UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Estimadores robustos para el modelo de regresión lineal con datos de alta dimensión

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Matemáticas

Ezequiel Smucler

Director de tesis: Dr. Víctor J. Yohai
Consejero de estudios: Dr. Víctor J. Yohai

Lugar de trabajo: Instituto de Cálculo

Buenos Aires, 2016
Fecha de defensa: 13 de octubre de 2016

Estimadores robustos para el modelo de regresión lineal con datos de alta dimensión

Resumen

Los estimadores de regresión penalizados son una herramienta popular para analizar conjuntos de datos ralos y de alta dimensión. Sin embargo, los estimadores de regresión penalizados definidos utilizando funciones de pérdida no acotadas, como la pérdida cuadrática, pueden verse muy afectados por la presencia de observaciones atípicas en la muestra, especialmente aquellas de alto leverage, y por lo tanto no son robustos.

Esta tesis consiste de dos partes. En la primera, proponemos una familia de estimadores penalizados para la estimación robusta en modelos lineales ralos y de alta dimensión basados en los MM-estimadores de Yohai (1987). Estudiamos las propiedades asintóticas de estos estimadores en modelos lineales con una cantidad fija de variables predictoras aleatorias. Proponemos un algoritmo para computar una subfamilia de los estimadores propuestos. Las ventajas relativas que ofrecen los estimadores propuestos son demostradas mediante un extenso estudio de simulación y el análisis de un conjunto de datos reales. Esta primer parte está basada en Smucler and Yohai (2015 b).

En la segunda parte, estudiamos las propiedades asintóticas de los estimadores propuestos en modelos lineales con un número de variables predictoras fijas que diverge, dentro del régimen $p \ll n$. Probamos la consistencia de los estimadores asumiendo solo $p/n \to 0$, y que si la función de penalización es elegida convenientemente entonces los estimadores resultantes tienen la propiedad oráculo definida en Fan and Li (2001). La misma técnica de demostración nos permite probar la consistencia y derivar la distribución asintótica de M-estimadores de regresión definidos utilizando una función de perdida acotada y un estimador de escala, en modelos lineales con un número de variables predictoras fijas que diverge. En particular, probamos la consistencia y derivamos la distribución asintótica de los S-estimadores (Rousseeuw and Yohai (1984)) y MM-estimadores de regresión.

Palabras clave: Regresión robusta, M-estimadores, S-estimadores, MM-estimadores, Estimados de Regresión Penalizados, Lasso, Modelos Ralos, Propiedad Oráculo, Modelos Estadísticos con un Número de Parámetros que Diverge.

Robust estimators for high-dimensional linear regression models

Abstract

Penalized regression estimators are a popular tool for the analysis of sparse and high-dimensional data sets. However, penalized regression estimators defined using unbounded loss functions, such as the quadratic loss, can be very sensitive to the presence of outlying observations, especially high leverage outliers, and hence are not robust.

This thesis consists of two parts. In the first one, we propose a family of penalized estimators for robust estimation in sparse and high-dimensional linear models based on the MM-estimators of Yohai (1987). We study the asymptotic properties of these estimators in linear models with a fixed number of random predictor variables. We propose an algorithm to compute a subset of this family. The relative advantages of these estimators are demonstrated through an extensive simulation study and the analysis of a real high-dimensional data set. This first part is based on Smucler and Yohai (2015 b).

In the second part, we study the asymptotic properties of the proposed estimators in linear models with a diverging number of fixed predictor variables in the $p \ll n$ regime. We prove the consistency of the estimators assuming only $p/n \to 0$ and that if the penalty function is chosen appropriately then the resulting estimators have the oracle property of Fan and Li (2001). The same proof technique allows us to prove the consistency and derive the asymptotic distribution of regression M-estimators defined using a bounded loss function and an estimate of scale, in linear models with a diverging number of fixed predictor variables. In particular, we prove the consistency and derive the asymptotic distribution of S-estimators (Rousseeuw and Yohai (1984)) and MM-estimators.

Keywords: Robust Regression, M-estimators, S-estimators, MM-estimators, Penalized Regression Estimators, Lasso, Sparsity, Oracle Property, Dimension Asymptotics.

Contents

Introducción

Regresión lineal

Una problema común a todas las áreas de la ciencia y de la tecnología es el de explicar la relación entre una variable respuesta y y un vector de variables predictivas $\mathbf{x}$. La manera más simple de modelar esta relación es mediante un modelo lineal. Más precisamente, supongamos que observamos $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ $i = 1, \ldots, n$, vectores de dimensión $(p+1)$, donde y_i es una variable respuesta y $\mathbf{x}_i \in \mathbb{R}^p$ es un vector de variables predictivas. Las variables predictivas pueden ser fijas, como es el caso en un experimento diseñado, o aleatorias, si por ejemplo corresponden a observaciones de campo. Asumimos que $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, satisfacen

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + u_i,$$

donde $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ es un vector de coeficientes de regresión a estimar y los u_i $i = 1, \ldots, n$ son errores aleatorios i.i.d. definidos en un mismo espacio de probabilidad. Para el caso de variables predictivas aleatorias, asumiremos que los vectores de variables predictivas son i.i.d. e independientes de los errores.

El estimador de mínimos cuadrados (EMC) de $\boldsymbol{\beta}_0$ está definido por

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}),$$

donde $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ es el residuo de la i-ésima observación. Es bien sabido que, bajo condiciones de regularidad, el EMC satisface

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N_p\left(\mathbf{0}, \sigma^2 \mathbf{C}^{-1}\right),$$

donde σ^2 es la varianza de los errores, $\mathbf{C} = \mathbb{E}\mathbf{x}\mathbf{x}^T$ para el caso de predictores aleatorios y $\mathbf{C} = \lim 1/n \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ para el caso de predictores fijos. Cuando los errores tienen distribución normal, el EMC de $\boldsymbol{\beta}_0$ coincide con el estimador de máxima verosimilitud. Sin embargo, es bien sabido que el EMC no es robusto, es decir, es muy sensible a desviaciones de los supuestos del modelo.

Estimadores de regresión robustos

El objetivo principal de la estadística robusta es proveer métodos para el análisis de datos que sean confiables aún en la presencia de datos atípicos (outliers). Un objetivo complementario es obtener estimadores robustos que sean casi tan buenos como el estimador clásico óptimo (digamos el de máxima verosimilitud) cuando la distribución de los datos coincide con una nominal, típicamente la distribución normal.

La robustez de un estimador se mide por su estabilidad cuando una fracción pequeña de las observaciones es reemplazada de forma arbitraria por datos atípicos que pueden no cumplir con el modelo asumido. Un estimador robusto no debería verse mayormente afectado por una fracción pequeña de datos atípicos. Una medida cuantitativa de la robustez de un estimador, introducida en Donoho and Huber (1983), es el punto de ruptura finito. Informalmente, el punto de ruptura finito de un estimador es la máxima fracción de datos atípicos que el estimador puede tolerar sin verse completamente arruinado. Para un estimador de regresión, esta medida se define de la siguiente manera. Dada una muestra $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, sea $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ y sea $\hat{\boldsymbol{\beta}}(\mathbf{Z})$ el estimador de regresión $\hat{\boldsymbol{\beta}}$ calculado en $\mathbf{Z}$. El punto de ruptura finito de $\hat{\boldsymbol{\beta}}$ está definido por

$$FBP(\hat{\boldsymbol{\beta}}) = \frac{m^*}{n},$$

donde

$$m^* = \max\left\{m \geq 0 : \hat{\boldsymbol{\beta}}(\mathbf{Z}_m) \text{ está acotado para todo } \mathbf{Z}_m \in \mathcal{Z}_m\right\},$$

y $\mathcal{Z}_m$ es el conjunto de todos los conjuntos de datos con al menos $n - m$ elementos en común con $\mathbf{Z}$. Sea $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})$ el estimador $\hat{\boldsymbol{\beta}}$ calculado en $(\mathbf{X}, \mathbf{y})$, donde $\mathbf{X}$ es la matriz con $\mathbf{x}_i$ como filas e $\mathbf{y} = (y_1, \ldots, y_n)$. El estimador $\hat{\boldsymbol{\beta}}$ se dice equivariante por transformaciones de regresión si para todo $\mathbf{b} \in \mathbb{R}^p$, $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\mathbf{b}) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}) + \mathbf{b}$. Se puede probar que el punto de ruptura de cualquier estimador de regresión equivariante por transformaciones de regresión es a lo sumo $[(n-p)/2]/n$, que es aproximadamente $1/2$ para $p \ll n$. Ver, por ejemplo, la Sección 5.4.1 de Maronna et al. (2006).

