound this quantity we will use Proposition 5.21 and inequality $1 - yg(x) \leq 1 + |g(x)|$, then if $t > 1 + K$ we have that $P(\{(y,x) : (1 - yg(x))_+ \geq t\}) = 0$ and hence, the integral in (5.31) becomes

$$\int_0^{1+K} \sqrt{\frac{P(\{(y,x) : (1-yg(x))_+ \geq t\})(1 - P(\{(y,x) : (1-yg(x))_+ \geq t\}))}{n}} dt,$$

and we can deduce that

$$E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{1+K}{\sqrt{n}(1-\alpha)}. \tag{5.34}$$

We are going now to work with $(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)))$. For this term we will use McDiarmid's inequality. In first place we have to check the conditions to use it. We define $F(\xi_1 \ldots, \xi_n) := R_{n,\alpha}(g)$ with $\xi_i = (Y_i, X_i)$, then if $W = (w_1, \ldots, w_n)$ is such that $\sum w_i = 1$ and $0 \leq w_i \leq \frac{1}{n(1-\alpha)}$ and $X_1', \ldots, X_n'$ are random variables independent form each other and independent from $X_1, \ldots, X_n$ we have

$$\begin{aligned} &|F(\xi_1 \ldots, \xi_i, \ldots, \xi_n) - F(\xi_1 \ldots, \xi_i', \ldots, \xi_n)| \\ &= \left| \min_W \sum_{j=1}^n w_j (1 - Y_j g(X_j))_+ - \min_W \sum_{j \in \{1,\ldots,n\} \setminus i} w_j (1 - Y_j g(X_j))_+ + w_i (1 - Y_i' g(X_i'))_+ \right| \\ &\leq \left| \hat{w}_i ((1 - Y_i g(X_i))_+ - (1 - Y_i' g(X_i'))_+) \right| \\ &\leq \frac{1}{n(1-\alpha)} \left| (1 - Y_i g(X_i))_+ - (1 - Y_i' g(X_i'))_+ \right| \\ &\leq \frac{1}{n(1-\alpha)} \left| Y_i g(X_i) - Y_i' g(X_i') \right| \\ &\leq \frac{1}{n(1-\alpha)} \left(|Y_i g(X_i)| + \left| Y_i' g(X_i') \right| \right) \leq \frac{2K}{n(1-\alpha)}, \end{aligned}$$

with $(\hat{w}_1, \ldots, \hat{w}_n) = \arg\min_{W} \sum_{j=\{1,\ldots,n\}\setminus i} w_j(1 - Y_j g(X_j))_+ + w_i(1 - Y_i' g(X_i'))_+$.

Taking $t = \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}}$ and applying McDiarmid's inequality we obtain

$$P\left(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}}\right) \leq e^{-z}.$$

That is, except for a set of probability less than e^{-z}

$$R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \leq \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}} \tag{5.35}$$

With this we have controlled the first two terms in (5.33). Now we look at $R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)$. Applying again McDiarmid's inequality, this time over the right side of the previous inequality, with $t = \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}$ we get that for all $\alpha' \in A$

$$P\left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}\right) \leq \frac{1}{n}e^{-z}.$$

Hence,

$$\begin{aligned}
&P\left(R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) - \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}} \geq 0\right) \\
&\leq P\left(\sup_{\alpha' \in A}\left(R_{\alpha'}(g) - R_{n,\alpha'}(g) - \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}\right) \geq 0\right) \\
&\leq \sum_{\alpha' \in A} P\left(R_{\alpha'}(g) - R_{n,\alpha'}(g) \geq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}\right) \\
&\leq \sum_{\alpha' \in A} P\left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}\right) \leq n\frac{1}{n}e^{-z} \leq e^{-z}.
\end{aligned}$$

Then, with probability at least $1 - e^{-z}$ we have that

$$R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \leq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\hat{\alpha})^2}} \leq \sqrt{\frac{2K^2\ln(n)}{n(1-\hat{\alpha})^2}} + \sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}}. \tag{5.36}$$

Choosing the penalization function as in the statement along with (5.34), (5.35) and (5.36), imply that with probability at least $1 - 2e^{-z}$

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}} + \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}} + \sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}}.$$

Grouping together terms we get

$$R_{\hat{\alpha}}(g) - R_{\alpha}(g) - \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}} - \frac{1+K}{n(1-\alpha)} \leq 2\sqrt{\frac{2K^2 z}{n(1-\alpha_{max})^2}}.$$

Reasoning as in the proof of Proposition 5.13, we conclude that

$$E(R_{\hat{\alpha}}(g)) \leq \min_{\alpha \in A}\left(R_{\alpha}(g) + \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}} + \frac{1+K}{\sqrt{n}(1-\alpha)}\right) + \frac{2K}{1-\alpha_{max}}\sqrt{\frac{2\pi}{n}}.$$

□

To complete the proof of Theorem 5.22 we need to justify that when two trimming levels are close, the difference between trimming risks in small. This is our next result.

Proposition 5.24. *Let α_1, α_2 be two trimming levels such that $\alpha_2 \in [\alpha_1, \alpha_1 + \frac{1}{n}]$, let $\alpha_{max} \in [0,1)$ be such that $\alpha_1 \leq \alpha_2 \leq \alpha_{max} < 1$ and let g be a given classifier. If we denote by $F(t) = P(\{(y,x) : (1 - yg(x))_+ \leq t\})$ and by $F_n(t) = P_n(\{(y,x) : (1 - yg(x))_+ \leq t\})$, then*

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) \leq \frac{F^{-1}(1-\alpha_1)}{n(1-\alpha_{max})} \quad \textit{and} \quad R_{n,\alpha_1}(g) - R_{n,\alpha_2}(g) \leq \frac{F_n^{-1}(1-\alpha_1)}{n(1-\alpha_{max})}.$$

Proof. Using equality (5.29), we start from

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) = \frac{1}{1-\alpha_1}\int_0^{+\infty}(1 - F(t) - \alpha_1)_+ dt - \frac{1}{1-\alpha_2}\int_0^{+\infty}(1 - F(t) - \alpha_2)_+ dt.$$

As for $t > F^{-1}(1-\alpha_1)$, $(1 - F(t) - \alpha_1)_+ = 0$ and the same happens for $t > F^{-1}(1-\alpha_2)$ and $(1 - F(t) - \alpha_2)_+$, the integrals above turn into

$$\begin{aligned}
&R_{\alpha_1}(g) - R_{\alpha_2}(g) = \frac{1}{1-\alpha_1}\int_0^{F^{-1}(1-\alpha_1)}(1 - \alpha_1 - F(t))dt \\
&- \frac{1}{1-\alpha_2}\int_0^{F^{-1}(1-\alpha_2)}(1 - \alpha_2 - F(t))dt \\
&= \frac{1}{(1-\alpha_1)(1-\alpha_2)}\int_0^{F^{-1}(1-\alpha_2)}\left[(1-\alpha_2)(1-\alpha_1-F(t)) - (1-\alpha_1)(1-\alpha_2-F(t))\right]dt \\
&+ \frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1 - \alpha_1 - F(t))dt \\
&= \frac{1}{(1-\alpha_1)(1-\alpha_2)}\int_0^{F^{-1}(1-\alpha_2)}F(t)(\alpha_2 - \alpha_1)dt + \frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1 - \alpha_1 - F(t))dt \\
&\leq \frac{(1-\alpha_2)F^{-1}(1-\alpha_2)}{n(1-\alpha_1)(1-\alpha_2)} + \frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1 - \alpha_1 - F(t))dt \\
&\leq \frac{F^{-1}(1-\alpha_2)}{n(1-\alpha_{max})} + \frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1 - \alpha_1 - F(t))dt.
\end{aligned}$$

Now, as for all $t \in [F^{-1}(1-\alpha_2), F^{-1}(1-\alpha_1)]$, $F(t) \geq 1-\alpha_2$, then

$$\begin{aligned}
&\frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1-\alpha_1-F(t))dt \\
&\leq \frac{1}{1-\alpha_1}\int_{F^{-1}(1-\alpha_2)}^{F^{-1}(1-\alpha_1)}(1-\alpha_1+\alpha_2-1)dt \\
&= \frac{(\alpha_2-\alpha_1)}{1-\alpha_1}(F^{-1}(1-\alpha_1)-F^{-1}(1-\alpha_2)) \\
&\leq \frac{(F^{-1}(1-\alpha_1)-F^{-1}(1-\alpha_2))}{n(1-\alpha_{max})}.
\end{aligned}$$

And, hence,

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) \leq \frac{F^{-1}(1-\alpha_2)}{n(1-\alpha_{max})} + \frac{(F^{-1}(1-\alpha_1)-F^{-1}(1-\alpha_2))}{n(1-\alpha_{max})} = \frac{F^{-1}(1-\alpha_1)}{n(1-\alpha_{max})}.$$

The proof is exactly the same for $R_{n,\alpha_1}(g) - R_{n,\alpha_2}(g)$ starting from (5.30).

□

Proof of Theorem 5.22. As we did in the proof of Theorem 5.12, we will replicate the proof of Proposition 5.23 up to inequality (5.35). From there, we can deduce with probability of at least $1-e^{-z}$

$$R_{\hat{\alpha}}(g) \leq R_{\alpha}(g) + pen(\alpha) + \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}} - pen(\hat{\alpha}) + [R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)]. \quad (5.37)$$

We define the set $A = \{0, \frac{1}{n}, \ldots, \frac{k_0}{n}\}$ where $k_0 = [n\alpha_{max}]$ and, as in the proposition, we have that for all $\alpha' \in A$, with a probability of at least $1-e^{-z}$,

$$E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \leq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}}. \quad (5.38)$$

If $\alpha' \in [0, \alpha_{max})$, then $\exists \alpha'' \in A$ such that $\alpha'' \leq \alpha' \leq \alpha_{max}$. By Proposition 5.24, in the set where (5.38) is satisfied we have that, if we denote $F_n(t) = P_n(\{(y,x) : (1-yg(x))_+ \leq t\})$, as for any p, $F^{-1}(p) \leq 1+K$, then

$$\begin{aligned}
&E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \\
&= E(R_{n,\alpha''}(g)) - R_{n,\alpha''}(g) + E(R_{n,\alpha'}(g)) - E(R_{n,\alpha''}(g)) + R_{n,\alpha''}(g) - R_{n,\alpha'}(g) \\
&= (E(R_{n,\alpha''}(g)) - R_{n,\alpha''}(g)) + E(R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) - (R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) \\
&\leq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha'')^2}} + \frac{F_n^{-1}(1-\alpha'')}{n(1-\alpha_{max})} \leq \sqrt{\frac{2K^2(\ln(n)+z)}{n(1-\alpha')^2}} + \frac{1+K}{n(1-\alpha_{max})} \\
&\leq \sqrt{\frac{2K^2\ln(n)}{n(1-\alpha')^2}} + \sqrt{\frac{2K^2 z}{n(1-\alpha')^2}} + \frac{1+K}{n(1-\alpha_{max})}.
\end{aligned}$$

And, hence,

$$R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \leq \sqrt{\frac{2K^2 \ln(n)}{n(1-\hat{\alpha})^2}} + \sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}} + \frac{1+K}{n(1-\alpha_{max})}.$$

Going back to (5.37) and choosing the penalization as in the statement we have that, with probability at least $1 - 2e^{-z}$,

$$\begin{aligned} R_{\hat{\alpha}}(g) &\leq R_{\alpha}(g) + pen(\alpha) + \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}} - pen(\hat{\alpha}) \\ &+ \sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}} + \frac{1+K}{n(1-\alpha_{max})}. \end{aligned}$$

Integrating with respect to z and bounding $\hat{\alpha}$ by α_{max} we come to

$$\begin{aligned} E(R_{\hat{\alpha}}(g)) &\leq \min_{\alpha \in [0, \alpha_{max})} \left(R_{\alpha}(g) + \frac{K}{1-\alpha}\sqrt{\frac{2\ln(n)}{n}} + \frac{1+K}{\sqrt{n}(1-\alpha)} \right) \\ &+ \frac{2K}{1-\alpha_{max}}\sqrt{\frac{2\pi}{n}} + \frac{1+K}{n(1-\alpha_{max})}. \end{aligned}$$

□

Now, we will consider the more realistic case in which instead of a class formed by a unique classifier we have to chose between several complex models each one of them with various classifiers. The penalization now will depend both, on the complexity of the model and in the trimming level. Again, to simplify the proof we divide the result.

Theorem 5.25. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P where $\xi_i = (Y_i, X_i)$ with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. Let $\{\mathcal{G}_m\}_{m \in \mathbb{N}}$ be a family of classifiers classes with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$ and such that $|g(X_i)| \leq K$, with $K \in \mathbb{R}_+$ a positive constant, for all $g \in \mathcal{G}_m$ and for all $m \in \mathbb{N}$. Take $\alpha_{max} \in (0, 1)$ and let Σ a positive constant and consider $\{x_m\}_{m \in \mathbb{N}}$ a family of non negative weights such that*

$$\sum_{m \in \mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

If we consider the penalization function

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{2K^2(\ln(n) + x_m)}{n(1-\alpha)^2}} + \frac{2(1+K)}{1-\alpha}\sqrt{\frac{4V_{\mathcal{G}_m}\ln(n+1) + \ln(2)^3}{n\ln(2)^2}}$$

and define

$$(\hat{\alpha}, \hat{m}) = \underset{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}}{\arg\min} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m), \quad g' := \underset{g\in\mathcal{G}_m}{\arg\min} R_\alpha(g),$$

then the following bound holds:

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha,m)\in[0,\alpha_{max}]\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{n(1-\alpha)} \right) \\ &+ \frac{K(1+\Sigma)}{(1-\alpha_{max})}\sqrt{\frac{\pi}{2n}} + \frac{1+K}{n(1-\alpha_{max})}. \end{aligned} \tag{5.39}$$

The proof of this theorem is based on the technical result below.