Un punto de ruptura igual a ε^* garantiza que para cualquier nivel de contaminación $\varepsilon \leq \varepsilon^*$ existe un compacto, digamos $K = K(\varepsilon)$, tal que el estimador en cuestión permanece en K cuando una fracción ε de las observaciones es modificada arbitrariamente. Sin embargo, este compacto podría ser muy grande. Luego, aún cuando un punto de ruptura alto es siempre una propiedad deseable, un estimador con punto de ruptura alto puede verse altamente afectado por una fracción pequeña de observaciones atípicas.

Es bien sabido que el EMC tiene punto de ruptura cero, esto es, una sola observación atípica puede arruinarlo completamente.

Un marco general para la estimación robusta en modelos lineales esta dado por la teoría de M-estimación. La noción de un M-estimador fue introducida por Huber (1964) para el caso de estimación de un parámetro de posición y extendida al modelo lineal en Huber (1973). Para

definir a los M-estimadores, primero introducimos un poco de notación. A lo largo de esta tesis, diremos que ρ es una ρ-función si

1. ρ es par y continua.
2. $\rho(x)$ es una función no decreciente de $|x|$.
3. $\rho(0) = 0$.
4. Si $\rho(v) < \lim_{x\to\infty} \rho(x)$ y $0 \leq u < v$ entonces $\rho(u) < \rho(v)$.
5. Si ρ es acotada, $\lim_{x\to\infty} \rho(x) = 1$.

Si ρ es una ρ-función derivable, llamaremos $\psi = \rho'$.

Una familia de ρ-funciones comunmente utilizada en la estadística robusta es la familia de funciones de pérdida de Huber, dada por

$$\rho_c^H(x) = x^2 I\{|x| \leq c\} + \left(2|x|c - c^2\right) I\{|x| \geq c\},$$

donde $c > 0$ es una constante de calibración. Las funciones de pérdida de Huber son un compromiso entre la pérdida cuadrática, que define al EMC, y la pérdida del valor absoluto, que define al estimador de mínimas desviaciones absolutas (EMDA). Otra familia de ρ-funciones comunmente utilizada es de funciones Bicuadráticas de Tukey, dada por

$$\rho_c^B(x) = 1 - \left(1 - \left(\frac{x}{c}\right)^2\right)^3 I\{|x| \leq c\},$$

donde $c > 0$ es una constante de calibración. Notar que ρ_c^B es acotada, más aún, si $|x| \geq c$, entonces $\rho_c^B(x) = 1$.

En la Figura 1 mostramos un gráfico de varias funciones de pérdida.

Para poder definir a los M-estimadores de regresión, primero necesitamos definir a los M-estimadores de escala. Sea ρ_0 una ρ-función acotada. Dada una muestra $\mathbf{u} = (u_1, ..., u_n)$ y $0 < b < 1$ el correspondiente M-estimador de escala $s_n(\mathbf{u})$ está definido, Huber (1981), por

$$s_n(\mathbf{u}) = \inf\left\{s > 0 : \frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{u_i}{s}\right) \leq b\right\}.$$

Es fácil ver que $s_n(\mathbf{u}) > 0$ si y solo si $\#\{i : u_i = 0\} < (1-b)n$, y en este caso

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{u_i}{s_n(\mathbf{u})}\right) = b.$$

Notar que los M-estimadores de escala son equivariantes por transformaciones de escala, es decir, para todo c, $s_n(c\mathbf{u}) = |c|s_n(\mathbf{u})$.

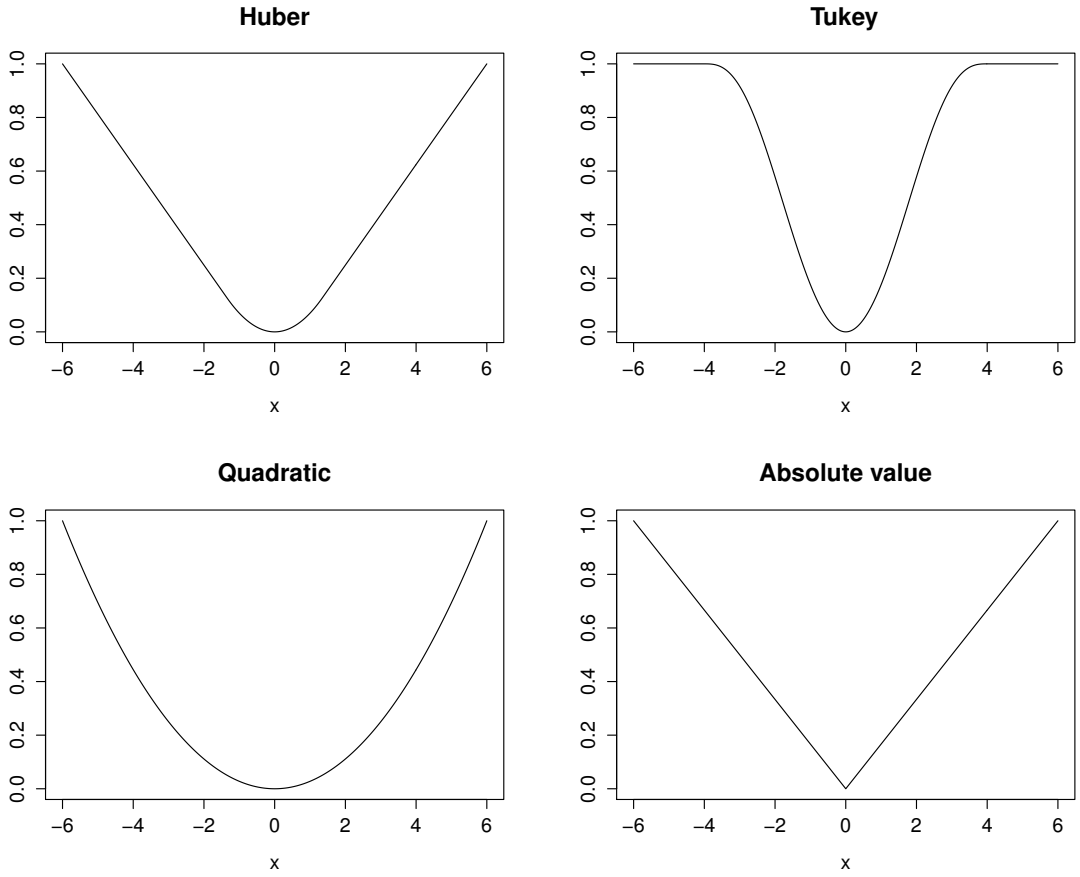

Figure 1: Gráficos de los funciones de pérdida de Huber, de Tukey, de la pérdidad cuadrática y de la pérdida del valor absoluto. Todas las funciones fueron reescaladas para que su máximo en el intervalo $[-6, 6]$ sea igual a 1.

Dada una ρ-función ρ, el correspondiente M-estimador de regresión está definido por

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right). \tag{1}$$

donde s_n es un estimador de escala de los residuos que puede haber sido estimado a priori o en simultáneo. Por ejemplo, s_n podría ser el M-estimador de escala de los residuos o la mediana del valor absoluto de los residuos de algún estimador de regresión inicial. No usar un estimador de escala en (1) es lo mismo que, haciendo un abuso de notación, tomar s_n como una constante igual a 1. Si s_n es equivariante por transformaciones de escala, dividir los residuos por s_n en (1) hace que el correspondiente M-estimador de regresión sea equivariante

por transformaciones de escala, esto es: para todo c, $\hat{\boldsymbol{\beta}}(\mathbf{X}, c\mathbf{y}) = c\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})$. Los M-estimadores de regresión son equivariantes por transformaciones de regresión cuando, por ejemplo, no se usa un estimador de escala para definirlos o cuando s_n es el M-estimador de escala de los residuos de un estimador de regresión equivariante por transformaciones de regresión. Estas propiedades de equivariancia son deseables, ya que nos permiten saber como cambian los estimadores cuando los datos sufren las correspondientes transformaciones.