Proposition 5.26. *Let $\xi_1, \ldots, \xi_n$ be n independent and identically distributed observations with distribution P where $\xi_i = (Y_i, X_i)$ with $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$. Let $\{\mathcal{G}_m\}_{m\in\mathbb{N}}$ a family of classes of classifiers with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$ in such a way that $|g(X_i)| \leq K$, with $K \in \mathbb{R}_+$, for all $g \in \mathcal{G}_m$ and for all m. Let $k_0 < n$ be a natural number and A the set of admissible trimmings, $A = \{0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{k_0}{n}\}$. Let Σ be a positive constant and consider $\{x_m\}_{m\in\mathbb{N}}$ a family of non negative weights such that*

$$\sum_{m\in\mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

Consider the penalization function

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{2K^2(\ln(n) + x_m)}{n(1-\alpha)^2}} + \frac{2(1+K)}{1-\alpha}\sqrt{\frac{4V_{\mathcal{G}_m}\ln(n+1) + \ln(2)^3}{n\ln(2)^2}}.$$

Suppose that

$$(\hat{\alpha}, \hat{m}) = \underset{(\alpha,m)\in A\times\mathbb{N}}{\arg\min} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m),$$

then, the following bound holds:

$$E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \leq \min_{(\alpha,m)\in A\times\mathbb{N}} \left(R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{\sqrt{n}(1-\alpha)} \right) + \frac{K(1+\Sigma)}{(1-\frac{k_0}{n})}\sqrt{\frac{\pi}{2n}}.$$

Proof. We will start from the basic inequality (5.32) to come to

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_\alpha(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + (R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m)) \\ &- pen(\hat{\alpha}, \mathcal{G}_{\hat{m}}) + (R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})). \end{aligned} \tag{5.40}$$

This inequality is the starting basic inequality in the proof of Proposition 5.17 so this proof will be similar. As before, we are looking for bounds for the quantities inside

the parenthesis. First, we will bound $(R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m))$ that, by definition of both quantities, is

$$R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m) = \min_{g\in\mathcal{G}_m} R_{n,\alpha}(g) - \min_{g\in\mathcal{G}_m} R_\alpha(g).$$

Suppose that $R_\alpha(g') := \min_{g\in\mathcal{G}_m} R_\alpha(g)$, then we have that

$$\begin{aligned}\min_{g\in\mathcal{G}_m} R_{n,\alpha}(g) - \min_{g\in\mathcal{G}_m} R_\alpha(g) &\leq R_{n,\alpha}(g') - R_\alpha(g') \\ &= (E(R_{n,\alpha}(g')) - R_\alpha(g')) + (R_{n,\alpha}(g') - E(R_{n,\alpha}(g'))).\end{aligned}$$

As we are in the same conditions than in Proposition 5.23 we can use the same reasoning and bound the first term by (5.34) and the second one by (5.35). Thus, with a probability greater than $1 - e^{-z}$ we have

$$R_{n,\alpha}(\mathcal{G}_m) - R_\alpha(\mathcal{G}_m) \leq \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}}. \tag{5.41}$$

For controlling the second line in (5.40), we need to adapt the proof of Proposition 5.17 because we don't have the same equalities for $R_\alpha(g)$ and $R_{n,\alpha}(g)$ as before. We begin with the same chain of inequalities we had then,

$$\begin{aligned}R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq \sup_{(\alpha',m')\in A\times\mathbb{N}} \left(R_{\alpha'}(\mathcal{G}_{m'}) - R_{n,\alpha'}(\mathcal{G}_{m'})\right) \\ &\leq \sup_{(\alpha',m')\in A\times\mathbb{N}} \left(\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g))\right).\end{aligned}$$

First we will focus in the supremum inside the parenthesis that becomes

$$\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) = E\left(\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g))\right) \tag{5.42}$$

$$+ \left[\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) - E\left(\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g))\right)\right]. \tag{5.43}$$

For bounding (5.43) we will use McDiarmid's inequality, we begin by proving that the bounded differences condition is met. Denote by $Z := f(\xi_1,\ldots,\xi_n) = \sup_{g\in\mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g))$ and define $Z^{(i)} := f(\xi_1,\ldots,\xi_i',\ldots,\xi_n)$, we want to prove that $|Z - Z^{(i)}| \leq c_i$ where c_i are constants. We will denote by $R^{(i)}_{n,\alpha'}(g)$ the empirical risk associated to the sample $\xi_1,\ldots,\xi_i',\ldots,\xi_n$, then

$$|(R_{\alpha'}(g) - R_{n,\alpha'}(g)) - (R_{\alpha'}(g) - R^{(i)}_{n,\alpha'}(g))| = |R^{(i)}_{n,\alpha'}(g) - R_{n,\alpha'}(g)| \leq \frac{2K}{n(1-\alpha')},$$

where the last inequality was obtained in the proof of Theorem 5.23. By Lemma 5.16 we have that $|Z - Z^{(i)}| \leq \frac{2K}{n(1-\alpha')}$ for all $i = 1, \ldots, n$ and we can apply McDiarmid's inequality to conclude that

$$P\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g) - R_{n,\alpha'}(g)) - E\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g) - R_{n,\alpha'}(g))\right)\right.$$
$$\left.\geq \sqrt{\frac{2K^2(\ln(n+1) + z + x_m)}{n(1-\alpha')^2}}\right) \leq \frac{1}{n}e^{-z-x_m}. \tag{5.44}$$

Now we bound the right hand side of (5.42). If for a classifier g we denote by $F_g(t) = P(\{(y,x) : (1 - yg(x))_+ \leq t\})$ and $F_{g,n}(t) = P_n(\{(y,x) : (1 - yg(x))_+ \leq t\})$, then by Proposition 5.20 and given that $(1 - yg(x))_+ \leq 1 + K$ we have,

$$\begin{aligned}
&R_{\alpha'}(g) - R_{n,\alpha'}(g)\\
&= \frac{1}{1-\alpha'}\int_0^{+\infty}\left[(1 - F_g(t) - \alpha')_+ - (1 - F_{g,n}(t) - \alpha')_+\right]dt\\
&= \frac{1}{1-\alpha'}\int_0^{1+K}\left[(1 - F_g(t) - \alpha')_+ - (1 - F_{g,n}(t) - \alpha')_+\right]dt\\
&\leq \frac{1}{1-\alpha'}\int_0^{1+K}|(1 - F_g(t) - \alpha')_+ - (1 - F_{g,n}(t) - \alpha')_+|dt\\
&\leq \frac{1}{1-\alpha'}\int_0^{1+K}|F_{g,n}(t) - F_g(t)|dt.
\end{aligned}$$

Taking the expected value of the supremum we get

$$E\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g) - R_{n,\alpha'}(g))\right) \leq \frac{1}{1-\alpha'}\int_0^{1+K}E\left(\sup_{g\in\mathcal{G}_{m'}}|F_{g,n}(t) - F_g(t)|\right)dt.$$

Fix t and call $A_g := \{(y,x) : t < 1 - yg(x)\}$ and define $\mathcal{A}_m = \{A_g : g \in \mathcal{G}_m\}$. By Lemma 5.27 we know that $V_{\mathcal{A}_m} \leq \frac{4V^2_{\mathcal{G}_m}}{\ln(2)^2}$. Then, by (2.7),

$$\begin{aligned}
\int_0^{1+K}E\left(\sup_{g\in\mathcal{G}_{m'}}|F_{g,n}(t) - F_g(t)|\right)dt &= \int_0^{1+K}E\left(\sup_{A_g\in\mathcal{A}_{m'}}|P(A_g) - P_n(A_g)|\right)dt\\
&\leq \int_0^{1+K}2\sqrt{\frac{V_{\mathcal{A}_{m'}}\ln(n+1) + \ln(2)}{n}}dt\\
&= 2(1+K)\sqrt{\frac{V_{\mathcal{A}_{m'}}\ln(n+1) + \ln(2)}{n}}\\
&\leq 2(1+K)\sqrt{\frac{\frac{4V^2_{\mathcal{G}_{m'}}}{\ln(2)^2}\ln(n+1) + \ln(2)}{n}}.
\end{aligned} \tag{5.45}$$

Equation (5.44) together with (5.45) say that $\forall \alpha' \in A$ and $\forall m' \in \mathbb{N}$,

$$P\left(\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g)-R_{n,\alpha'}(g)) \geq \sqrt{\frac{2K^2(\ln(n+1)+z+x_m)}{n(1-\alpha')^2}} + \frac{1+K}{1-\alpha'}\sqrt{\frac{\frac{4V^2_{\mathcal{G}_{m'}}}{\ln(2)^2}\ln(n+1)+\ln(2)}{n}}\right) \leq \frac{1}{n}e^{-z-x_m}.$$

Using this we can deduce that,

$$\begin{aligned}
&P\left(\bigcup_{(\alpha',m')\in A\times\mathbb{N}}\sup_{g\in\mathcal{G}_{m'}}(R_{\alpha'}(g)-R_{n,\alpha'}(g)) \geq \sqrt{\frac{2K^2(\ln(n+1)+z+x_m)}{n(1-\alpha')^2}}\right.\\
&\left.+\frac{2(1+K)}{1-\alpha'}\sqrt{\frac{\frac{4V^2_{\mathcal{G}_{m'}}}{\ln(2)^2}\ln(n+1)+\ln(2)}{n}}\right)\\
&\leq \sum_{\alpha'\in A}\sum_{m'\in\mathbb{N}} P\left(R_{\alpha'}(g)-R_{n,\alpha'}(g) \geq \sqrt{\frac{2K^2(\ln(n+1)+z+x_m)}{n(1-\alpha')^2}}\right.\\
&\left.+\frac{2(1+K)}{1-\alpha'}\sqrt{\frac{\frac{4V^2_{\mathcal{G}_{m'}}}{\ln(2)^2}\ln(n+1)+\ln(2)}{n}}\right)\\
&\leq \sum_{\alpha'\in A}\sum_{m'\in\mathbb{N}}\frac{1}{n}e^{-z-x_{m'}} \leq \sum_{m'\in\mathbb{N}} e^{-z-x_{m'}} \leq \Sigma e^{-z}.
\end{aligned}$$

We can conclude that with probability at least $1-\Sigma e^{-z}$

$$\begin{aligned}
&R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq \sqrt{\frac{2K^2(\ln(n+1)+z+x_m)}{n(1-\hat{\alpha})^2}}\\
&+\frac{2(1+K)}{1-\hat{\alpha}}\sqrt{\frac{4V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}}\\
&\leq \sqrt{\frac{2K^2(\ln(n)+x_{\hat{m}})}{n(1-\hat{\alpha})^2}}+\sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}}+\frac{2(1+K)}{1-\hat{\alpha}}\sqrt{\frac{4V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}}.
\end{aligned}$$

With all this we can ensure that except in a set of probability not greater than $(\Sigma+1)e^{-z}$

$$\begin{aligned}
R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m)+pen(\alpha,\mathcal{G}_m)+\sqrt{\frac{2K^2 z}{n(1-\alpha)^2}}+\frac{1+K}{\sqrt{n}(1-\alpha)}\\
&- pen(\hat{\alpha},\mathcal{G}_{\hat{m}})+\sqrt{\frac{2K^2(\ln(n)+x_{\hat{m}})}{n(1-\hat{\alpha})^2}}+\sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}}\\
&+ \frac{2(1+K)}{1-\hat{\alpha}}\sqrt{\frac{4V_{\mathcal{G}_{\hat{m}}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}}.
\end{aligned}$$

Considering the penalization in the statement as

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{2K^2 z}{n(1-\alpha)^2}} + \sqrt{\frac{2K^2 z}{n(1-\hat{\alpha})^2}} \\ &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{\sqrt{n}(1-\alpha)} + \sqrt{\frac{8K^2 z}{n(1-\frac{k_0}{n})^2}}. \end{aligned}$$

Grouping and integrating with respect to z we arrive to,

$$E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) \leq \min_{(\alpha,m)\in A\times\mathbb{N}} \left(R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{1+K}{\sqrt{n}(1-\alpha)} \right) + \frac{K(1+\Sigma)}{(1-\frac{k_0}{n})} \sqrt{\frac{\pi}{2n}}.$$

□

In this proof we have used the following result in which a relationship is established between the dimension VC of the class $\mathcal{G}_m$ and that of the sets of the form $\mathcal{A}_m = \{\{(y,x) : t < 1 - yg(x)\} : g \in \mathcal{G}_m\}$.