El EMC se obtiene tomando $\rho(x) = x^2$ en (1), mientras que el EMDA se obtiene tomando $\rho(x) = |x|$. Notar que para estos dos casos, no es necesario utilizar un estimador de escala para que los correspondientes estimadores de regresión sean equivariantes por transformaciones de escala. Para obtener estimadores robustos, generalmente se utiliza una ρ-función en (1) que crezca menos rápido que la función cuadrática.

Para el caso de una función de pérdida convexa y derivable, por ejemplo la de Huber, (1) es esencialmente equivalente a

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{s_n}\right)\mathbf{x}_i = \mathbf{0}, \tag{2}$$

donde $\psi = \rho'$; ver la Sección 7.3 de Huber (1981) y la Sección 4.4 de Maronna et al. (2006). En este caso, el M-estimador resultante es llamado un M-estimador monótono. Cuando ψ tiende a cero en infinito, el estimador resultante es llamado un M-estimador redescendiente y en este caso algunas soluciones de (2) pueden no corresponder a soluciones de (1)

En Huber (1964), Huber muestra que los M-estimadores de posición definidos utilizando la función de pérdida de Huber tiene una propiedad de optimalidad minimax: si la constante de calibración es escogida convenientemente, los estimadores resultantes minimizan la máxima varianza asintótica sobre entornos de contaminación de la distribución normal. Huber (1973) estudió las propiedades asintóticas de M-estimadores monótonos definidos por (1) para el caso de variables predictoras fijas, pero sin incluir un estimador de escala. En particular, probó la consistencia y normalidad asintótica de estos estimadores. Se puede probar que si solo consideramos la posibilidad de observaciones atípicas en la variable respuesta, los M-estimadores monótonos pueden tener un punto de ruptura alto. Sin embargo, si consideramos la posibilidad de observaciones atípicas en las variables predictivas, el punto de ruptura de los M-estimadores monótonos es cero; ver la Sección 5.16.1 de Maronna et al. (2006).

El estimador de la mínima mediana de los cuadrados (EMMC) es un estimador equivariante por transformaciones de regresión que puede alcanzar la cota óptima de 1/2 para su punto de ruptura asintótico. El EMMC fue propuesto en primera instancia por Hampel (1975) y desarrollado posteriormente por Rousseeuw (1984). Está definido por

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \text{mediana } (r_i(\boldsymbol{\beta})^2)_{i=1}^n.$$

El EMMC converge a tasa $n^{1/3}$ y por lo tanto su eficiencia asintótica para el caso de errores normales es cero. Ver Davies (1990) y Kim and Pollard (1990).

Rousseeuw and Yohai (1984) introdujeron los S-estimadores. Los S-estimadores combinan la tasa usual de convergencia de $\sqrt{n}$ con un punto de ruptura alto. Están definidos por

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} s_n(\mathbf{r}(\boldsymbol{\beta}))$$

donde $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$ y $s_n()$ es un M-estimador de escala. Es fácil verificar que los S-estimadores son equivariantes por transformaciones de regresión y de escala. Sea $\hat{s}_n = s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$ y sea ρ_0 la ρ-función usada para definir $s_n()$. Los S-estimadores satisfacen

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_0\left(\frac{r_i(\boldsymbol{\beta})}{\hat{s}_n}\right),$$

ver la Sección 5.6.1 de Maronna et al. (2006). Luego, los S-estimadores son M-estimadores, (1), donde la función de pérdida ρ es acotada y la escala es estimada simultáneamente. En la práctica, ρ_0 suele elegirse de manera que $\rho_0(x) = 1$ si $|x| \geq m$ para cierto m. Por ejemplo, ρ_0 podría ser la función de pérdida de Tukey. La distribución de los S-estimadores de regresión fue derivada, bajo condiciones muy generales, por Fasano et al. (2012) para el caso de variables predictoras aleatorias y por Davies (1990) para el caso de variables predictoras fijas. Los S-estimadores siempre pueden calibrarse para que tengan el mayor punto de ruptura finito posible para estimadores equivariantes por transformaciones de regresión, ver la Sección 5.6.1 de Maronna et al. (2006). Sin embargo, los S-estimadores no pueden combinar un punto de ruptura alto con una eficiencia asintótica alta en la distribución normal. Ver Hössjer (1992).

Los MM-estimadores, introducidos en Yohai (1987), son estimadores de regresión que pueden calibrarse para alcanzar simultáneamente un alto punto de ruptura y una alta eficiencia asintótica en la distribución normal. Supongamos que $\hat{\boldsymbol{\beta}}_1$ es una estimador robusto, pero no necesariamente altamente eficiente. En la práctica, $\hat{\boldsymbol{\beta}}_1$ suele ser un S-estimador. Sea $s_n()$ un M-estimador de escala definido usando una ρ-función ρ_0 y b. Sea ρ_1 otra ρ-función que satisface $\rho_1 \leq \rho_0$. El MM-estimador de regresión está definido por

$$\hat{\boldsymbol{\beta}}_{MM} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right).$$

Notar que los MM-estimadores son M-estimadores, (1), definidos usando una función de pérdidad acotada y un estimador de escala preliminar. Los MM-estimadores son equivariantes por transformaciones de escala y de regresión, siempre que $\hat{\boldsymbol{\beta}}_1$ satisfaga estas propiedades. Yohai (1987) prueba que bajo condiciones de regularidad y para el caso de variables predictivas aleatorias, los MM-estimadores satisfacen

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MM} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p\left(\mathbf{0}, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2}\mathbf{V}_{\mathbf{x}}^{-1}\right),$$

donde $\psi_1 = \rho_1'$, $\mathbf{V}_{\mathbf{x}} = \mathbb{E}\mathbf{x}\mathbf{x}^{\mathrm{T}}$, s_0 está definida por

$$\mathbb{E}\rho_0\left(\frac{u}{s_0}\right) = b,$$

$$a(\psi) = \mathbb{E}\psi^2\left(\frac{u}{s_0}\right)$$

y

$$b(\psi) = \mathbb{E}\psi'\left(\frac{u}{s_0}\right).$$

Se puede obtener un resultado análogo para el caso de variables predictivas fijas, ver Salibian-Barrera (2006).

Maronna et al. (2006) recomiendan usar un S-estimador con punto de ruptura máximo como el estimador inicial al calcular MM-estimadores. El MM-estimador resultante también tendrá punto de ruptura máximo. Recomiendan tomar $\rho_0 = \rho^B_{c_0}$ y $\rho_1 = \rho^B_{c_1}$ donde $c_1 \geq c_0$ y ρ^B_c es la Bicuadrática de Tukey. La constante de calibración c_0 debe elegirse para que el M-estimador de escala resultante sea consistente para el desvío estandar de los errores en el caso de errores normales. La elección de c_1 debe apuntar a balancear robustez y eficiencia. Maronna et al. (2006) recomiendan tomar c_1 de manera que el MM-estimador resultante tenga una eficiencia asintótica del 85% para el caso de errores normales. La razón para elegir una eficiencia del 85% es la siguiente: a este nivel de eficiencia el máximo sesgo asintótico del MM-estimador coincide con el del S-estimador para el caso de errores y covariables normales. Ver la Sección 5.9 de Maronna et al. (2006).

El hecho de que los MM-estimadores pueden calibrarse para tener tanto un punto de ruptura alto como una eficiencia asintótica alta en la distribución normal hace que sean una de las alternativas más populares dentro de los método robustos de regresión.

Modelos ralos y estimadores penalizados

Los problemas de estimación en modelos lineales ralos y de alta dimensión, donde la razón entre el número de variables predictivas y el número de observaciones, digamos p/n, es alta, pero la razón entre el número de variables predictivas que de hecho son relevantes y el número de observaciones, digamos k/n, es baja, se han hecho cada vez más comunes en áreas como la bioinformática y la quimiometría. En un modelo de regresión ralo y de alta dimensión se tienen un gran número de posibles variables predictivas, posiblemente incluso un número mayor que el número de observaciones, pero se cree que la mayoría de ellas no proporcionan información relevante para predecir la respuesta, es decir, que la mayoría de las coordenadas del verdadero vector de regresión son cero o tienen coeficientes muy pequeños.

En este tipo de modelos de regresión, debido a la posible alta dimensión de los datos, es difícil descubrir observaciones atípicas usando criterios simples. Los estimadores de regresión robustos tradicionales, como los MM-estimadores, no producen modelos ralos y pueden tener un mal comportamiento en lo que respecta a robustez y a eficiencia cuando p/n es alto. Ver Maronna and Yohai (2015) y Smucler and Yohai (2015 a). Más aún, no se pueden calcular cuando $p > n$.

Los M-estimadores de regresión penalizados proveen un marco de trabajo general para la estimación en modelos lineales ralos y de alta dimensión. Sea ρ una ρ-función y sea s_n un estimador de escala. Los M-estimadores de regresión penalizados están definidos por

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) + \sum_{j=1}^{p} p_{\lambda_j^n}(|\beta_j|), \tag{3}$$

donde $p_{\lambda_j^n}$ es una función no negativa, llamada función de penalización, que depende de ciertos parámetros de penalización $\lambda_j^n \geq 0$. El término $\sum_{j=1}^{p} p_{\lambda_j^n}(|\beta_j|)$ mide en cierto sentido la complejidad del modelo de regresión estimado. Cuando el modelo contiene una ordenada al origen, generalmente no se la penaliza. Si la función de penalización es elegida apropiadamente, el M-estimador penalizado correspondiente estará bien definido aún si $p > n$ y producirá modelos ralos. En la práctica, los parámetros de penalización se eligen mediante algún esquema basado en los datos de la muestra, como la validación cruzada.

Aunque nuestra formulación de los M-estimadores penalizados incluye un estimador de escala, el análisis teórico y práctico de estos estimadores se ha limitado hasta ahora a estimadores definidos por (3) pero sin utilizar un estimador de escala. Hasta nuevo aviso, asumiremos que no se utilizó un estimador de escala para estandarizar los residuos en (3).

Tomando $\rho(x) = x^2$ en (3), obtenemos la familia de estimadores de mínimos cuadrados penalizados. Una familia importante de funciones de penalización está dada por $p_{\lambda_j^n}(|\beta_j|) = \lambda_n|\beta_j|^q$, donde $q > 0$. Estas funciones de penalización se llaman de tipo Bridge y fueron introducidas en Frank and Friedman (1993). Notar que en este caso

$$\sum_{j=1}^{p} p_{\lambda_j^n}(|\beta_j|) = \lambda_n \sum_{j=1}^{p} |\beta_j|^q = \lambda_n \|\boldsymbol{\beta}\|_q^q,$$

de modo que el término de penalización es proporcional a la "norma" ℓ_q de los coeficientes. Llamaremos al estimador que resulta de tomar la pérdida cuadrática y una penalidad tipo Bridge en (3), LS-Bridge. Dos casos particulares muy importantes son:

- La penalidad ℓ_2, que se obtiene tomando $q = 2$ y que junto con la pérdida cuadrática da lugar al estimador LS-Ridge, introducido en Hoerl and Kennard (1970).

- La penalidad ℓ_1, que se obtiene tomando $q = 1$ y que junto con la pérdida cuadrática da lugar al estimador LS-Lasso, introducido en Tibshirani (1996).

Notar que los estimadores de regresión definidos usando la penalidad ℓ_2, o más en general, una función de penalización suave, no producen modelos ralos. Ver Fan and Li (2001).

Por otro lado, el estimador LS-Lasso si produce modelos ralos. Más aún, si $p > n$, el número de coeficientes distintos de cero de la estimación LS-Lasso es a lo sumo n para cualquier parámetro de penalización positivo, ver la Sección 2.6 de Buhlmann and van de Geer (2011).

El problema de optimización que define al estimador LS-Lasso es convexo y existen algoritmos muy eficientes para resolverlo, por ejemplo el algoritmo LARS desarrollado por Efron et al. (2004) o Coordinate Descent Optimization. Los estimadores tipo LS-Bridge con $q < 1$ también producen modelos ralos, pero su cálculo es más complicado computacionalmente.

Otra función de penalización comunmente utilizada es la función SCAD, propuesta en Fan and Li (2001). Está dada por

$$p'_{\lambda,a}(|\beta|) = \lambda \left\{ I\{|\beta| \leq \lambda\} + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I\{|\beta| > \lambda\} \right\}$$

donde $a > 2$ y $p_{\lambda,a}(0) = 0$. La función de penalización SCAD tiene varias propiedades teóricas interesantes, ver Fan and Li (2001). Tomando como función de pérdida en (1.6) a la función cuadrática y como función de penalización a la función SCAD, se obtiene el estimador LS-SCAD. El estimador LS-SCAD produce modelos ralos, pero su cómputo es un tanto complicado.

Las propiedades teóricas de los estimadores de mínimos cuadrados penalizados han sido estudiadas extensamente en los último años. De especial interés es la llamada *propiedad oráculo* definida en Fan and Li (2001): un estimador tiene la propiedad oráculo si los coeficientes estimados que se corresponden a coeficientes nulos del parámetro de regresión verdadero se estiman como exactamente cero con probabilidad que tiende a uno, mientras que al mismo tiempo los coeficientes que se corresponden a coeficientes no nulos del parámetro de regresión verdadero se estiman con la misma eficiencia asintótica que tendriamos si hubieramos conocido el modelo correcto a priori.

Knight and Fu (2000) derivan la distribución asintótica de los estimadores LS-Bridge para todo $q > 0$ y prueban que para $q < 1$ estos estimadores pueden tener la propiedad oráculo. También prueban que para $q = 1$, las distribuciones asintóticas de las coordenadas del LS-Lasso que se corresponden a coeficientes nulos de $\boldsymbol{\beta}_0$ pueden poner probabilidad positiva en cero. El LS-Lasso no tiene la propiedad oráculo en general; ver Zou (2006) y Buhlmann and van de Geer (2011). Además, el LS-Lasso tiene un problema de sesgo: puede achicar excesivamente a los coeficientes grandes.

Para remediar este problema Zou (2006) introdujo el LS-Lasso adaptivo, donde se usan pesos adaptivos para penalizar los distintos coeficientes, y mostró que este estimador puede tener la propiedad oráculo. Sea $\hat{\boldsymbol{\beta}}_{ini}$ un estimador inicial (por ejemplo el LS-Lasso) y $\varsigma > 0$. Entones el LS-Lasso adaptivo está definido por

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{ini,j}|^\varsigma}.$$

Notar que para los coeficientes que se corresponden a coeficientes grandes de $\hat{\boldsymbol{\beta}}_{ini}$, el LS-Lasso adaptivo utiliza una penalización pequeña. Zou (2006) muestra que el estimador LS-Lasso adaptivo puede computarse utilizando cualquier algoritmo que compute el estimador LS-Lasso.

En Fan and Li (2001), los autores muestran que existe un *mínimo local* de la función objetivo que define al LS-SCAD que tiene la propiedad oráculo.

Los estimadores de mínimos cuadrados penalizados no son robustos y pueden tener una eficiencia baja cuando los errores tienen colas pesadas. En un intento por resolver estos problemas, se han propuesto M-estimadores penalizados definidos utilizando una función de pérdida convexa. Por ejemplo, en Wang et al. (2007) los autores proponen tomar la pérdida del valor absoluto, $\rho(x) = |x|$, y una penalidad como la del Lasso adaptivo. En ese trabajo, muestran que el estimador que proponen puede tener la propiedad oráculo. Li et al. (2011) estudian estimadores definidos utilizando la pérdida de Huber o la del valor absoluto y la penalidad SCAD. También se han propuesto estimadores basados en rangos, ver por ejemplo Johnson and Peng (2008) y Leng (2010). Todos estos estimadores apuntan a lograr robustez frente a observaciones atípicas en la variables respuesta. Desafortunadamente, no son robustos con respecto a contaminaciones en la variables predictivas.