Lemma 5.27. *Let $\mathcal{A}_m = \{\{(y,x) : t < 1 - yg(x)\} : g \in \mathcal{G}_m\}$ with t fixed and $\mathcal{G}_m$ be a class of classifiers, then*

$$V_{\mathcal{A}_m} \leq \frac{4V_{\mathcal{G}_m}^2}{\ln(2)^2}.$$

Proof. The set $\mathcal{A}_m$ can be written

$$\mathcal{A}_m = \left\{ \{x : g(x) < 1 - t\} \times \{1\} \bigcup \{x : g(x) > t - 1\} \times \{-1\} : g \in \mathcal{G}_m \right\}.$$

Define

$$\tilde{\mathcal{A}}_m = \left\{ \{x : g(x) < 1 - t\} \times \{1\} \bigcup \{x : \tilde{g}(x) > t - 1\} \times \{-1\} : g \in \mathcal{G}_m, \tilde{g} \in \mathcal{G}_m \right\},$$

obviously $\mathcal{A}_m \subset \tilde{\mathcal{A}}_m$ and, hence, $S_{\mathcal{A}_m}(n) \leq S_{\tilde{\mathcal{A}}_m}(n)$. If we consider the sets

$$\begin{aligned} \tilde{\mathcal{A}}_{m,1} &= \{\{x : g(x) < 1 - t\}\}_{g\in\mathcal{G}_m}, \\ \tilde{\mathcal{A}}_{m,2} &= \{\{x : g(x) > t - 1\}\}_{g\in\mathcal{G}_m}, \end{aligned}$$

then $\tilde{\mathcal{A}}_m = \tilde{\mathcal{A}}_{m,1} \cup \tilde{\mathcal{A}}_{m,2}$ and, by Theorem 13.5 in Devroye et al. (1996), $S_{\mathcal{A}_m}(n) \leq S_{\tilde{\mathcal{A}}_{m,1}}(n) S_{\tilde{\mathcal{A}}_{m,2}}(n)$. Let us prove that $V_{\tilde{\mathcal{A}}_{m,i}} \leq V_{\mathcal{G}_m}$ with $i = 1, 2$.
Take $I \subset \{(u_1, x_1), \ldots, (u_n, x_n)\}$; there exists g such that $I = \{(u_1, x_1), \ldots, (u_n, x_n)\} \cap \{g(x) > u\}$. Then, $g(x_i) > u_i$ if $(u_i, x_i) \in I$ but $g(x_i) \leq u_i$ if $(u_i, x_i) \notin I$. If $V_{\mathcal{G}_m} = l$, then there is not a set of $l+1$ points $(u_1, x_1), \ldots, (u_{l+1}, x_{l+1})$ that can be fully extracted with intersections with sets of the form $\{g(x) > u\}$. In particular, there is not a set of $l+1$ points $(t-1, x_1), \ldots, (t-1, x_{l+1})$ that can be fully extracted with intersections of

sets of the form $\{g(x) > t-1\}$. This implies that $V_{\tilde{\mathcal{A}}_{m,2}} \leq V_{\mathcal{G}_m}$, which in turn implies that $S_{\tilde{\mathcal{A}}_{m,2}}(l+1) < 2^{l+1}$. By section (iv) from Lemma 2.6 in van der Vaart and Wellner (1996) and Exercise 2.6.10 in the same book, $S_{\tilde{\mathcal{A}}_{m,1}}(l) = S_{\tilde{\mathcal{A}}_{m,2}}(l)$. Then

$$S_{\mathcal{A}_m}(l+1) \leq S_{\tilde{\mathcal{A}}_{m,1}}(l+1) S_{\tilde{\mathcal{A}}_{m,2}}(l+1),$$

and, by Theorem 13.3 in Devroye et al. (1996)

$$S_{\tilde{\mathcal{A}}_{m,1}}(l+1) S_{\tilde{\mathcal{A}}_{m,2}}(l+1) \leq n^{2V_{\mathcal{G}_m}}.$$

As for all n

$$\frac{n}{\ln(n)} > \frac{2V_{\mathcal{G}_m}}{\ln(2)},$$

$n^{2V_{\mathcal{G}_m}} < 2^n$ and

$$\frac{n}{\ln(n)} > \sqrt{n} > \frac{2V_{\mathcal{G}_m}}{\ln(2)},$$

then

$$V_{\mathcal{A}_m} \leq \frac{4V_{\mathcal{G}_m}^2}{\ln(2)^2}.$$

□

Proof of Theorem 5.25. As in every extension to a continuous interval before, we are going to replicate the proof of the previous result defining for that the set $A = \{0, \frac{1}{n}, \ldots, \frac{k_0}{n}\}$ where $k_0 = [n\alpha_{max}]$. We arrive to (5.41) in an identical way as how we did it in Proposition 5.26 because the type of interval does not affect the proof up to that point. As in the proposition we have that with a probability of at least $1 - \frac{1}{n}e^{-z-x_m}$

$$\begin{aligned} \sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) &\leq \sqrt{\frac{2K^2(\ln(n+1) + z + x_m)}{n(1-\alpha')^2}} \\ &+ \frac{2(1+K)}{1-\alpha'} \sqrt{\frac{4V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)^3}{n \ln(2)^2}}, \end{aligned}$$

with $\alpha' \in A$. Then for all $\alpha'' \in [0, \alpha_{max}]$ we can find $\alpha' \in A$ in such a way that $\alpha' \leq \alpha'' <$

$\alpha' + \frac{1}{n}$ and

$$\begin{aligned}
&\sup_{g\in\mathcal{G}_{m'}} (R_{\alpha''}(g) - R_{n,\alpha''}(g)) \\
&= \sup_{g\in\mathcal{G}_{m'}} ([R_{\alpha'}(g) - R_{n,\alpha'}(g)] + [R_{\alpha''}(g) - R_{\alpha'}(g)] + [R_{n,\alpha'} - R_{n,\alpha''}(g)]) \\
&\leq \sqrt{\frac{2K^2(\ln(n+1)+z+x_m)}{n(1-\alpha')^2}} + \frac{2(1+K)}{1-\alpha'}\sqrt{\frac{4V_{\mathcal{G}_{m'}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}} \\
&+ \frac{1+K}{n(1-\alpha_{max})} \\
&\leq \sqrt{\frac{2K^2(\ln(n+1)+x_m)}{n(1-\alpha')^2}} + \sqrt{\frac{2K^2 z}{n(1-\alpha')^2}} \\
&+ \frac{2(1+K)}{1-\alpha'}\sqrt{\frac{4V_{\mathcal{G}_{m'}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}} + \frac{1+K}{n(1-\alpha_{max})},
\end{aligned}$$

where we have applied Proposition 5.24 and that the trimmed generalization risk decreases when the trimming level is increased. From here, proceeding like in Proposition 5.26 we have that

$$\begin{aligned}
R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \quad &\leq \quad \sqrt{\frac{2K^2(\ln(n+1)+x_m)}{n(1-\alpha')^2}} + \sqrt{\frac{2K^2 z}{n(1-\alpha')^2}} \\
&+ \quad \frac{2(1+K)}{1-\alpha'}\sqrt{\frac{4V_{\mathcal{G}_{m'}}\ln(n+1)+\ln(2)^3}{n\ln(2)^2}} + \frac{1+K}{n(1-\alpha_{max})},
\end{aligned}$$

with a probability greater than $1 - \Sigma e^{-z}$.
Taking the penalization as in the statement and integrating with respect to z we have (5.39). □

5.3.2 Algorithm

Next we will describe the algorithm that implements the above criterion in the statistical language R. To do this, we will use the coordinate descent minimization algorithm that is described, for example, in Bühlmann and van de Geer (2011) and in section 2.5.2 and the C-Steps algorithm that can be found in Rousseeuw and Driessen (2006) and Alfons et al. (2013) and that we reviewed in section 2.5.3.

Assume we have a sample consisting of n pairs $\xi_i = (Y_i, X_i)$ where Y_i is a label that has values 1 or -1 and X_i is the attribute which has associated label Y_i and will be a vector in $\mathbb{R}^d$. The classifiers we are going to consider are linear classifiers, so we have that

$g(X_i) = X_i\beta$ where $\beta = (\beta_1, \ldots, \beta_p)$ and we will choose the family of weights $x_m = \ln(p)$. By Theorem 5.25 we know that, under certain conditions, if we choose

$$\begin{aligned}(\hat{\alpha}, \hat{m}) &= \underset{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\ldots,p\}}{\arg\min} R_{n,\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) \qquad (5.46)\\ &= \underset{(\alpha,m)\in[0,\alpha_{max}]\times\{1,\ldots,p\}}{\arg\min} \left[\min_W \sum_{i=1}^{n} w_i(1 - Y_i g(X_i))_+ + \sqrt{\frac{2K^2(\ln(n) + \ln(p))}{n(1-\alpha)^2}}\right.\\ &+ \left.\frac{2(1+K)}{1-\alpha}\sqrt{\frac{4V_{\mathcal{G}_m}\ln(n+1) + \ln(2)^3}{n\ln(2)^2}}\right]\end{aligned}$$

with $W = \left\{(w_1, \ldots, w_n) : 0 \leq w_i \leq \frac{1}{n(1-\alpha)}; \sum w_i = 1\right\}$, the empirical trimmed generalization risk for the model $\hat{m}$ will be small.

First of all, we will explain how to calculate $\min_W \sum_{i=1}^{n} w_i(1 - Y_i g(X_i))_+$ using the iterative C-Steps algorithm and in each iteration we will use the coordinate descent algorithm. Then, we will explain the method to obtain the minimum for the penalized criteria.

The first step is to apply Algorithm 7 to calculate the minimum empirical trimmed risk for a given class of linear functions. In each iteration of the method, the vector of parameters that we want to estimate is of the form

$$\hat{\beta}_k = \underset{\beta\in\mathbb{R}^m}{\arg\min}\, Q(H_k, \beta) = \underset{\beta\in\mathbb{R}^m}{\arg\min} \sum_{i\in H_k} (1 - Y_i X_i \beta)_+ \qquad (5.47)$$

and the risks that used to be quadratic residuals are now

$$r_{ki} = (1 - Y_i X_i \hat{\beta}_k)_+.$$

To calculate $\hat{\beta}_k$ in each iteration we will use Algorithm 6. The function that we are going to minimize is

$$Q(H, \beta) = \frac{1}{n}\sum_{i\in H}(1 - Y_i X_i \beta)_+,$$

which is a non-differentiable function, so we will have to work with a subgradient of this one that will be of the form

$$G_j(H, \beta) = \frac{1}{n}\sum_{i\in H}(1 - Y_i X_i) I_{(Y_i X_i \beta \leq 1)}.$$

In Friedman et al. (2007) it is proven that the coordinate descent algorithm is efficient for problems of this type leading to an optimal solution in a short time.

The aim of the algorithm is to minimize (5.46). So far we have seen how to obtain the best classifier of the trimmed sample when we have preset the level of trimming and

the class in which to look for that classifier. For computational convenience we'll look for the optimal trimming size within $A = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, \alpha_{max}\}$. As we have already seen in Proposition 5.26 the oracle inequality is valid in this case and in fact our results guarantee that, essentially, nothing is lost by restricting the search to this set.

For each element in the grid formed by A and the set $\{1, \ldots, p\}$, we will calculate (5.47) and $pen(\alpha, \mathcal{G}_m)$. To calculate the penalization we still need to fix two elements, $V_{\mathcal{G}_m}$ and K. The first of these is calculated as in Example 5.1. We will treat K as a parameter to be chosen by cross validation following the method proposed in section 7.10 in Hastie et al. (2009), dividing the sample into 5 parts. With all this we just have to calculate the sum of (5.47) and $pen(\alpha, \mathcal{G}_m)$ and choose the smallest value. The trimming level and the dimension that provides the minimum amount shall be $(\hat{\alpha}, \hat{m})$.

5.3.3 Simulations

To evaluate the performance of our algorithm, we have carried out a series of simulations. First of all, we wanted to examine the way in which the selection of the dimension works (there is no trimming involved). To do this we have simulated data from two normal d-dimensional distributions where the signal is concentrated in the first d_{real} dimensions and the rest of dimensions are noise. The first distribution is a normal one centered at $(0, \ldots, 0)$, the second one is centered at $\bar{\mu} = (\mu, \ldots, \mu, 0, \ldots, 0)$ and the two with covariance matrix the identity. As we want to prove the effectiveness of our method when samples are not linearly separable, we will take μ so that the risk of classification when we choose the Bayes rule is 0.1. If, as before, we denote this rule by g_B,

$$\begin{aligned} R(g_B) &= P(g_B(X) \neq Y) = \frac{1}{2}P(g_B(X) = -1|Y = 1) + \frac{1}{2}P(g_B(X) = 1|Y = -1) \\ &= P(g_B(X) = 1|Y = -1) = P\left(\mu^T Z > \frac{1}{2}\|\bar{\mu}\|^2\right) = P\left(\frac{\bar{\mu}^T Z}{\|\bar{\mu}\|} > \frac{1}{2}\|\mu\|\right) \\ &= 1 - \Phi(\frac{\|\bar{\mu}\|}{2}) \end{aligned}$$

where $Z \backsim N_d(0, Id)$. So we'll choose μ so that $1 - \Phi(\frac{\mu d_{real}}{\sqrt{2}}) = 0.1$.

Different combinations of sample size n, data dimension d, and effective dimension d_{real} have been considered. For each of these combinations, five datasets have been simulated with 50% of observations for each group. The optimal K has been selected by cross-validation. The results are presented in tables 5.1, 5.2, 5.3. The tables contain the estimated dimension and estimation of the optimal classifier coefficients β_i. Note that these coefficients are not unique since multiplying each by the same number would provide us with an equivalent classifier (uniqueness is given up to multiplications).

We have taken into account three possible situations, in the first of them we have the signal concentrated in 3 variables and we have added 7 noise variables (table 5.1), in this case it can be seen how, even for small samples, the estimation of the dimension is correct. We want to point out how close to the Bayes classifier the classifiers chosen by the algorithm are when the sample size is large enough. Secondly, we have kept the signal concentrated in 3 variables, but we have increased the noise to 97 variables. Despite the large increase in the size of the problem it can be observed, table 5.2, that the estimation of the real dimension is correct and, again, the coefficients β also comes close to the value of the Bayes classifier when the dimension grows. Although the increase in the number of noise variables does not affect the results obtained, it does affect the computational complexity of the problem. This is why we have not included results for very large samples as in the previous example. Finally, we wanted to consider a case in which the number of variables on which the signal is concentrated is the same as the noise variables. In this case we have considered that we have 5 variables of each type. The results obtained for this case are not as good as the previous ones for small samples. But you can see in table 5.3 that when the dimension is large enough, the estimation of the dimension is correct and the coefficients β_i approximate those of Bayes classifier.

We now deal with the simultaneous performance of optimal trimming and dimension selection. Our goal is to study the behavior of the partial classification rule when the sample contains poorly classified individuals. The main objective of the partial classification rules is to ensure a correct classification of most of the data, bearing in mind that the possible presence of outliers could seriously damage the classification capacity of the rules obtained by classical methods. A particularly important situation in practice is when the training sample contains poorly classified individuals, which is often the case in medical studies for example. The partial SVM classification ability to handle these situations is illustrated by the simulation corresponding to table 5.4. In this case the simulated data are of dimension 10 with the signal concentrated in the first 3 dimensions and the other 7 as additional noise. The data simulated with the label 1 is mostly normal of mean $(0, \ldots, 0)$ and identity covariance matrix. The simulated data with label -1 is mostly normal with mean $(2, 2, 2, 0, \ldots, 0)$ and identity covariance matrix. In both groups we have included a 10% of individuals with characteristics specific to the other group (these are our misdiagnosed). To test the utility of our algorithm, we have compared the results obtained with trimming and dimension selection with those obtained when we do not allow trimming and only allow to select the dimension as in the previous examples.

Table 5.4 has a similar structure to the previous ones but includes, in addition to the

Table 5.1: Dimension selection with 3 real variables and 7 of noise

n	d	real d	estimated d	Beta
200	10	3	2	-1.67 1.28 1.04
			3	-1.88 0.70 0.80 0.86
			3	-2.25 1.08 1.18 0.86
			3	-2.44 0.76 1.13 1.19
			3	-2.40 1.04 0.79 1.16
500	10	3	3	-2.24 0.95 1.05 1.08
			3	-2.55 1.16 1.07 0.87
			3	-2.40 0.86 0.93 1.06
			3	-2.31 0.98 1.04 0.88
			3	-2.10 1.03 1.02 1.06
1000	10	3	3	-2.10 0.95 0.93 1.10
			3	-2.37 1.02 0.92 1.05
			3	-2.34 1.04 0.97 1.00
			3	-2.36 1.04 0.92 1.07
			3	-2.44 0.96 0.96 1.04
2000	10	3	3	-2.46 0.95 1.06 0.98
			3	-2.51 0.99 1.05 0.93
			3	-2.37 1.00 1.02 1.06
			3	-2.41 0.95 1.03 1.02
			3	-2.41 0.92 1.04 1.07
5000	10	3	3	-2.39 1.04 0.95 0.99
			3	-2.39 0.97 0.99 1.03
			3	-2.37 1.01 1.00 1.00
			3	-2.41 0.96 1.02 1.03
			3	-2.37 1.03 0.97 0.97
10000	10	3	3	-2.37 1.01 1.00 1.01
			3	-2.37 1.01 0.98 1.01
			3	-2.35 1.02 0.99 1.00
			3	-2.40 0.97 0.99 1.04
			3	-2.41 1.01 0.99 0.99

Table 5.2: Dimension selection with 3 real variables and 97 of noise

n	d	real d	estimated d	Beta
200	100	3	3	-2.64 0.80 1.04 1.26
			3	-2.84 1.16 1.14 1.18
			3	-2.06 0.74 1.01 0.74
			3	-1.97 0.63 1.11 1.06
			3	-2.16 0.81 1.24 0.80
500	100	3	3	-2.36 0.92 0.96 1.10
			3	-2.43 1.06 1.01 1.11
			3	-2.68 1.25 1.08 0.94
			3	-2.41 1.00 1.06 0.98
			3	-2.64 1.08 0.97 1.28
1000	100	3	3	-2.41 0.93 0.96 1.14
			3	-2.46 0.99 1.07 0.95
			3	-2.39 1.06 1.02 0.86
			3	-2.32 1.05 0.88 0.90
			3	-2.17 0.97 0.98 1.03
2000	100	3	3	-2.37 0.99 1.01 1.00
			3	-2.31 0.98 0.99 0.99
			3	-2.36 0.97 1.03 0.97
			3	-2.45 0.99 1.05 0.98
			3	-2.41 0.91 1.08 0.98
5000	100	3	3	-2.39 0.99 1.04 0.97
			3	-2.37 0.99 1.02 0.99
			3	-2.35 1.00 0.99 1.00
			3	-2.41 1.09 0.94 0.96
			3	-2.44 0.98 1.03 0.96

Table 5.3: Dimension selection with 5 real variables and 5 of noise

n	d	real d	estimated d	Beta
200	10	5	4	-3.61 1.36 1.28 0.93 1.27
			5	-3.64 0.94 1.00 0.85 1.09 1.19
			4	-3.91 1.38 0.94 1.00 1.10
			4	-3.40 0.88 0.91 1.14 1.11
			3	-2.27 1.03 1.14 1.03
500	10	5	4	-3.27 0.99 1.15 1.03 0.82
			5	-4.00 1.16 0.98 1.03 1.10 0.89
			3	-2.50 1.14 1.01 1.13
			5	-4.23 1.00 0.96 0.95 1.19 1.05
			5	-4.54 1.12 1.19 1.08 1.00 1.01
1000	10	5	5	-4.04 1.06 0.98 1.12 0.98 1.13
			5	-4.12 0.79 1.03 0.92 1.00 1.20
			5	-3.70 0.94 0.82 0.90 1.07 0.96
			5	-3.74 1.05 1.13 0.85 0.85 1.05
			5	-3.77 0.85 1.15 0.90 0.81 1.01
2000	10	5	5	-4.01 1.13 0.84 1.03 1.05 0.87
			5	-3.89 1.01 0.96 0.99 1.03 0.92
			5	-4.09 1.06 0.99 0.98 0.96 1.00
			5	-3.92 0.95 0.89 1.05 1.04 1.12
			5	-3.94 0.97 1.03 0.93 0.99 1.03
5000	10	5	5	-4.11 1.01 0.98 1.01 1.02 0.96
			5	-3.93 1.00 0.98 0.99 0.99 0.99
			5	-4.03 0.93 0.99 1.04 0.98 0.99
			5	-3.95 0.98 1.05 0.95 0.97 0.97
			5	-4.00 1.04 1.04 0.99 0.96 1.01
10000	10	5	5	-4.10 0.99 1.00 0.98 1.01 1.01
			5	-3.97 1.01 0.95 1.02 0.96 1.00
			5	-3.99 0.99 1.00 0.93 1.06 0.98
			5	-3.97 1.03 1.00 0.95 1.00 0.96
			5	-3.96 0.99 0.96 0.99 0.99 0.98

Table 5.4: Optimal trimming and dimension selection with misclassified individuals

n	No trimming allowed			With trimming			
	est d	Emp R	Beta	est d	trim	Emp R	Beta
500	2	0.484	-1.95 0.74 0.95	4	0.06	0.193	-2.80 1.07 0.92 1.11 -0.43
	2	0.495	-2.02 0.76 0.91	3	0.06	0.202	-2.79 0.77 0.92 0.85
	2	0.464	-2.00 0.74 0.92	3	0.08	0.130	-3.06 0.89 1.15 0.92
	2	0.461	-1.99 0.80 0.96	3	0.08	0.106	-3.14 0.84 1.13 0.92
	2	0.441	-1.88 0.86 0.76	3	0.1	0.080	-2.87 1.30 0.92 0.93
1000	2	0.483	-1.91 0.73 0.96	3	0.08	0.128	-2.81 0.92 1.04 0.92
	1	0.549	-1.02 1.02	3	0.08	0.121	-3.14 1.09 1.01 0.91
	1	0.556	-1.11 1.00	3	0.08	0.133	-2.90 1.00 0.93 0.97
	2	0.474	-1.95 0.81 0.90	3	0.1	0.079	-3.11 1.12 1.04 0.99
	2	0.470	-2.06 0.78 0.93	3	0.08	0.116	-3.27 0.97 0.98 1.14
2000	2	0.457	-1.97 0.82 0.88	3	0.1	0.072	-3.03 1.13 0.92 1.00
	2	0.479	-2.05 0.78 0.93	3	0.06	0.194	-3.22 0.91 1.10 0.91
	2	0.459	-1.96 0.79 0.93	3	0.08	0.135	-3.10 0.98 0.93 1.04
	2	0.459	-1.97 0.77 0.94	3	0.06	0.194	-3.13 0.91 1.08 0.93
	1	0.510	-0.98 0.95	3	0.08	0.120	-2.76 0.97 0.95 0.98
5000	2	0.463	-2.02 0.85 0.89	3	0.08	0.124	-3.06 1.08 0.97 0.98
	2	0.489	-2.04 0.82 0.91	3	0.1	0.090	-3.06 1.06 0.98 1.06
	2	0.466	-1.99 0.78 0.93	3	0.1	0.080	-2.96 1.03 1.04 0.99
	2	0.465	-2.03 0.81 0.93	3	0.06	0.184	-3.15 0.93 1.02 0.93
	2	0.467	-1.95 0.79 0.95	3	0.06	0.181	-3.01 0.95 1.05 0.96

estimated dimension and estimated coefficients, the empirical risk (which we recall is a superior bound for the generalization error) associated with the chosen rule and the level of trimming selected in the case of partial classification. In the application of the procedure, $\alpha_{max} = 0.12$ has been set. This table shows that the untrimmed SVM classification can produce very poor results with an empirical risk, which we use as an approximation of the population risk, around 0.5. This is an indicator of how bad classical classification can be in this scenario. In addition, we can see that even for large sample sizes the dimension is not selected correctly. On the other hand, we see that when we allow ourselves to trim a part of the sample, the classification rule is close to the optimal one for large samples and, in any case, a correct estimation of the dimension is made. Note that not in all cases the algorithm needs to remove all contamination to provide good results and that in none of

the cases does the level of trimming exceed that of contamination.

Finally, we wanted to check how the algorithm works under a different type of contamination. We now consider a design similar to the previous one but with the same contamination in both groups. In a medical study this would correspond to the situation in which certain variables have predictive capacity for most individuals but not for a small subgroup. More specifically, we will consider that the simulated data comes from the same normal distributions as in the previous case. Now the contamination, which remains at 10% for each population, comes from a normal distribution centered in $(1, 1, 1, 0, \ldots, 0)$ and with covariance matrix 3 times the identity. As before, we want to see if the results obtained with the algorithm are good and, to do this, we will compare them with those obtained if we do not allow trimmings.

Table 5.5: Optimal trimming and dimension selection with contamination

n	No trimming allowed			With trimming			
	est d	Emp R	Beta	est d	Trim	Emp R	Beta
500	2	0.340	-1.90 0.86 0.95	3	0.08	0.061	-2.81 1.01 1.16 0.89
	1	0.446	-0.97 1.12	3	0.02	0.181	-2.73 0.89 0.77 0.95
	3	0.247	-2.83 0.97 0.83 1.01	3	0.02	0.161	-3.02 1.13 0.97 1.08
	2	0.338	-2.30 1.04 0.95	3	0.02	0.054	-2.96 1.29 0.74 0.94
	3	0.268	-3.07 0.84 0.94 1.11	3	0.02	0.168	-3.07 0.95 0.97 1.15
1000	3	0.256	-3.19 0.96 0.84 0.97	3	0.04	0.113	-3.195 1.08 0.88 1.01
	3	0.247	-2.89 0.79 0.99 1.01	3	0.02	0.154	-2.84 0.89 0.99 1.06
	3	0.275	-3.14 0.96 0.86 0.93	3	0.06	0.059	-3.14 1.07 1.09 0.97
	3	0.224	-2.23 0.73 0.72 0.79	3	0.04	0.088	-2.89 1.01 1.15 0.97
	3	0.292	-2.89 0.76 0.84 1.03	3	0.02	0.182	-2.89 0.86 0.89 1.11
2000	2	0.327	-2.07 0.92 0.99	3	0.02	0.159	-3.19 1.06 0.96 1.04
	3	0.269	-3.01 0.81 0.95 1.01	3	0.04	0.106	-3.00 0.97 1.06 1.03
	3	0.276	-3.07 0.87 0.98 0.97	3	0.04	0.113	-2.96 1.03 1.01 0.98
	3	0.294	-2.95 0.84 0.97 0.96	3	0.04	0.119	-2.94 1.05 1.01 1.04
	2	0.359	-2.01 0.93 0.93	3	0.02	0.171	-3.01 0.97 0.93 1.07
5000	3	0.274	-3.03 0.85 0.92 0.99	3	0.1	0.026	-2.98 1.15 1.00 0.97
	3	0.271	-3.04 0.86 0.92 0.97	3	0.04	0.115	-3.03 1.00 1.04 1.03
	3	0.263	-3.00 0.83 0.99 0.91	3	0.04	0.109	-3.00 1.00 1.03 1.02
	3	0.276	-2.95 0.88 0.94 0.97	3	0.04	0.118	-2.93 1.04 1.03 1.01
	3	0.268	-2.98 0.88 0.91 1.01	3	0.02	0.163	-2.88 0.97 0.92 1.04

The results shown in table 5.5 are similar to those in table 5.4. Classical methods have

more difficulty in estimating the correct signal size and produce worse results in terms of both estimated risk and the quality of the estimation of the optimal rule.

5.3.4 Example with real data

Once the method has been tested with simulated data, we will use it with a real sample. We are going to analyze a set of medical data in which 16 features of 43 patients with liver disease are analyzed. This data set was provided by Professor Remy Burcelin of the Institute of Cardiovascular and Metabolic Diseases at Paul Sabatier University in Toulouse. This is a pilot study and the sample size is probably too small for partial classification methods to produce significant benefits. The collaboration with Professor Remy Burcelin continues, but for the time it was a question of assessing the potential of the partial SVM method for detecting incorrect diagnoses. We would like to thank prof. Burcelin for the opportunity he has provided to show the applicability of our methods on real data.

Patients are classified in two groups, those who have fibrosis ($Y = 1$) and those who have a mild form of the disease or no fibrosis at all ($Y = -1$). This disease is characterized by episodes of illness interspersed with episodes of good health, so there are often individuals who have been misclassified. We want to analyze the data to determine which variables we should consider and which individuals are misclassified and likely to be re-diagnosed.

Since our algorithm is designed in such a way that it only considers keeping subsets of variables arranged in ascending order (for reasons of computational efficiency), we will reorder the variables in such a way that we will first consider those that have the least apparent error when considering the model in which we classify only one variable. Once the variables have been rearranged, we will scale them since each one is measured on a different scale and the variables that are measured in larger magnitudes will have a much greater risk associated with the hinge loss.

When it comes to normalizing variables, since we do not want the solution to be affected by outliers, we have used the median instead of the mean and the interquartile range instead of the standard deviation.

After the described normalization we have applied the partial SVM classification algorithm with $\alpha_{max} = 0.2$ and possible trimming levels in $A = \{0, 0.01, 0.02, \ldots, 0.2\}$. The parameter K in the penalization has been chosen by cross-validation in the same way as in previous simulation studies. As mentioned above, the sample size seems too small to guarantee a good performance of the method. The chosen rule retains 13 variables (only 3 are deleted). The estimation of the coefficients of the selected rule is probably very

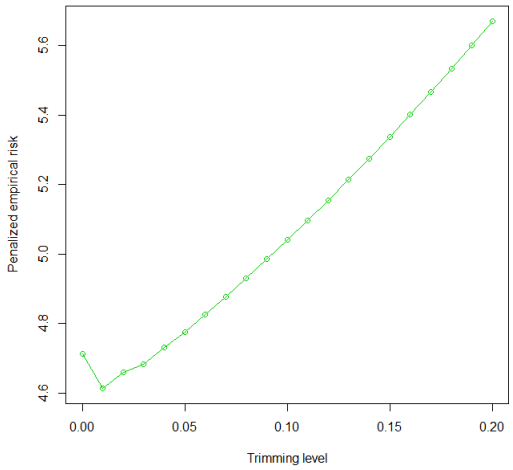


Figure 5.1: Minimal penalized risk for each trimming level in A.

unstable. On the positive side, we note that the partial SVM method has allowed one of the individuals in the study to be trimmed. In Figure 5.1, shows the minimum penalized empirical risk that is reached with each trimming level in A. Analyzing the individual trimmed a posteriori we have been able to verify that it presented incorrect values for some of the variables such as, for example, a cholesterol of 0mg/dL. Therefore the partial SVM method has succeeded in detecting an individual whose values do not correspond to their real values even for such a small sample.