Khan et al. (2007) proponen una versión robusta del algoritmo LARS, llamado RLARS. Sin embargo, como este procedimiento no está basado en la minimización de una función objetivo, el análisis de sus propiedades teóricas no parece accesible. En Wang and Li (2009) los autores proponen un estimador tipo Wilcoxon pesado con una penalización SCAD y muestran que su estimador tiene la propiedad oráculo. Sin embargo, veremos en la Sección 2.3 que este estimador puede verse muy afectado por la presencia de observaciones atípicas. Alfons et al. (2013) proponen el estimador Sparse-LTS, un estimador de mínimos cuadrados podados con una penalidad ℓ_1. En un estudio de simulación, los autores muestran que el Sparse-LTS puede ser robusto respecto de contaminaciones tanto en la variables respuesta como en las variables predictivas. Este estimador puede ser calculado para $p > n$. Sin embargo, los autores no dan ningún resultado asintótico para el Sparse-LTS. Wang et al. (2013) proponen un estimador penalizado basado en un pérdida de tipo exponencial. Prueban que un mínimo local de la función objetivo usada para definir el estimador tiene la propiedad oráculo. Por otro lado, este estimador no puede calcularse para el caso de $p > n$. En Ollerer et al. (2014) los autores estudian las funciones de influencia de los M-estimadores de regresión penalizados. Gijbels and Vrinssen (2015) proponen versiones tipo garrote no-negativo de varios estimadores de regresión robustos, incluyendo S y MM-estimadores. Sin embargo, los autores no dan ningún resultado teórico para estos estimadores y estos no pueden calcularse para el caso $p > n$. Avella-Medina (2016) propuso M-estimadores penalizados para modelos lineales generalizados y modelos aditivos generalizados y derivó la función de influencia de M-estimadores penalizados bajo condiciones generales. Maronna (2011) propone S y MM-estimadores de regresión con una penalidad ℓ_2. En un extenso estudio de simulación, muestra que estos estimadores son robustos en una amplia variedad de escenarios. Recordar de cualquier forma que los estimadores con una penalidad ℓ_2 no producen modelos ralos. Maronna (2011) no da ningún resultado asintótico para los estimadores que propone.

En la primera parte de esta tesis, que consiste del Capítulo 2, definimos los estimadores MM-Bridge y MM-Bridge adaptivos: MM-estimadores con una penalidad ℓ_q y una penalidad

ℓ_t adaptiva respectivamente. Calculamos el punto de ruptura de los estimadores MM-Bridge y damos una cota inferior para el punto de ruptura de los estimadores MM-Bridge adaptivos. Para el caso de un número fijo de variables predictivas aleatorias, probamos la consistencia fuerte de los estimadores MM-Bridge y MM-Bridge adaptivos. Derivamos la distribución asintótica de los estimadores MM-Bridge para todo q y mostramos que para $q < 1$ pueden tener la propiedad oráculo. Probamos también que los estimadores MM-Bridge adaptivos pueden tener la propiedad oráculo para todo $t \leq 1$. Proponemos un algoritmo para computar tanto los estimadores MM-Bridge con $q = 1$, que llamamos estimadores MM-Lasso, como los estimadores MM-Bridge adaptivo con $t = 1$, que llamamos MM-Lasso adaptivo. Nuestro algoritmo utiliza el S-estimador con penalidad ℓ_2 de Maronna (2011) como estimador inicial y resuelve iterativemente un problema del tipo Lasso con pesos. Los estimadores MM-Lasso y MM-Lasso adaptivos pueden calcularse para el caso $p > n$.

Modelos con un número de parámetros diverge

En los últimos cuarenta años, ha habido un cambio de paradigma en el análisis asintótico de ciertos problemas estadísticos multivariados. El creciente número de problemas estadísticos con un gran número de parámetros ha motivado el estudio de las propiedades asintóticas de estimadores en modelos con un número de parámetros que crece con el tamaño de la muestra.

Para el caso de la regresión lineal, consideremos una sucesión de modelos de regresión

$$y_{i,n} = \mathbf{x}_{i,n}^T \boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n.$$

donde $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$ son vectores fijos, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ es un parámetro vectorial a estimar y $u_{i,n}$ son variables aleatorias i.i.d. definidas en un mismo espacio de probabilidad. Notar que p_n puede depender de n en un manera tal que p_n tiende a infinito a cierta tasa.

A continuación, repasamos brevemente la historia del estudio de las propiedades de estimadores para modelos de regresión lineal con un número de parámetros que diverge. El primer análisis de este problema aparece en Huber (1973). Huber (1973) estudia las propiedades asintóticas de M-estimadores de regresión monótonos definidos sin utilizar un estimador de escala. Motivado por problemas que surgen en la cristalografía de rayos X, Huber propuso estudiar las propiedades de estos estimadores cuando $p = p_n \to \infty$. Huber prueba la normalidad asintótica de los contrastes lineales de estos estimadores cuando $p^3/n \to 0$. Este resultado fue mejorado por Yohai and Maronna (1979), quienes, bajo esencialmente las mismas hipótesis que Huber (1973), prueban la normalidad asintótica de los contrastes lineales asumiendo solo $p^{5/2}/n \to 0$ y también la consistencia con tasa $\sqrt{n/p}$ asumiendo que $p^2/n \to 0$. Yohai and Maronna (1979) también dan resultados análogos para M-estimadores monótonos definidos utilizando un estimador de escala. Portnoy (1984) y Portnoy (1985) estudian las propiedades asintóticas de las soluciones de (2), sin incluir un estimador de escala y donde la función de pérdida no es necesariamente convexa. Para el caso de una pérdida convexa, y bajo ciertas hipótesis técnicas sobre las covariables, Portnoy prueba la consistencia con

tasa $\sqrt{n/p}$ y la normalidad asintótica de los contrastes lineales asumiendo $(p \log p)/n \to 0$ y $(p \log n)^{3/2}/n \to 0$ respectivamente. Para el caso de una pérdida no convexa, prueba que valen resultados análogos para *alguna* solución de (2). Mammen (1988) obtiene desarrollos asintóticos para las soluciones de (2), sin incluir un estimador de escala y donde la función de pérdida es convexa, asumiendo solo que $p^{3/2} \log n/n \to 0$. También prueba que valen resultados análogos para alguna solución de (2) cuando la función de pérdida no es necesariamente convexa y una escala se estima simultáneamente. Welsh (1989) obtiene resultados bajo hipótesis menos estrictas sobre la regularidad de la función de pérdida ρ pero bajo condiciones más estrictas sobre la tasa a la que puede crecer p. Bai and Wu (1994) y Bai and Wu, part II (1994) generalizan los resultados enumerados anteriormente, al relajar las condiciones de regularidad impuestas sobre ρ o la tasa a la cual puede crecer p. Por ejemplo, para el caso de un función de pérdida convexa y lo suficientemente suave, los autores prueban la consistencia y normalidad asintótica de los M-estimadores asumiendo que $p/n \to 0$ y $p^2/n \to 0$ respectivamente.

Ninguno de los resultados mencionados anteriormente se puede aplicar directamente a M-estimadores definidos utilizando una función de pérdida acotada o a estimadores con un alto punto de ruptura como los S y MM-estimadores. Davies (1990) prueba la consistencia de los S-estimadores de regresión asumiendo que $(p \log n)/n \to 0$.

Más recientemente, El Karoui et al. (2013), El Karoui (2013), Donoho and Montanari (2015 a),Donoho and Montanari (2015 b) y Nevo and Ritov (2016) estudian las propiedades de M-estimadores monótonos asumiendo que $p/n \to m \in (0, 1)$.

Fan and Peng (2004) estudian las propiedades asintóticas de estimadores de máxima verosimilitud penalizados con un número de parámetros que diverge. Prueban que existe un *máximo local* de la función objetivo que define a los estimadores que es consistente con tasa $\sqrt{n/p}$ asumiendo que $p^4/n \to 0$ y que ese máximo local tiene la propiedad oráculo asumiendo que $p^5/n \to 0$. Para el caso particular de regresión lineal con pérdida cuadrática, este resultado fue mejorado por Huang and Xie (2007), quienes prueban que el *mínimo global* tiene la propiedad oráculo asumiendo que p crece con n a cierta tasa que depende, entre otras cosas, del número de coordenadas no nulas del vector de regresión verdadero. Huang et al. (2008) prueban que para todo $q > 0$ los estimadores LS-Bridge son consistentes con tasa $\sqrt{n/p}$ y que para todo $q < 1$ tienen la propiedad oráculo siempre que p crezca a cierta tasa que depende, entre otras cosas, del número de coordenadas no nulas del vector de regresión verdadero. Zou and Zhang (2009) prueban que el LS-Lasso adaptivo converge con tasa $\sqrt{n/p}$ y que tiene la propiedad oráculo asumiendo que $(\log p/ \log n) \to \nu < 1$. Huang, Ma and Zhang (2008) prueban que el LS-Lasso adaptivo puede tener la propiedad oráculo aún cuando $p \gg n$. Li et al. (2011) estudian las propiedades asintóticas de M-estimadores penalizados definidos utilizando una función de pérdida convexa y probaron que existe un mínimo local de la función objetivo usada para definir los estimadores que es consistente con tasa $\sqrt{n/p}$ si $(p \log n)/n \to 0$ y que ese mínimo local tiene la propiedad oráculo si $p^2/n \to 0$.