5.4 Extensions

In this section we include a series of alternative approaches to the problems seen throughout this chapter that we have considered and that we have either discarded for not giving convenient results, or we leave as work to develop in the future.

First, we have considered another way of choosing the penalty based on the trimmed model and not on the trimming level. What we are proposing is to move from the paradigm of Massart (2007) to the LASSO style of Bühlmann and van de Geer (2011). The reason for considering this alternative is the computational advantage, since at the same time that a single estimator calculation is made, the model selection is made. To study the behavior of the problem we will focus, as before, on the simple case in which we have a single g classifier. In this case, we will look for a trimming Q that minimizes the Q-risk plus a penalization related to the amount of deviation of Q from P.

So far we have considered the problem of calculating

$$\begin{aligned} \hat{\alpha} &= \min_{\alpha} R_\alpha(g) + Pen(\alpha), \quad \text{where} \\ R_\alpha(g) &= \min_{Q \in \mathcal{R}_\alpha(P)} E_Q(\ell(g; Y, X)), \end{aligned}$$

with $\ell(g; Y, X)$ a positive loss function. Now, the problem we are going to consider is of the form

$$\min_{Q \in \mathcal{R}_{\alpha_{max}}(P)} E_Q(\ell(g; Y, X)) + Pen(Q),$$

where the penalization we consider will be a measure of the discrepancy between P and Q and α_{max} a pre-set value in $(0, 1]$. As we have done throughout the chapter we will work with the empirical version of the problem, that is,

$$\min_{(w_1,\ldots,w_n)} \sum_{i=1}^{n} w_i \ell(g; y_i, x_i) + Pen(w_1, \ldots, w_n),$$

where the vector $(w_1, \ldots, w_n)$ consists of the weights associated with the probability $\sum_{i=1}^{n} w_i \delta_{(y_i,x_i)}$.

We have explored the possible advantages or disadvantages of this approach by considering two different penalties to measure the similarity between P and Q, the Kullback divergence and the Wasserstein distance.

5.4.1 Penalization based on Kullback's divergence

Kullback's divergence,

$$K(Q, P) = \int \ln\left(\frac{dQ}{dP}\right) dQ,$$

is one of the most widely used measures in statistics to assess deviation from a model. In this section we will study the problem of minimizing

$$E_Q(\ell(g; Y, X)) + LK(Q, P)$$

and, concretely, we will study the behaviour of the empirical version of this criteria, i.e., the problem of minimizing

$$\sum_{i=1}^{n} w_i \ell(g; y_i, x_i) + LK(W, P_n),$$

with $W = (w_1, \ldots, w_n)$. Here, trimmings of P_n are probabilities with the same support as P_n which give a mass of w_i to the point (y_i, x_i) and can be codified by the vector W. Now

$$K(W, P_n) = \sum_{i=1}^{n} w_i \ln(w_i) + \ln(n),$$

in the following we ignore the term $\ln(n)$ in our calculations, this will not affect the final result. In our analysis we have set $\alpha_{max} = 1$.

Obviously we can express the penalization function as the sum of n separated functions each of them depending only in the value of a single w_i. Hence, we may apply the decomposition principle described in Rockafellar (1997) to obtain a simple expression of the optimal weights. Now we are going to describe the principal aspects in that theory.

The decomposition principle we are talking about has to do with the general problem of minimizing a convex function with convex restrictions. More concretely, we consider the optimization problem

$$\begin{array}{lll} \text{(P1)} \ \min\limits_{x \in C} & f_0(x) & \\ \text{s.t.} & f_j(x) \leq 0, & j = 1, \ldots, r, \\ & f_j(x) = 0, & j = r+1, \ldots, m, \end{array}$$

where $f_0, f_1, \ldots, f_m$ are convex functions (and, by simplicity, we assume here that are also differentiable). The Lagrangian associated to problem (P1) is

$$L(\lambda, x) = \begin{cases} f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x) \text{ if } \lambda \in E_r, x \in C \\ -\infty \text{ if } \lambda \notin E_r, x \in C \\ +\infty \text{ if } x \notin C \end{cases},$$

where the convex set C is the domain of f_0 and

$$E_r = \{\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m : \lambda_i \geq 0, \ i = 1, \ldots, r\}.$$

It is said that a vector λ is a Kuhn-Tucker vector if it belongs to E_r and, besides, has the property that the minimum of $f_0(x)$ with the restrictions is attained in the same points as the minimum of $f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)$. Each variable λ_i is known as the Lagrange multiplier associated with the ith restriction in (P1).

The following result which establishes the existence of a Kuhn-Tucker vector is Corollary 28.2.2 in Rockafellar (1997).

Corollary 5.28. *Let (P) be an ordinary convex optimization problem with only linear restrictions. If the optimal value of (P) is not $-\infty$ and (P) has a feasible solution in the relative interior of C, with C the domain of the function we are optimizing, then there exists a Kuhn-Tucker vector for (P).*

With the following result, that corresponds to Theorem 28.3 in Rockafellar (1997), sufficient and necessary conditions for the optimality of $\bar{x} \in \mathbb{R}^d$ for (P1) are established.

Theorem 5.29. *Take $\lambda \in \mathbb{R}^m$ and $\bar{x} \in \mathbb{R}^d$. A necessary and sufficient condition for λ to be a Kuhn-Tucker vector and for $\bar{x}$ to be an optimal solution for (P1), is that $(\lambda, \bar{x})$ satisfies Kuhn-Tucker's optimality conditions, this is,*

$$\begin{aligned}
&\lambda_i \geq 0,\ f_i(\bar{x}) \leq 0 \text{ and } \lambda_i f_i(\bar{x}) = 0,\ i = 1, \ldots, r,\\
&f_i(\bar{x}) = 0,\ i = r+1, \ldots, m,\\
&0 \in [\partial f_0(\bar{x}) + \lambda_1 \partial f_1(\bar{x}) + \cdots + \lambda_m \partial f_m(\bar{x})].
\end{aligned}$$

As we have guaranteed the existence of a Kuhn-Tucker vector by Corollary 5.28, we have to look for a vector $\bar{x} \in \mathbb{R}^d$ that satisfies the conditions in the theorem.

We are focusing now in the simpler problem in which both the restrictions and the objective function in (P1) are separable functions. This means, $f_i(x) = q_{i1}(x_1) + \cdots + q_{in}(x_n)$ with $i = 0, \ldots, m$. Now (P1) is equivalent to

$$\min_{x \in \mathbb{R}^n} f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)$$

without restrictions, where $\lambda = (\lambda_1, \ldots, \lambda_m)$ is a Kuhn-Tucker vector. Take $\lambda_0 = 1$, then

$$\begin{aligned}
f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x) &= \sum_{i=0}^{m} \sum_{j=1}^{n} \lambda_i q_{ij}(x_j)\\
&= (q_{01}(x_1) + \lambda_1 q_{11}(x_1) + \cdots + \lambda_m q_{m1}(x_1)) + \cdots\\
&+ (q_{0n}(x_n) + \lambda_1 q_{1n}(x_n) + \cdots + \lambda_m q_{mn}(x_n)).
\end{aligned}$$

This establish a decomposition principle. The optimization problem over $\mathbb{R}^n$ can be reduced to n independent optimization problems in $\mathbb{R}$.

We apply this idea to the minimization of a convex function of separated variables in the positive octant over the simplex $\{x : \sum x_i = 1, x_i \geq 0\}$. For dividing the problem it is used the method introduced in chapter 28 in Rockafellar (1997). Define the functions

$$f_{0j}(x_j) = \begin{cases} q_{0j}(x_j) & \text{if} \quad x_j \geq 0 \\ \\ +\infty & \text{if} \quad x_j < 0 \end{cases} \tag{5.48}$$

for each $j \in \{1, \ldots, n\}$ and

$$f_{1j}(x_j) = x_j \qquad \text{with } j = 1, \ldots, n-1, \tag{5.49}$$

$$f_{1n}(x_n) = x_n - 1. \tag{5.50}$$

We can express problem (P1) in terms of this functions,

$$\begin{array}{lll} (\mathrm{P}) & \min\limits_{x\in\mathbb{R}^n} & f_0(x) = f_{01}(x_1) + \cdots + f_{0n}(x_n) \\ & \text{s.t.} & f_1(x) = f_{11}(x_1) + \cdots + f_{1n}(x_n) = 0. \end{array}$$

Given that the infimum of (P) is finite and that in the interior of the domain of f_0 exist points which make $f_1(x) = 0$ we can apply Corollary 5.28 to ensure the existence of a Kuhn-Tucker vector, that, in this case in which we only have a restriction, is formed by a single coefficient λ. This coefficient is, see page 288 in Rockafellar (1997),

$$\lambda = \arg\min_v \{v + f_{01}^*(-v) + \ldots + f_{0n}^*(-v)\}, \tag{5.51}$$

where for each $j = 1, \ldots, n$

$$f_{0j}^*(z) = \sup_{x\in\mathbb{R}} \{xz - f_{0j}(x)\}. \tag{5.52}$$

If the coefficient λ can be calculated, solving the problem (P1) is equivalent to solving n problems of the form

$$\min_{x_j} \{f_{0j}(x_j) + \lambda f_{1j}(x_j)\}.$$

We apply now this principle to our problem in which we have to minimize the function

$$H_K(w_1, \ldots, w_n) = \sum_{i=1}^{n} \left(w_i \ell(g; y_i, x_i) + L w_i \ln(w_i)\right).$$

The optimization problem is now of the form

$$\begin{array}{lll} (\mathrm{P}) & \min\limits_{(w_1,\ldots,w_n)\in\mathbb{R}^n} & H_K(w_1, \ldots, w_n) = q_1(w_1) + \cdots + q_n(w_n) \\ & \text{s.t.} & w_1 + \cdots + w_n - 1 = 0, \\ & & w_i \geq 0,\ i = 1, \ldots, n, \end{array}$$

with

$$q_j(w_j) = w_j c_j + L w_j \ln(w_j)$$

and $c_j = \ell(g; y_j, x_j)$ with $y_j \in \mathbb{R}$ and $x_j \in \mathbb{R}^d$.

Functions (5.48) now will be

$$f_{0j}(w_j) = \begin{cases} w_j c_j + L w_j \ln(w_j) & \text{if} \quad w_j \geq 0 \\ \\ +\infty & \text{if} \quad w_j < 0 \end{cases},$$

while functions (5.49) and (5.50) remain unchanged. To calculate Kuhn-Tucker's coefficient we need to calculate first (5.52):

$$f_{0j}^*(z) = \sup_{x \geq 0}\{xz - xc_j - Lx\ln(x)\}.$$

If we call $i(x) := xz - xc_j - Lx\ln(x)$,

$$\frac{\partial i(x)}{\partial x} = z - c_j - L(\ln(x) + 1) = 0 \text{ if and only if } \hat{x} = e^{\frac{z-c_j}{L}-1}.$$

Obviously $\hat{x} \geq 0$, hence

$$f_{0j}^*(z) = ze^{\frac{z-c_j}{L}-1} - c_j e^{\frac{z-c_j}{L}-1} - Le^{\frac{z-c_j}{L}-1}\left(\frac{z-c_j}{L} - 1\right)$$

and

$$f_{0j}^*(-\lambda) = -\lambda e^{\frac{-\lambda-c_j}{L}-1} - c_j e^{\frac{-\lambda-c_j}{L}-1} - Le^{\frac{-\lambda-c_j}{L}-1}\left(\frac{-\lambda-c_j}{L} - 1\right).$$

last, we have to calculate (5.51). Taking $l(v) := v + f_{01}^*(-v) + \ldots + f_{0n}^*(-v)$ then

$$\begin{aligned}
\frac{\partial l(\lambda)}{\partial \lambda} &= 1 - \sum_{j=1}^{n}\left[e^{\frac{-\lambda-c_j}{L}-1} + \frac{\lambda}{L}e^{\frac{-\lambda-c_j}{L}-1} - \frac{c_j}{L}e^{\frac{-\lambda-c_j}{L}-1}\right. \\
&+ \left. e^{\frac{-\lambda-c_j}{L}-1}\left(\frac{-\lambda-c_j}{L} - 1\right) - e^{\frac{-\lambda-c_j}{L}-1}\right] \\
&= 1 - \sum_{j=1}^{n} e^{\frac{-\lambda-c_j}{L}-1}\left[-\frac{\lambda}{L} - \frac{c_j}{L} + \frac{\lambda}{L} + \frac{c_j}{L} + 1\right] = 1 - \sum_{j=1}^{n} e^{\frac{-\lambda-c_j}{L}-1} \\
&= 1 - \frac{e^{\frac{-\lambda}{L}}}{e}\sum_{j=1}^{n} e^{\frac{-c_j}{L}} = 0.
\end{aligned}$$

Solving this equation we have that the Lagrange multiplier for our problem is

$$\lambda^* = L\left(\ln\left(\sum_{j=1}^{n} e^{\frac{-c_j}{L}}\right) - 1\right).$$

Now we can solve the n equivalent problems. For $j \in \{1, \ldots, n-1\}$, we want to calculate

$$\min_{w_j \geq 0} w_j c_j + Lw_j\ln(w_j) + w_j L\left(\ln\left(\sum_{i=1}^{n} e^{\frac{-c_i}{L}}\right) - 1\right).$$

If we call $h(x)$ the function we want to minimize,

$$\frac{\partial h(w_j)}{\partial w_j} = c_j + L(\ln(w_j) + 1) + L\left(\ln\left(\sum_{i=1}^{n} e^{\frac{-c_i}{L}}\right) - 1\right) = 0 \longleftrightarrow \widehat{w}_j = \frac{e^{\frac{-c_j}{L}}}{\sum_{i=1}^{n} e^{\frac{-c_i}{L}}}.$$

As $f_{1n(w_n)}$ is different from the rest, the function to minimize for $j = n$ is

$$\min_{w_n \geq 0} w_n c_j + L w_n \ln(w_n) + (w_n - 1) L \left(\ln \left(\sum_{i=1}^{n} e^{\frac{-c_i}{L}} \right) - 1 \right).$$

But, as what we add to the function is constant with respect to w_n, when we derive it will vanish and we will obtain the same result for the optimum of the nth observation. Hence

$$\widehat{w}_j = \frac{e^{\frac{-c_j}{L}}}{\sum_{i=1}^{n} e^{\frac{-c_i}{L}}} \geq 0$$

for $j = 1, \ldots, n$.

Given that the value of the weights in the optimum is determined by an exponential function, it does not matter how bad an observation is, this procedure will never assign a weight 0 and, hence, it will never be completely removed. We conclude that a penalization based on Kullback's divergence does not seem a good election for detecting and eliminating outliers. In the case of classification with 0/1 loss we can see more clearly that this type of penalization does not remove atypical observations. For this reason we abandon this line and chose to study another type of penalizations.

5.4.2 Penalization based on Wasserstein distance

Once we have discarded the option of Kullback's divergence, we intend to use as a penalization Wasserstein distance between distributions P_X and Q_X that are the marginal distributions for the attributes. This is, if $B \in \mathbb{R}^d$, then

$$P_X(B) = P(\{(y, x) : x \in B\}) = P(\{(0, x) : x \in B\}) + P(\{(1, x) : x \in B\}),$$

and, if we define the object (p_0, P_0, P_1) as in Lemma 5.1, then

$$P_X(B) = p_0 P_0(B) + (1 - p_0) P_1(B).$$

In this section we are working with the *penalized trimmed generalization error* associated to a probability $Q \in \mathcal{R}_\alpha(P)$ for a given α that we define as

$$R(Q) := Q(\{(y, x) : g(x) \neq y\}) + L \mathcal{W}_2^2(P_X, Q_X) \tag{5.53}$$

where L is a real positive factor and g is a given classification function.

We want to see that exists a probability $Q \in \mathcal{R}_\alpha(P)$ that minimizes this quantity and, besides, that this probability is unique. Observe that there is not a probability Q different

from P in $\mathcal{R}_\alpha(P)$ such that $R(Q) = 0$, except that $\exists Q \in \mathcal{R}_\alpha(P)$ with $Q_X = P_X$ and $Q(\{(y,x) : g(x) \neq y\}) = 0$.

As we can obtain a null penalized trimmed generalization error now we will see that we can find a probability that minimizes that error.

Proposition 5.30. *If P is such that $P(\partial A_g) = 0$ with $A_g = \{(y,x) : g(x) \neq y\}$, then the map $Q \mapsto R(Q)$ is continuous in $\mathcal{R}_\alpha(P)$ with respect to $\mathcal{W}_2$ and convex in $\mathcal{F}_2(\{0,1\} \times \mathbb{R}^d)$ (the set of probabilities in $\{0,1\} \times \mathbb{R}^d$ with finite second moment). Even more, if for every $Q_1, Q_2 \in \mathcal{R}_\alpha(P)$ we have that $Q_{1,X} = Q_{2,X} \Rightarrow Q_1 = Q_2$ the convexity will be strict.*

Observe that the condition $P(\partial A_g) = 0$ is equivalent to $P_0(\partial\{x : g(x) \neq 0\}) = 0$ and $P_1(\partial\{x : g(x) \neq 1\}) = 0$ and this conditions is met if g is linear and P_0 and P_1 have a density.

To simplify the proof of the previous result we need first the following lemma.

Lemma 5.31. *Let Q, Q_n be two probabilities in $\{0,1\} \times \mathbb{R}^d$ with $Q \equiv (q_0, Q_0, Q_1)$ and $Q_n \equiv (q_{n,0}, Q_{n,0}, Q_{n,1})$ as in Lemma 5.1, then*

$$Q_n \to^{\mathcal{W}_2} Q \iff \begin{cases} q_{n,0} \to q_0 \\ Q_{n,0} \to^{\mathcal{W}_2} Q_0 \\ Q_{n,1} \to^{\mathcal{W}_2} Q_1 \end{cases} \tag{5.54}$$

Proof.

($\Leftarrow$)

For the proof of this implication we will apply Proposition 2.7. The conditions over $Q_{n,0}$ and $Q_{n,1}$ imply that for $i = 0, 1$,

$$\int_{\mathbb{R}^d} \|x\|^2 dQ_{n,i}(x) \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^d} \|x\|^2 dQ_i(x).$$

By (2.2),

$$\begin{aligned} \int_{\{0,1\}\times\mathbb{R}^d} \|(y,x)\|^2 dQ_n(y,x) &= q_{n,0} \int_{\mathbb{R}^d} \|x\|^2 dQ_{n,0}(x) \\ &+ (1 - q_{n,0}) \int_{\mathbb{R}^d} (\|x\|^2 + 1) dQ_{n,1}(x) \\ &\overset{n\to\infty}{\longrightarrow} \int_{\{0,1\}\times\mathbb{R}^d} \|(y,x)\|^2 dQ(y,x). \end{aligned}$$

For applying Proposition 2.7 we still have to verify that $Q_n \rightharpoonup Q$. For this we will apply the definition of weak convergence,

$$Q_n \rightharpoonup Q \text{ if and only if } \int_{\{0,1\}\times\mathbb{R}^d} h dQ_n \longrightarrow \int_{\{0,1\}\times\mathbb{R}^d} h dQ \quad \forall h \text{ bounded and continuous.}$$

Applying decomposition (2.2),

$$\begin{aligned}\int_{\{0,1\}\times\mathbb{R}^d} h dQ_n &= q_{0,n}\int_{\mathbb{R}^d} h(0,x)dQ_{n,0}(x) + (1-q_{0,n})\int_{\mathbb{R}^d} h(1,x)dQ_{n,1}(x)\\ &\longrightarrow q_0\int_{\mathbb{R}^d} h(0,x)dQ_0(x) + (1-q_0)\int_{\mathbb{R}^d} h(1,x)dQ_1(x)\\ &= \int_{\{0,1\}\times\mathbb{R}^d} h dQ.\end{aligned}$$

($\Rightarrow$)
AS $Q_n \to^{W_2} Q$ we have that

$$\int_{\{0,1\}\times\mathbb{R}^d} h(y,x)dQ_n(y,x) \longrightarrow \int_{\{0,1\}\times\mathbb{R}^d} h(y,x)dQ(y,x) \tag{5.55}$$

for every function h bounded and continuous and, concretely, for $h(y,x) = \|(y,x)\|^2$.
On the other hand

$$\int_{\{0,1\}\times\mathbb{R}^d} h(y,x)dQ(y,x) = q_0\int_{\mathbb{R}^d} h(0,x)dQ_0(x) + (1-q_0)\int_{\mathbb{R}^d} h(1,x)dQ_1(x)$$

and the same stands for Q_n.
We take $h(y,x) = \max(\min(1-y,1),0)$ which is bounded and continuous. Then (5.55) implies that $q_{0,n} \to q_0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be any bounded and continuous function and, now, $h(y,x) = \max(\min(1-y,1),0)f(x)$ which is bounded and continuous in $\{0,1\}\times\mathbb{R}^d$. Besides, $h(0,x) = f(x)$ and $h(1,x) = 0$, which by (5.55) implies that

$$q_{0,n}\int_{\mathbb{R}^d} f(x)dQ_{0,n}(x) \longrightarrow q_0\int_{\mathbb{R}^d} f(x)dQ_0(x).$$

In this way, given that $q_0 > 0$, we can deduce that $Q_{0,n} \rightharpoonup Q_0$. In a similar way we can prove that

$$q_{0,n}\int_{\mathbb{R}^d} \|x\|^2 dQ_{0,n}(x) \longrightarrow q_0\int_{\mathbb{R}^d} \|x\|^2 dQ_0(x)$$

and we conclude that $Q_{0,n}\mathcal{W}_2 Q_0$. Convergence $Q_{1,n} \to^{\mathcal{W}_2} Q_1$ is proved in a similar way. □

Proof of Proposition 5.30. First we prove the convexity. The map $Q \mapsto Q(\{(y,x) : g(x) \neq y\})$ es lineal. Hence, it suffices to prove that the map $Q \mapsto \mathcal{W}_2^2(P_X, Q_X)$ is strictly convex. Given that P_X has a density by hypothesis, the map $\tilde{Q} \mapsto \mathcal{W}_2^2(P_X, \tilde{Q})$ is strictly convex in Q (see Álvarez-Esteban et al. (2011)). Hence, for each $\gamma \in (0,1)$ and $Q_1, Q_2 \in \mathcal{R}_\alpha(P)$,

$$\mathcal{W}_2^2(P_X, \gamma Q_{1,X} + (1-\gamma)Q_{2,X}) \leq \gamma\mathcal{W}_2^2(P_X, Q_{1,X}) + (1-\gamma)\mathcal{W}_2^2(P_X, Q_{2,X})$$

and the inequality is strict except if $Q_{1,X} = Q_{2,X}$. But in this case we will have $Q_1 = Q_2$. This proves that R is strictly convex in $\mathcal{R}_\alpha(P)$.
To prove the continuity of R we will prove the continuity of each of the terms in the sum separately. Let us see first the continuity for $\mathcal{W}_2^2(P,.)$. Let us define the mapping

$$H : Q \mapsto \mathcal{W}_2(P,Q),$$

we want to prove that if $\mathcal{W}_2(Q,Q_n) \to 0$ then $H(Q_n) \to H(Q)$. Obviously

$$|H(Q_n) - H(Q)| = |\mathcal{W}_2(Q_n,P) - \mathcal{W}_2(Q,P))| \leq |\mathcal{W}_2(Q_n,Q)|$$

and the continuity of that part is proved.
On the other hand, condition $P(\partial A_g) = 0$ implies that $Q(\partial A_g) = 0$ for every $Q \in \mathcal{R}_\alpha(P)$ (for every $\alpha < 1$). Then if $Q_n \rightharpoonup Q$ where $Q \in \mathcal{R}_\alpha(P)$, necessarily $Q_n(A_g) \to Q(A_g)$. So, the mapping $Q \mapsto Q(A_g)$ is continuous in $\mathcal{R}_\alpha(P)$ for the metric $\mathcal{W}_2$.
The map $Q \mapsto Q_X$ is continuous for $\mathcal{W}_2$ and the map $R \mapsto \mathcal{W}_2^2(P_X,R)$ too. This proves that $Q \mapsto R(Q)$ is continuous in $\mathcal{R}_\alpha(P)$. □

By Proposition 2.1 we know that the set $\mathcal{R}_\alpha(P)$ is compact. Hence, as we are looking for a probability in a compact set that minimizes a continuous and strictly convex function, we can ensure that the minimum exists and will be unique. We will denote by $\hat{P}_\alpha := \underset{Q\in\mathcal{R}_\alpha(P)}{\arg\min}\, R(Q)$, then when $\hat{P}_\alpha$ is unique it can be estimated consistently by means of $\hat{P}_{n,\alpha} := \underset{Q\in\mathcal{R}_\alpha(P_n)}{\arg\min}\, R_n(Q)$ where $\hat{P}_{n,\alpha}$ is the *empirical penalized trimming generalization error* associated with a distribution $Q \in \mathcal{R}_\alpha(P_n)$ with given α and we define it as

$$R_n(Q) := \sum_{i=1}^n w_i \delta_{\xi_i} + \lambda \mathcal{W}_2^2(P_{n,X}, Q_X),$$

where $Q_X = \sum_{i=1}^n w_i \delta_{X_i}$ with $0 \leq w_i \leq \frac{1}{n(1-\alpha)}$.

As

$$\begin{aligned} \mathcal{W}_2^2(P_{n,X},Q_X) = \min_{\pi_{i,j}} \quad & \sum_{i=1}^n \sum_{j=1}^n \pi_{ij} \parallel X_i - X_j \parallel^2 \\ \text{s.t.} \quad & \sum_{j=1}^n \pi_{ij} = w_i,\ i = 1,\dots,n \\ & \sum_{i=1}^n \pi_{ij} = \frac{1}{n},\ j = 1,\dots,n. \end{aligned}$$

Then

$$\begin{aligned} R_n(Q) = \min_{\pi_{i,j}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij}\ell(g; y_i, x_{ij}) + L\sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij} \parallel X_i - X_j \parallel^2 \\ \text{s.t.} \quad & \sum_{j=1}^{n} \pi_{ij} \leq \frac{1}{n(1-\alpha)},\ i = 1, \ldots, n \\ & \sum_{i=1}^{n} \pi_{ij} = \frac{1}{n},\ j = 1, \ldots, n, \end{aligned}$$

or, what is the same

$$\begin{aligned} R_n(Q) = \min_{\pi_{i,j}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij} c_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{n} \pi_{ij} \leq \frac{1}{n(1-\alpha)},\ i = 1, \ldots, n \\ & \sum_{i=1}^{n} \pi_{ij} = \frac{1}{n},\ j = 1, \ldots, n, \end{aligned}$$

with $c_{ij} = \ell(g; y_i, x_{ij}) + L \parallel X_i - X_j \parallel^2$ where $\ell(g; y_i, x_{ij})$ is a positive loss function. Hence, we can express $R_n(Q)$ as an optimal partial transportation problem in the same way as the ones we saw in section 3.2 and, for its resolution, we can apply the algorithm described in section 3.2.1.

Further that the computability of the criteria, the analysis of the statistical properties of this penalized method is left as future work.

5.4.3 Other loss functions

Another alternative we are considering is exploring new loss functions. Instead of 0/1 loss or SVM loss we are going to use a L_2 loss as the one used in regression problems. Our objective is to find a function that minimizes the *regression error* defined as

$$R(g) = E(\parallel Y - g(X) \parallel^2).$$

We want to study the effect that looking for a function that explains the relation between variables for a fraction of the population instead of for the whole population will have in the regression error. This is, instead of calculating the error in terms of a probability P we will do it in terms of a trimming of P.