En Buhlmann and van de Geer (2011) se puede encontrar un excelente análisis de las

propiedades teóricas de estimadores regularizados de regresión cuando $p \gg n$. Más recientemente, en Loh (2015) la autora estudia las propiedades teóricas de M-estimadores de regresión penalizados cuando $p \gg n$. Muestra que, bajo condiciones de regularidad, todos los puntos estacionarios del problema de optimización que define a los estimadores que estan suficientemente cerca de $\boldsymbol{\beta}_{0,n}$ convergen a la misma tasa que el LS-Lasso para el caso de errores sub-Gaussianos. Más aún, muestra que si la función de penalización es no convexa y elegida convenientemente, esos puntos estacionarios son iguales a la solución oráculo. Sin embargo, estos resultados no pueden aplicarse a los estimadores que estudiamos en esta tesis.

En la segunda parte de esta tesis, que consiste del Capítulo 3, estudiamos las propiedad asintóticas de versiones más generales de los estimadores MM-Bridge y MM-Bridge adaptivos, más precisamente, M-estimadores penalizados con una penalidad tipo Bridge y una función de pérdida acotada, en modelos lineales con un número de parámetros que diverge en el régimen $p \ll n$. Probamos su consistencia bajo condiciones muy generales, asumiendo solo que $p/n \to 0$. Asumiendo las mismas hipótesis sobre las covariables que asume Portnoy (1984) y que $(p \log n)/n \to 0$ probamos que los estimadores son consistentes con tasa $\sqrt{n/p}$. Estos resultados incluyen como caso particular la consistencia de M-estimadores definidos usando una pérdida acotada y un estimador de escala y por lo tanto prueban la consistencia de los S y MM-estimadores. Probamos que los estimadores MM-Bridge con $q < 1$ y los estimadores MM-Bridge adaptivos con $t \leq 1$ de hecho son consistentes con tasa $\sqrt{n/k}$, donde k es el número de coeficientes no nulos del verdadero vector de regresión, y que tienen la propiedad oráculo. Con la misma técnica, podemos probar la consistencia y normalidad asintótica de M-estimadores definidos usando una pérdida acotada y un estimador de escala y por lo tanto prueban la de los S y MM-estimadores.

El resto de esta tesis esta organizada de la siguiente manera. En el Capítulo 2 definimos los estimadores que proponemos, estudiamos sus puntos de ruptura, estudiamos sus propiedad asintóticas para el caso de p fijo y predictores aleatorios y proponemos un algoritmo para computarlos. En un estudio de simulación extensivo, estudiamos su rendimiento en lo que respecta a estabilidad en la presencia de observaciones atípicas y también a sus propiedades de predicción y de selección de variables cuando los datos siguen el modelo asumido. También aplicamos los estimadores propuestos a un conjunto de datos reales. En el Capítulo 3 estudiamos las propiedades asintóticas de M-estimadores penalizados definidos usando una pérdida acotada y un estimador de escala, en modelos de regresión con un número de variables predictivas fijas que diverge. Las pruebas de los resultados originales de esta tesis se encuentran en el Capítulo 4.

Chapter 1

Introduction

1.1 Linear regression

A problem common to all branches of science and technology is to explain the relation between a response variable y and a vector of predictor variables $\mathbf{x}$. The simplest way to model that relation is via a linear model. More precisely, suppose that we observe $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ $i = 1, \ldots, n$, $(p+1)$-dimensional vectors, where y_i is a response variable and $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of predictor variables. The predictor variables may be fixed, as is the case in a designed experiment, or random, if for example they are observational variables. We assume that $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, satisfy

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + u_i,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the vector of regression coefficients to be estimated and the u_i $i = 1, \ldots, n$ are i.i.d. random error variables defined in a common probability space. For the case of random predictor variables, we will assume that the vectors of predictor variables are i.i.d. and independent of the errors.

The least-squares estimator (LSE) of $\boldsymbol{\beta}_0$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}),$$

where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ is the residual of the i-th observation. It is well known that, under regularity assumptions, the LSE satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} N_p\left(\mathbf{0}, \sigma^2 \mathbf{C}^{-1}\right),$$

where σ^2 is the error variance, $\mathbf{C} = \mathbb{E}\mathbf{x}\mathbf{x}^T$ for the case of random predictors and $\mathbf{C} = \lim 1/n \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ for the case of fixed predictors. When the errors are normally distributed, the LSE of $\boldsymbol{\beta}_0$ is equal to the maximum likelihood estimator. However, it is well known that the LSE is not robust, it is very sensitive to deviations from the model assumptions.

1.2 Robust regression estimators

The main goal of robust statistics is to provide methods for statistical analysis that are reliable in the presence of atypical data or outliers. A further goal is to obtain robust estimators that are almost as good as the optimal classical estimator (say the maximum likelihood estimator) when the data distribution coincides with the nominal one, typically the normal distribution.

The robustness of an estimator is measured by its stability when a small fraction of the observations is arbitrarily replaced by outliers that may not follow the assumed model. A robust estimator should not be much affected by a small fraction of outliers. A popular quantitative measure of an estimator's robustness, introduced by Donoho and Huber (1983), is the finite-sample replacement breakdown point. Very loosely speaking, the finite-sample replacement breakdown point of an estimator is the maximum fraction of outliers that the estimator may tolerate without losing all meaning. For a regression estimator, this measure is defined as follows. Given a sample $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, let $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ and let $\hat{\boldsymbol{\beta}}(\mathbf{Z})$ note the regression estimator $\hat{\boldsymbol{\beta}}$ computed in $\mathbf{Z}$. The finite-sample replacement breakdown point of $\hat{\boldsymbol{\beta}}$ is then defined as

$$FBP(\hat{\boldsymbol{\beta}}) = \frac{m^*}{n},$$

where

$$m^* = \max\left\{m \geq 0 : \hat{\boldsymbol{\beta}}(\mathbf{Z}_m) \text{ is bounded for all } \mathbf{Z}_m \in \mathcal{Z}_m\right\},$$

and $\mathcal{Z}_m$ is the set of all datasets with at least $n-m$ elements in common with $\mathbf{Z}$. Let $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})$ note the estimator $\hat{\boldsymbol{\beta}}$ computed in the dataset $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is the matrix with $\mathbf{x}_i$ as rows and $\mathbf{y} = (y_1, \ldots, y_n)$. The estimator $\hat{\boldsymbol{\beta}}$ is said to be regression equivariant if for any $\mathbf{b} \in \mathbb{R}^p$, $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\mathbf{b}) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}) + \mathbf{b}$. It can be shown that any regression equivariant estimator has a breakdown point of at most $[(n-p)/2]/n$, which is approximately $1/2$ for $p \ll n$. See, for example, Section 5.4.1 of Maronna et al. (2006).

A breakdown point equal to ε^* guarantees that for any given contamination fraction $\varepsilon \leq \varepsilon^*$ there exists a compact set, say $K = K(\varepsilon)$, such that the estimator in question remains in K whenever a fraction of ε observations is arbitrarily modified. However, this compact set may be very large. Thus, even though a high breakdown point is always a desirable property, an estimator with a high breakdown point can still be severely affected by a small fraction of outlying observations.

It is well known that the LSE has zero breakdown point, that is, a single outlying observation can completely ruin it.

A general framework for robust estimation in the linear model is provided by M-estimators. The notion of an M-estimator was first introduced in the landmark paper Huber (1964) for the case of the estimation of a location parameter and extended to the linear model in Huber (1973). To define them, we first introduce some notation. Throughout this thesis, we will say that ρ is a ρ-function if

1. ρ is even and continuous.

2. $\rho(x)$ is a nondecreasing function of $|x|$.

3. $\rho(0) = 0$.

4. If $\rho(v) < \lim_{x\to\infty} \rho(x)$ and $0 \leq u < v$ then $\rho(u) < \rho(v)$.

5. If ρ is bounded, $\lim_{x\to\infty} \rho(x) = 1$.

If ρ is a differentiable ρ-function we will set $\psi = \rho'$.