Definition 5.4. *For a given trimming level α we define the* trimmed regression error *as*

$$R_\alpha(g) = \inf_{Q \in \mathcal{R}_\alpha(P)} E_Q(\parallel Y - g(X) \parallel^2).$$

In the same way we did for classification problem we want to find the relationship existent between $R(g)$ and $R_\alpha(g)$. In first place we will characterize $\hat{Q} \in \mathcal{R}_\alpha(P)$, which is the trimmed probability that gives us the minimal trimmed error, this is,

$$\hat{Q} = \arg\min_{Q \in \mathcal{R}_\alpha(P)} R_\alpha(P).$$

Repeating essentially the argument of Proposition 5.20 we can check that the minimizers $\hat{Q}$ are trimmings concentrated in zones of low values of the loss function, so

$$\hat{Q}(\{(y,x) :\| y - g(x) \|^2 > t\}) = \frac{(P(\{(y,x) :\| y - g(x) \|^2 > t\}) - \alpha)_+}{1-\alpha}. \tag{5.56}$$

Besides, if $F(t) = P(\{(y,x) :\| y - g(x) \|^2 \leq t\})$ and $\lambda = F^{-1}(1-\alpha)$, then

$$R_\alpha(g) = \frac{1}{1-\alpha}\left[R(g) - \lambda\alpha - \int_\lambda^\infty (1 - F(t))dt\right].$$

Note that given a trimming level $\alpha \in [0,1)$ and a function g,

$$R_\alpha(g) \leq R(g).$$

Normally we do not have a fixed regression rule, we look for the best regressor in a class $\mathcal{G}$. This is, we look for that function which gives the optimal trimmed regression error in the class,

$$\min_{(\alpha,g) \in (0,\alpha_{max}] \times \mathcal{G}} R_\alpha(g).$$

In the rest of this section we will assume that the class $\mathcal{G}$ is the class of linear functions. This is, is the class formed by functions of the form

$$g(x) = \beta^T x$$

with $\beta \in \mathbb{R}^d$. More specifically, we will discuss the computational difficulties associated with the practical implementation of a regression method with trimmings.

As in classification, it is very rare that we know the distribution of the variables, so we will have to settle for working with the empirical version of these.

We now have n observations $X_i \in \mathbb{R}^d$ and n dependent variables $Y_i \in \mathbb{R}$, the goal is to minimize in β and $\overline{W} = (w_1, \ldots, w_n)$ the function

$$G_{\alpha,n}(\beta, \overline{W}) = \sum_{i=1}^n (Y_i - \beta^T X_i)^2 w_i,$$

with $\beta \in \mathbb{R}^d$, $\overline{W} \in \mathbb{R}^n$, $0 \leq w_i \leq \frac{1}{n(1-\alpha)}$ and $\sum w_i = 1$.

A possible strategy is to fix $\overline{W}$ to obtain $H(\overline{W}) = \min_{\beta \in \mathbb{R}^d} G_{\alpha,n}(\beta, \overline{W})$. We can write $G_{\alpha,n}(\beta, \overline{W})$ in a matrix form. Take $W = \text{diag}(\overline{W})$,

$$H(\overline{W}) = \min_{\beta \in \mathbb{R}^d} (W^{1/2}Y - W^{1/2}X\beta)^T (W^{1/2}Y - W^{1/2}X\beta),$$

reaching the minimum in

$$\beta = \left((W^{1/2}X)^T (W^{1/2}X)\right)^{-1} (W^{1/2}X)^T W^{1/2}Y = \left(X^T W X\right)^{-1} X^T W^{1/2} W^{1/2} Y.$$

Hence,

$$H(\overline{W}) = Y^T W^{1/2} \left(I - W^{1/2}X \left(X^T W X\right)^{-1} X^T W^{1/2}\right) W^{1/2} Y.$$

To complete the minimization we should now minimize $H(\bar{W})$. We observe that H is a concave function (because it is the minimum of a collection of linear functions).

More precisely, the calculation of the risk of trimmed regression can be expressed as the problem of concave minimization.

$$\begin{array}{ll} \min & Y^T W^{1/2} \left(I - W^{1/2}X \left(X^T W X\right)^{-1} X^T W^{1/2}\right) W^{1/2} Y \\ \text{s.t.} & w_1 + \cdots + w_n = 1 \\ & w_1 - \dfrac{1}{n(1-\alpha)} \leq 0 \\ & \cdots \\ & w_n - \dfrac{1}{n(1-\alpha)} \leq 0 \\ & w_i \geq 0 \quad i = 1, \ldots, n. \end{array} \tag{5.57}$$

There are methods for calculating solutions to such problems. One possibility is the so-called simple external approach algorithm (see chapter 6 in Tuy (1998)). We briefly describe the application of this algorithm. We start from the polytope with which we are going to start the algorithm is $P_1 = \mathbb{R}^n_+ \bigcap \{w_1 + \ldots + w_n = 1\}$ being its vertexes set the one formed by the vectors of the canonical base in $\mathbb{R}^n$, that is, $V_1 = \{e_1, \ldots, e_n\}$. For each of these vectors, its neighbours are formed by the rest of vectors of V_1, this is, $N(e_i) = \{e_j : j \neq i, j = 1, \ldots, n\}$ with $i = 1, \ldots, n$.

In each iteration the algorithm selects which restriction is the most violated and incorporates it into the feasible polytope, eliminating the vertex that does not meet that restriction of the feasible polytope and looks for the vertices of the new polytope. We will assume by convenience in the notation that the restriction that enters into the feasible polytope in iteration k is the $k+1$ restriction of the problem (5.57). At the beginning of

iteration k the set of vertices is of the form

$$V_k = \{e_k, \ldots, e_n\} \bigcup \left\{ \frac{e_i}{n(1-\alpha)} + \left(1 - \frac{1}{n(1-\alpha)}\right) e_j : i = 1, \ldots, k-1; j = k, \ldots, n \right\}$$
$$\bigcup \left\{ \frac{e_i + e_j}{n(1-\alpha)} + \left(1 - \frac{2}{n(1-\alpha)}\right) e_l : i = 1, \ldots, k-2; j = i+1, \ldots, k-1; l = k, \ldots, n \right\}$$
$$\bigcup \cdots \bigcup \left\{ \frac{e_1 + \ldots + e_{k-1}}{n(1-\alpha)} + \left(1 - \frac{k-1}{n(1-\alpha)}\right) e_l : l = k, \ldots, n \right\}.$$

The algorithm will at least need to perform $\lceil n(1-\alpha) \rceil$ iterations to achieve the optimum that will be

$$\left\{ \frac{e_{i_1} + \ldots + e_{i_t}}{n(1-\alpha)} + \left(1 - \frac{t}{n(1-\alpha)}\right) e_{i_{t+1}} : \right.$$
$$\left. i_{t+1} \in \{1, \ldots, n\}; i_1, \ldots, i_t \in \{1, \ldots, n\} \setminus i_{t+1}; i_1 < \cdots < i_t \right\},$$

if $t \notin \mathbb{N}$ or

$$\left\{ \frac{e_{i_1} + \ldots + e_{i_t}}{n(1-\alpha)} : i_1, \ldots, i_t \in \{1, \ldots, n\}; i_1 < \cdots < i_t \right\}$$

if $t \in \mathbb{N}$ where $t = \lfloor n(1-\alpha) \rfloor$.

The resulting algorithm is computationally very expensive making the calculation unfeasible even for samples of size 10. This makes it completely useless for solving our problem. A possible alternative would be to use a concentration algorithm of the type described in section 2.5.3 and used in section 5.3.2. The adaptation of this idea, as well as the getting oracle inequalities in the style of sections 5.2.1 and 5.3.1 for the problem of partial regression is left as future work.

Chapter 6

Conclusions and future work

6.1 Conclusiones y trabajo futuro

Finalizamos este documento con un breve esbozo de las principales conclusiones de esta investigación, junto con una descripción de las futuras líneas de investigación vinculadas a las ideas y resultados presentados en este trabajo.

Nuestro principal objetivo en este proyecto fue explorar el uso de ideas y técnicas de recorte en diferentes aplicaciones estadísticas. Se han considerado dos configuraciones principales: validación de modelos y aprendizaje supervisado. El primero condujo a un enfoque nuevo y robusto a lo que llamamos *validación esencial de modelos*. El segundo llevó a un enfoque diferente de la clasificación binaria en el que buscamos un clasificador asumiendo que posiblemente estamos tratando con datos con ruido y nos conformamos con un clasificador que funciona bien para la mayoría de los datos. Durante la investigación tuvimos que tratar con problemas de transporte óptimo y esto llevó a considerar una aplicación adicional para el alineamiento en una configuración de distribuciones deformadas.

La *validación esencial de modelos* consiste en evaluar si un modelo es apropiado para una fracción dada de los datos. Esto se hace en términos de la distancia de Wasserstein entre los conjuntos de recorte. Es conocida la relación entre la distancia de Wassertein y los planes de transporte óptimos, pero el problema al que nos enfrentamos no era exactamente el problema de transporte clásico, era un problema de transporte parcial que tiene una estructura similar pero diferente. En este trabajo hemos estudiado este problema de transporte parcial con vistas a las aplicaciones estadísticas. En particular, hemos prestado atención al cálculo de las versiones empíricas de los costes parciales de transporte. Esto incluyó dos configuraciones diferentes. En primer lugar, tratamos el problema completa-

mente discreto que es más adecuado para aplicaciones con dos o más muestras de muestra. Hemos demostrado que este problema de transporte parcial discreto podría reformularse como un problema de transporte equilibrado al que se pueden aplicar algoritmos eficientes de programación lineal. En el caso del transporte parcial entre una medida empírica y un modelo continuo (el caso que surge naturalmente en la validación esencial de modelos) tuvimos que tratar un problema semidescreto. En este caso demostramos que el problema podría ser reformulado en términos de minimización convexa. Propusimos un algoritmo de gradiente estocástico convergente e ilustramos su funcionamiento a través de algunos experimentos numéricos. El algoritmo estocástico podría mejorarse con una estrategia cuasi Monte-Carlo para una evaluación más eficiente del gradiente. Sin embargo, este enfoque tenía algunas limitaciones. Las búsqueda de nuevas estrategias conducentes a soluciones más eficientes del problema de minimización se dejan para investigación futura. Estos resultados se han aplicado al problema de validación esencial de modelos con un enfoque basado en el paradigma de selección de modelos. Hemos proporcionado garantías de su desempeño con una desigualdad oráculo. Consideramos como trabajo futuro también el estudio de nuevas formas para la función de penalización.

Mientras estudiábamos estos problemas de transporte, resultó que los métodos numéricos que estábamos desarrollando para resolver el problema del transporte parcial eran una herramienta útil para resolver los problemas de alinemaiento de distribuciones. Hemos desarrollado un criterio basado en la minimización de la distancia de Wasserstein para estimar los parámetros de deformación con el fin de recuperar la forma original de las distribuciones deformadas.

En la teoría de clasificación, muchas reglas de clasificación se ven afectadas por la presencia de puntos atípicos que son muy difíciles de clasificar. Cuando se trata de observaciones de alta dimensión o cuando el número de observaciones es grande, esta situación ocurre con bastante frecuencia y puede obstaculizar drásticamente el rendimiento de los clasificadores que tienen en cuenta todos los datos. Uno puede estar tentado a centrarse en estos puntos y modificar la regla de clasificación para aumentar su clasificación para estos puntos especiales. Este es el punto de vista de los algoritmos de potenciación, por ejemplo, tal como se describe en Freund and Schapire (1995) por ejemplo. Sin embargo, esto se hace a menudo a expensas de la complejidad de la regla y de su capacidad de generalizarse. Por lo tanto, una solución práctica y quizás pragmática es considerar algunos de estos puntos como valores atípicos y simplemente eliminarlos. Los estadísticos son reacios a descartar las observaciones, pero en muchas aplicaciones, en particular cuando se enfrentan a grandes cantidades de observaciones, esto permite elaborar reglas más fáciles

de interpretar y que pueden proporcionar una mejor comprensión del fenómeno estudiado, siempre que no se retiren demasiados datos de la muestra de formación. Esta es la elección típica que se hace en varios documentos, pero no se dice mucho sobre la forma en que se seleccionan los valores anómalos y su impacto en el rendimiento de la clasificación.

Esta es la razón por la que hemos proporcionado en este documento un marco estadístico para una clasificación robusta mediante la eliminación de algunas observaciones. Hemos proporcionado un método que para considerar los datos como valores anómalos se basa en su error de clasificación por un clasificador o una clase dada de clasificadores. En este marco, este procedimiento permite seleccionar de forma dirigida por los datos una proporción óptima de observaciones que deben eliminarse para lograr un mejor error de clasificación. El nivel de recorte y el mejor clasificador se seleccionan simultáneamente y hemos obtenido una desigualdad oráculo para evaluar la calidad de este procedimiento. Creemos que este resultado puede proporcionar algunas pautas para eliminar los valores atípicos para los problemas de clasificación con garantías teóricas.

Primero nos basamos en la minimización de una función de pérdida 0/1 penalizada que es computacionalmente difícil de manejar. A continuación se estudia una versión de este procedimiento de recorte para funciones convexas que conduce a una forma factible de calcular los pesos. Hemos obtenido una forma de detectar valores anómalos y eliminarlos de tal manera que el error de clasificación con este nuevo conjunto de datos se controlará teóricamente. También hemos proporcionado un algoritmo que obtiene la regla óptima y elimina los valores atípicos en un tiempo factible. Como trabajo futuro en esta dirección queremos mejorar la cota dada en el Lema 5.27 que puede aportar mejoras en el rendimiento del algoritmo antes mencionado. Por último, en la sección 5.4, hemos introducido dos problemas que nos gustaría continuar en el futuro. El primero es la selección de modelos para la clasificación basada en el modelo recortado y no en el nivel de recorte, es decir, considerando una función de penalización que depende de la desviación del modelo recortado respecto al modelo original y no en el nivel de recorte. Aunque hemos encontrado que penalizar con la divergencia de Kullback conduce a malos resultados, creemos que usar la distancia de Wasserstein entre los modelos como penalización puede dar buenos resultados usando algoritmos de transporte óptimos para resolver el problema. Por otro lado, hemos tocado el problema de la regresión parcial, pero sólo hemos dado los primeros pasos en el proceso. Como trabajo futuro nos gustaría extender los resultados obtenidos para el problema de la clasificación parcial a esta configuración de regresión obteniendo un método para seleccionar los regresores óptimos para una fracción de los datos, así como los cotas oráculo y un algoritmo eficiente.

6.2 Conclusions and future work

We end this document with a brief outline of the main conclusions of this research, together with a description of future lines of investigation linked to the ideas and results presented in this work.

Our main goal in this project was to explore the use of trimming ideas and techniques in different statistical applications. Two main setups have been considered: model validation and supervised learning. The first one led to a new, robust approach to what we call *essential model validation*. The second led to a different approach to binary classification in which we look for a classifier assuming that we are possibly dealing with noisy data and settle for a classifier that works well for most of the data. During the research we had to deal with optimal transportation problems and this led to consider a further application to registration in a setup of warped distributions.

Essential model validation consists in assessing whether a model is appropriate for a given fraction of the data. This is done in terms of Wasserstein distance between trimming sets. It is well known the relationship between Wasserstein distance and optimal transportation plans, but the problem we were facing was not exactly the classical transportation problem, it was a partial transportation problem which has a similar but different structure. In this work we have studied this partial transportation problem with a view towards statistical applications. In particular we have paid attention to the computation of empirical versions of the partial transportation cost. This included two different setups. First we dealt with the completely discrete problem that is more suitable for two or k-sample applications. We have shown that this discrete partial transportation problem could be reformulated as a balanced transportation problem to which efficient linear programming algorithms can be applied. We turned this algorithms into efficient C-code callable from R. In the case of partial transportation between an empirical measure and a continuous model (the case which arises naturally in essential model validation) we had to deal with a semidiscrete problem. In this case we showed that the problem could be recast in terms of convex minimization. We proposed a convergent stochastic gradient algorithm and illustrated its performance through some numerical experiments. The stochastic algorithm could be enhanced with a quasi Monte-Carlo strategy for a more efficient evaluation of the gradient. However, this approach had some limitations. New strategies leading to more efficient solutions of the minimization problem are left for future research. This results have been applied to the essential model validation problem with an approach based on the model selection paradigm. We have provided guarantees of its performance with an oracle inequality. We consider as future work as well the study of new forms for the

penalization function.

While we were studying these transportation problems it turned out that the numerical methods we were developing for solving the partial transportation problem were a useful tool to solve distribution registration problems. We have developed a criterion based on Wasserstein distance minimization to estimate deformation parameters in order to recover the original form of warped distributions.

In classification theory, many classification rules are affected by the presence of atypical points which are very difficult to classify. When dealing with high dimensional observations or when the number of observations is large, this situation occurs quite often and may drastically hamper the performance of classifiers which take into account all the data. One may be tempted to focus on these points and modify the classification rule to increase their classification ranking for these special points. This is the point of view of boosting algorithms for instance, as described in Freund and Schapire (1995) for example. Yet this is often done at the expense of the complexity of the rule and its ability to be generalized. Hence a practical and maybe pragmatic solution is to consider some of these points as outliers and to simply remove them. Statisticians are reluctant to discard observations, yet in many applications, in particular when confronted to large amounts of observations, this enables to produce rules that are easier to interpret and that can provide a better understanding of the phenomenon which is studied, provided not too many data are removed from the training sample. This is the typical choice made in several papers but not much is said about the way the outliers are selected and its impact on the classification performance.

This is the reason why we have provided in this work a statistical framework to robust classification by removal of some observations. We have provided a method that considers data as outliers based on their classification error by a classifier or a given class of classifiers. Within this framework this procedure enables to select in a data driven way an optimal proportion of observations to be removed in order to achieve a better classification error. The level of trimming and the best classifier are selected simultaneously and we have obtained an oracle inequality to assess the quality of this procedure. We think that this result may provide some guidelines to remove outliers for classification problems with theoretical guarantees.

First we relied on a minimization of a penalized 0/1 loss function which is computationally difficult to handle. A version of this trimming procedure for convex functions that lead to a feasible way of computing the weights is studied later. We have obtained a way of detecting outliers and removing them such that the classification error with this new

data set will be theoretically controlled. We also have provided an algorithm that gets the optimal rule and removes outliers in a feasible time. As future work in this direction we want to improve the bound given in Lemma 5.27 that can bring improvements in the performance of the algorithm just mentioned. Finally, in section 5.4, we have introduced two problems that we would like to continue in the future. The first one is model selection for classification based on the trimmed model and not on the trimming level, that is, considering a penalization function that depends on the deviation of the trimmed model from the original one and not in the trimming level. Although we have found that penalizing with Kullback's divergence leads to bad results, we believe that using Wasserstein distance between the models as penalization may give good results using optimal transportation algorithms to solve the problem. On the other hand we have touched upon the partial regression problem but we have only taken the first steps in the process. As future work we would like to extend the results obtained for the partial classification problem to this regression setup obtaining a method to select optimal regressors for a fraction of the data as well as oracle bounds and an efficient algorithm.

Bibliography

[1] Agulló-Antolín, M., Cuesta-Albertos, J. A., Lescornel, H., & Loubes, J. (2015). A parametric registration model for warped distributions with Wasserstein's distance. *J. Multivariate Anal.*, 135:117–130.

[2] Agulló-Antolín, M., del Barrio, E., & Loubes, J.-M. (2017). A data driven trimming procedure for robust classification. *arXiv preprint arXiv:1701.05065*.

[3] Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Stat.*, 7(1):226–248.

[4] Álvarez-Esteban, P. C. (2009). Aplicaciones de los recortes imparciales en la comparación de distribuciones.

[5] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2008). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, 103(482):697–704.

[6] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2011). Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):358–375.

[7] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2012). Similarity of samples and trimming. *Bernoulli*, 18(2):606–634.

[8] Amit, Y., Grenander, U., & Piccioni, M. (1991). Structural Image Restoration through deformable template. *Journal of the American Statistical Association*, 86:376–387.

[9] Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley Publications in Statistics. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London.

[10] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances.* Princeton University Press, Princeton, N.J.

[11] Bartlett, P. L. & Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840.

[12] Bazaraa, M. S., Jarvis, J. J., & Sherali, H. D. (2010). *Linear programming and network flows*, (fourth ed.). John Wiley & Sons, Inc., Hoboken, NJ.

[13] Bercu, B. & Fraysse, P. (2012). A Robbins-Monro procedure for estimation in semiparametric regression models. *Annals of Statistics*, 40(2):666–693.

[14] Berkelaar, M. (2013). *lpSolve: Interface to Lpsolve v. 5.5 to solve linear/integer programs.* R package version 5.6.7.

[15] Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536.

[16] Bickel, P. J. & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6):1196–1217.

[17] Bickel, P. J. & Lehmann, E. L. (1975). Descriptive statistics for nonparametric models. II. Location. *Ann. Statist.*, 3(5):1045–1069.

[18] Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

[19] Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375.

[20] Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357.

[21] Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

[22] Chen, X., Wang, Z. J., & McKeown, M. J. (2010). Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Trans. Inform. Theory*, 56(10):5131–5149.

[23] Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press.

[24] Cuesta, J. & Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields*, 78(4):523–534.

[25] Cuesta-Albertos, J. A. & Fraiman, R. (2006). Impartial trimmed means for functional data. In: *Data depth: robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 121–145. Amer. Math. Soc., Providence, RI.

[26] Cuesta-Albertos, J. A., García-Escudero, L. A., & Gordaliza, A. (2002). On the asymptotics of trimmed best k-nets. *J. Multivariate Anal.*, 82(2):486–516.

[27] Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553–576.

[28] Dantzig, G. B. (1951). Application of the simplex method to a transportation problem. In: *Activity Analysis of Production and Allocation*, Cowles Commission Monograph No. 13, pages 359–373. John Wiley & Sons, Inc., New York, N. Y.; Chapman & Hall, Ltd., London.

[29] Dantzig, G. B. (1963). *Linear programming and extensions.* Princeton University Press, Princeton, N.J.

[30] Debruyne, M. (2009). An outlier map for support vector machine classification. *Ann. Appl. Stat.*, 3(4):1566–1580.

[31] del Barrio, E., Deheuvels, P., & van de Geer, S. (2007). *Lectures on empirical processes.* EMS Series of Lectures in Mathematics. European Mathematical Society (EMS), Zürich. Theory and statistical applications, With a preface by Juan A. Cuesta Albertos and Carlos Matrán.

[32] del Barrio, E. & Loubes, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299v2.*

[33] del Barrio, E. & Matrán, C. (2013). Rates of convergence for partial mass problemas. *Probability Theory and related fields*, 155(3-4):521–542.

[34] Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

[35] Devroye, L. & Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York.

[36] Donoho, D. & Huber, P. J. (1983). The notion of breakdown point. In: *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 157–184. Wadsworth, Belmont, CA.

[37] Dudley, R. M. (2002). *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press. Revised reprint of the 1989 original.

[38] Dupuy, J., Loubes, J., & Maza, E. (2011). Non parametric estimation of the structural expectation of a stochastic increasing function. *Statistics and Computing*, pages 1–16.

[39] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499. With discussion, and a rejoinder by the authors.

[40] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[41] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Human Genetics*, 8(4):376–386.

[42] Fournier, N. & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738.

[43] Fraiman, R. & Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

[44] Freund, Y. & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In: *European conference on computational learning theory*, pages 23–37. Springer.

[45] Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

[46] Gallon, S., Loubes, J., & Maza, E. (2013). Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142.

[47] Gamboa, F., Loubes, J., & Maza, E. (2007). Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1:616–640.

[48] García-Escudero, L. A., Gordaliza, A., & Matrán, C. (1999). Asymptotics for trimmed k-means and associated tolerance zones. *J. Statist. Plann. Inference*, 77(2):247–262.

[49] García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *J. Comput. Graph. Statist.*, 12(2):434–449.

[50] García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36(3):1324–1345.

[51] Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, 64(2):162–180.

[52] Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, 42:1887–1896.

[53] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. The approach based on influence functions.

[54] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*, (second ed.). Springer Series in Statistics. Springer, New York. Data mining, inference, and prediction.

[55] Herbei, R. & Wegkamp, M. H. (2006). Classification with reject option. *Canad. J. Statist.*, 34(4):709–721.

[56] Hitchcock, F. L. (1941). The distribution of a product from several sources to numerous localities. *J. Math. Phys. Mass. Inst. Tech.*, 20:224–230.

[57] Hodges, Jr., J. L. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I: Statistics, pp. 163–186. Univ. California Press, Berkeley, Calif.

[58] Hodges, Jr., J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B.*, 16:261–268.

[59] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.

[60] Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

[61] Huber, P. J. (1996). *Robust statistical procedures*, (second ed.), volume 68 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[62] Kantorovich, L. V. (1958). On the translocation of masses. *Management Science*, 5(1):1–4.

[63] Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422.

[64] Koopmans, T. C. (1949). Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, pages 136–146.

[65] Korukoğlu, S. & Ballı, S. (2011). A improved vogel's approximatio method for the transportation problem. *Mathematical and Computational Applications*, 16(1):370–381.

[66] Lin, Y. (2002). Support vector machines and the Bayes rule in classification. *Data Min. Knowl. Discov.*, 6(3):259–275.

[67] Lindsay, B. & Liu, J. (2009). Model assessment tools for a model false world. *Statist. Sci.*, 24(3):303–318.

[68] Lugosi, G. (2002). Pattern classification and learning theory. In: *Principles of nonparametric learning*, pages 1–56. Springer.

[69] Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53. Supplementary materials available online.

[70] Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and methods.

[71] Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[72] Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale.

[73] Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609.

[74] R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. ISBN 3-900051-07-0.

[75] Rachev, S. T. & Rüschendorf, L. (1998). *Mass transportation problems. Vol. II.* Probability and its Applications (New York). Springer-Verlag, New York. Applications.

[76] Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, (2nd ed.). Springer.

[77] Rao, C. (1952). Advanced statistical methods in multivariate analysis.

[78] Reinfeld, N. V. & Vogel, W. R. (1960). *Matematicheskoe programmirovanie.* Translated from the English by G. N. Andrianov and B. N. Mihalevskiĭ; edited by A. A. Konyus. Izdat. Inostr. Lit., Moscow.

[79] Rockafellar, R. T. (1997). *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ. Reprint of the 1970 original, Princeton Paperbacks.

[80] Rousseeuw, P. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, 79(388):871–880.

[81] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In: *Mathematical statistics and applications, Vol. B (Bad Tatzmannsdorf, 1983)*, pages 283–297. Reidel, Dordrecht.

[82] Rousseeuw, P. (1997). Introduction to positive-breakdown methods. In: *Robust inference*, volume 15 of *Handbook of Statist.*, pages 101–121. North-Holland, Amsterdam.

[83] Rousseeuw, P. & Driessen, K. V. (2006). Computing LTS regression for large data sets. *Data Min. Knowl. Discov.*, 12(1):29–45.

[84] Rousseeuw, P. & Hubert, M. (2013). High-breakdown estimators of multivariate location and scatter. In: *Robustness and complex data structures*, pages 49–66. Springer, Heidelberg.

[85] Rudas, T., Clogg, C. C., & Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc. Ser. B*, 56(4):623–639.

[86] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

[87] Trouve, A. & Younes, L. (2005). Metamorphoses Through Lie Group Action. *Foundations of Computational Mathematics*, 5(2):173–198.

[88] Tuy, H. (1998). *Convex analysis and global optimization*, volume 22 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht.

[89] van der Vaart, A. W. & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[90] Vapnik, V. (1982). *Estimation of dependences based on empirical data*. Springer Series in Statistics. Springer-Verlag, New York-Berlin. Translated from the Russian by Samuel Kotz.

[91] Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Information Science and Statistics. Springer, New York. Reprint of the 1982 edition, Afterword of 2006: Empirical inference science.

[92] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

[93] Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Society.

[94] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[95] Vimond, M. (2010). Efficient estimation for a subclass of shape invariant models. *Ann. Statist.*, 38(3):1885–1912.

[96] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.*, 15(2):642–656.