A popular family of ρ-functions used in robust regression is Huber's family of loss functions, given by

$$\rho_c^H(x) = x^2 I\{|x| \leq c\} + \left(2|x|c - c^2\right) I\{|x| \geq c\},$$

where $c > 0$ is some tuning constant. Huber's loss functions are a compromise between the quadratic loss, that defines the LSE, and the absolute value loss, that defines the well known Least Absolute Deviations (LAD) estimator. Another popular family of ρ-functions is Tukey's Bisquare family of ρ-functions, given by

$$\rho_c^B(x) = 1 - \left(1 - \left(\frac{x}{c}\right)^2\right)^3 I\{|x| \leq c\},$$

where $c > 0$ is some tuning constant. Note that ρ_c^B is bounded, moreover, if $|x| \geq c$, then $\rho_c^B(x) = 1$. In Figure 1.1 we show a plot of several popular loss functions.

In order to define regression M-estimators in full generality, we first need to define M-estimators of scale. Let ρ_0 be a bounded ρ-function. Given a sample $\mathbf{u} = (u_1, ..., u_n)$ and $0 < b < 1$ the corresponding M-estimate of scale $s_n(\mathbf{u})$ is defined, Huber (1981), by

$$s_n(\mathbf{u}) = \inf\left\{s > 0 : \frac{1}{n}\sum_{i=1}^n \rho_0\left(\frac{u_i}{s}\right) \leq b\right\}.$$

It is easy to prove that $s_n(\mathbf{u}) > 0$ if and only if $\#\{i : u_i = 0\} < (1-b)n$, and in this case

$$\frac{1}{n}\sum_{i=1}^n \rho_0\left(\frac{u_i}{s_n(\mathbf{u})}\right) = b.$$

Note that M-estimators of scale are scale equivariant, in the sense that for any c, $s_n(c\mathbf{u}) = |c|s_n(\mathbf{u})$.

Given a ρ-function ρ, the corresponding regression M-estimator is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right). \tag{1.1}$$

where s_n is an estimate of scale of the residuals that may be estimated a priori or simultaneously. For example, s_n could be the M-estimate of scale of the residuals or the median of

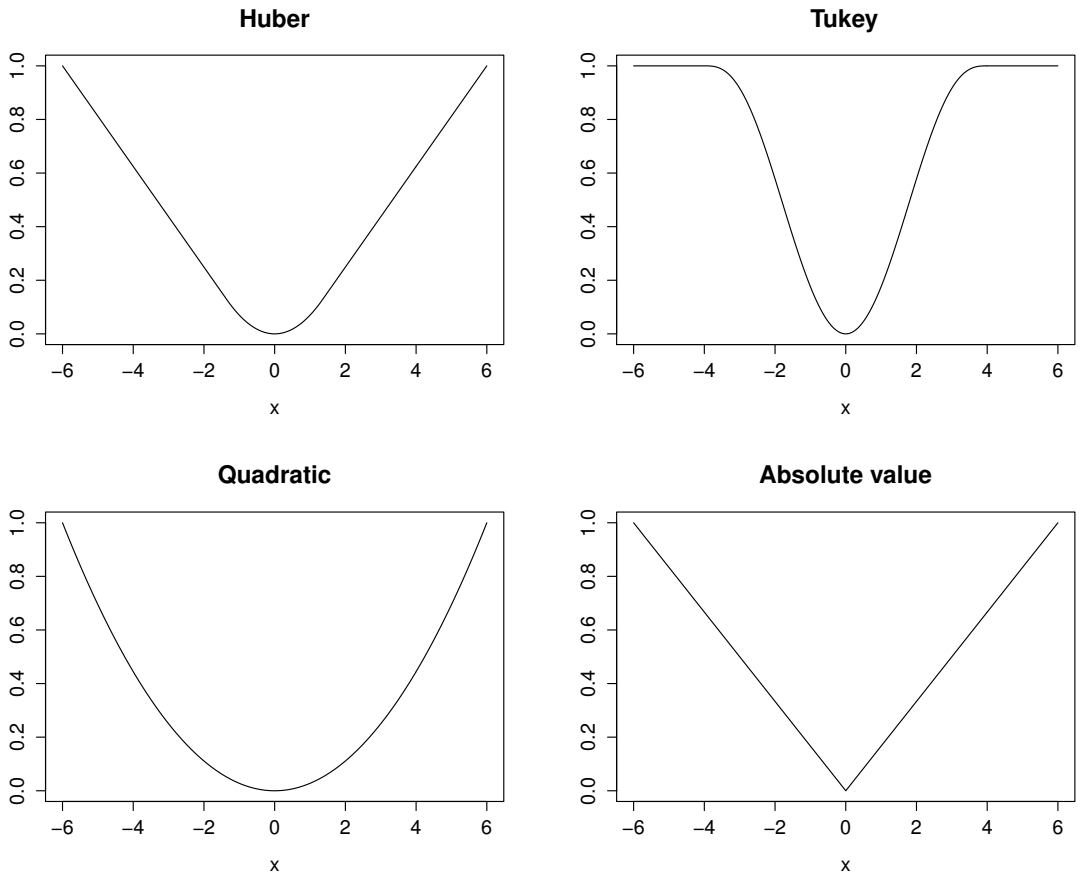

Figure 1.1: Plots of Huber's loss function, Tukey's loss function, the quadratic loss function and the absolute value loss function. All functions were scaled so as to have a maximum equal to 1 on the $[-6, 6]$ interval.

the absolute values of the residuals of some initial regression estimator. Not using an estimate of scale in (1.1) is the same as, somewhat abusing notation, taking s_n to be a constant equal to 1. If s_n is scale equivariant, dividing the residuals by s_n in (1.1) makes the corresponding regression M-estimator scale equivariant, that is: for any c, $\hat{\boldsymbol{\beta}}(\mathbf{X}, c\mathbf{y}) = c\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})$. M-estimators are regression equivariant when, for example, an estimate of scale is not used in the definition of the estimators or when s_n is the M-estimate of scale of the residuals of some regression equivariant estimator. These equivariance properties are desirable, since they allow us to know how the estimates change under these transformations of the data.

The LSE is obtained by taking $\rho(x) = x^2$ in (1.1), whereas the LAD estimator is obtained by taking $\rho(x) = |x|$. Note that for these two cases, a scale estimate is not needed to make the resulting estimators scale equivariant. To obtain robust estimators, one generally uses a

ρ-function in (1.1) that increases less rapidly than the quadratic function.

For the case of a convex and differentiable loss function, for example Huber's loss function, (1.1) is essentially equivalent to

$$\sum_{i=1}^{n} \psi \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{s_n} \right) \mathbf{x}_i = \mathbf{0}, \tag{1.2}$$

where $\psi = \rho'$; see Section 7.3 of Huber (1981) and Section 4.4 of Maronna et al. (2006). In this case, the resulting M-estimator is called a monotone regression M-estimator. When ψ tends to zero at infinity the resulting estimator is called a redescending regression M-estimator and in this case some solutions of (1.2) may not correspond to solutions of (1.1).

Huber (1964) showed that M-estimators of location defined using Huber's loss function have a minimax optimality property: loosely speaking, if the tuning constant is appropriately chosen, they minimize the maximum asymptotic variance over gross-error neighbourhoods of the normal distribution. Huber (1973) studied the asymptotic properties of monotone regression M-estimators defined by (1.1) for the case of fixed predictor variables, but without including the estimate of scale. He showed that, under regularity assumptions, these estimators are $\sqrt{n}$-consistent and asymptotically normal. It can be shown that, if we only entertain the possibility of outliers in the response variable, monotone M-estimators defined by (1.2) with a bounded ψ may have a high breakdown point. This holds, for example, for the cases of one-way or two-way ANOVA designs; see Section 4.6 of Maronna et al. (2006). However, if outliers in the predictor variables are a possibility, the breakdown point of monotone regression M-estimators is zero; see Section 5.16.1 of Maronna et al. (2006).

The least median of the squares estimator (LMSE) is a regression equivariant estimator that has the optimal 1/2 asymptotic breakdown point. The LMSE was first proposed by Hampel (1975) and further developed by Rousseeuw (1984). It is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \text{median } (r_i(\boldsymbol{\beta})^2)_{i=1}^n.$$

The LMSE is $n^{1/3}$-consistent and so its asymptotic efficiency for the case of normal errors is 0. See Davies (1990) and Kim and Pollard (1990).

S-estimators, introduced in Rousseeuw and Yohai (1984), combine the usual $\sqrt{n}$ rate of consistency with a high breakdown point. They are defined by

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} s_n(\mathbf{r}(\boldsymbol{\beta}))$$

where $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$ and $s_n()$ is an M-estimator of scale. It is easy to verify that S-estimators are scale and regression equivariant. Let $\hat{s}_n = s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$ and let ρ_0 be the ρ-function used to define $s_n()$. Then, S-estimators satisfy

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_0 \left(\frac{r_i(\boldsymbol{\beta})}{\hat{s}_n} \right),$$

see Section 5.6.1 of Maronna et al. (2006). Hence, S-estimators are M-estimators in the sense of (1.1), where the loss function ρ is bounded and the scale is estimated simultaneously. In practice, ρ_0 is usually chosen so that it satisfies $\rho_0(x) = 1$ if $|x| \geq m$ for some m. For example, ρ_0 could be Tukey's Bisquare loss. The asymptotic distribution of regression S-estimators was derived, under very general conditions, by Fasano et al. (2012) for the case of random predictors and by Davies (1990) for the case of fixed predictors. S-estimators can always be tuned so as to attain the maximum possible finite-sample replacement breakdown point for regression equivariant estimators; see Section 5.6.1 of Maronna et al. (2006). However, S-estimators cannot combine a high breakdown point with a high efficiency at the normal distribution, see Hössjer (1992).

MM-estimators, introduced in Yohai (1987), are regression estimators that can be tuned to attain both a high breakdown point and an arbitrarily high asymptotic efficiency at the normal distribution. Suppose $\hat{\boldsymbol{\beta}}_1$ is a highly robust, but not necessarily highly efficient, initial estimator. In practice, $\hat{\boldsymbol{\beta}}_1$ will usually be an S-estimator. Let $s_n()$ be an M-estimator of scale defined using a bounded ρ-function ρ_0 and b. Let ρ_1 be another ρ-function that satisfies $\rho_1 \leq \rho_0$. Then the MM-estimator is defined by

$$\hat{\boldsymbol{\beta}}_{MM} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))} \right).$$

Note that MM-estimators are M-estimators, as in (1.1), defined using a bounded loss function and a preliminary estimate of scale. MM-estimators are scale and regression equivariant whenever $\hat{\boldsymbol{\beta}}_1$ satisfies these properties. We note that the original definition of MM-estimators is actually more general, but for technical convenience we will work with this definition. Yohai (1987) proved that under regularity assumptions and for the case of random predictor variables MM-estimators satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MM} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p \left(\mathbf{0}, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \mathbf{V}_{\mathbf{x}}^{-1} \right),$$

where $\psi_1 = \rho_1'$, $\mathbf{V}_{\mathbf{x}} = \mathbb{E}\mathbf{x}\mathbf{x}^{\mathrm{T}}$, s_0 is defined by

$$\mathbb{E}\rho_0 \left(\frac{u}{s_0} \right) = b, \tag{1.3}$$

$$a(\psi) = \mathbb{E}\psi^2 \left(\frac{u}{s_0} \right) \tag{1.4}$$

and

$$b(\psi) = \mathbb{E}\psi' \left(\frac{u}{s_0} \right). \tag{1.5}$$

An analogous result can be obtained for the case of fixed predictors, see Salibian-Barrera (2006).

Maronna et al. (2006) recommend the use of an S-estimator with maximal breakdown point as the initial estimator when computing MM-estimators. The resulting MM-estimator will also have maximal breakdown point. They recommend taking $\rho_0 = \rho^B_{c_0}$ and $\rho_1 = \rho^B_{c_1}$ with $c_1 \geq c_0$, where ρ^B_c is Tukey's Bisquare loss. The tuning constant c_0 should be chosen so that the resulting M-estimator of scale be consistent for the error standard deviation in the case of normal errors. The choice of c_1 should aim at striking a balance between robustness and efficiency. Maronna et al. (2006) recommend to choose c_1 so that the MM-estimator has an asymptotic efficiency of 85% at the normal distribution. The reason for choosing an 85% asymptotic efficiency at the normal distribution is the following: at this level of the efficiency the MM-estimator has the same maximum asymptotic bias as the initial S-estimator for the case of normal errors and normal covariates. See Section 5.9 of Maronna et al. (2006).

The fact that MM-estimators can be tuned to attain both a high breakdown point and an arbitrarily high asymptotic efficiency at the normal distribution has made them one of the most popular alternatives robust regression has to offer.

For the sake of brevity, several interesting methods and theoretical concepts have been omitted from our brief account of robust regression. See Maronna et al. (2006) and the references therein.

1.3 Sparsity and penalized estimators

In modern regression analysis, sparse and high-dimensional estimation problems where the ratio of the number of predictor variables to the number of observations, say p/n, is high, but the ratio of the number of actually relevant predictor variables to the number of observations, say k/n, is low, have become increasingly common in areas such as bioinformatics and chemometrics. In a sparse and high-dimensional regression model, we have many candidate predictor variables, possibly even more than the number of observations, but we believe that most of them do not provide relevant information to predict the response, that is, that most of the coordinates of the true regression vector have either zero or very small coefficients.

In this type of regression scenarios, due to the high-dimensional nature of the data, it is difficult to discover outlying observations using simple criteria. Traditional robust regression estimators, such as MM-estimators, do not produce sparse models and can have a bad behaviour with regards to robustness and efficiency when p/n is high. See Maronna and Yohai (2015) and Smucler and Yohai (2015 a). Moreover, they cannot be computed for $p > n$.

A general framework for estimation in sparse and high-dimensional linear models is that of penalized regression M-estimators. Let ρ be a ρ-function and let s_n be an estimate of scale. Penalized regression M-estimators are defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) + \sum_{j=1}^{p} p_{\lambda_j^n}(|\beta_j|), \tag{1.6}$$

where $p_{\lambda_j^n}$ is a non-negative function, called the penalty function, that depends on some penalty parameters $\lambda_j^n \geq 0$. The term $\sum_{j=1}^p p_{\lambda_j^n}(|\beta_j|)$ measures in some sense the complexity of the estimated regression model. When the model contains an intercept, it is generally not penalized. If the penalty function is chosen appropriately, then the corresponding penalized M-estimator will be defined even if $p > n$ and will produce sparse models. In practice, the penalty parameters are usually chosen via some data-driven procedure such as cross-validation.

Even though our formulation of penalized M-estimators includes an estimate of scale, to the best of our knowledge both the theoretical and practical analysis of penalized M-estimators has been concerned, up to now, only with estimators defined by (1.6) but without using an estimate of scale. Until further notice, we will assume that no scale estimate is used to standardize the residuals in (1.6).

Taking $\rho(x) = x^2$ in (1.6), we obtain the family of penalized least-squares estimators. An important family of penalty functions is given by $p_{\lambda_j^n}(|\beta_j|) = \lambda_n |\beta_j|^q$, where $q > 0$. These are called the Bridge penalty functions, introduced in Frank and Friedman (1993). Note that in this case

$$\sum_{j=1}^p p_{\lambda_j^n}(|\beta_j|) = \lambda_n \sum_{j=1}^p |\beta_j|^q = \lambda_n \|\boldsymbol{\beta}\|_q^q,$$

so that the penalty term is proportional to the q-th power of the ℓ_q "norm" of the coefficients. We will call the estimator that results from taking the quadratic loss and a Bridge penalty in (1.6) LS-Bridge. Two very important special cases are:

- The ℓ_2 penalty, obtained by taking $q = 2$, which together with the quadratic loss results in the LS-Ridge estimator introduced in Hoerl and Kennard (1970).

- The ℓ_1 penalty, obtained by taking $q = 1$, which together with the quadratic loss results in the LS-Lasso estimator introduced in Tibshirani (1996).

We note that penalized estimators defined using the ℓ_2 penalty, or more generally smooth penalty functions, do not produce sparse models, see Fan and Li (2001).

On the other hand, the LS-Lasso does produce sparse models. Moreover, if $p > n$, then the number of non-zero coefficients of the LS-Lasso solution is at most n for any positive penalty parameter, see Section 2.6 of Buhlmann and van de Geer (2011). The optimization program that defines the LS-Lasso estimator is convex and there exist very efficient algorithms to solve it, for example the LARS algorithm developed by Efron et al. (2004) or Coordinate Descent Optimization. LS-Bridge estimators with $q < 1$ also produce sparse models, but their computation is more involved.

Another popular penalty function is the Smoothly Clipped Absolute Deviation Penalty (SCAD), proposed in Fan and Li (2001). It is given by

$$p'_{\lambda,a}(|\beta|) = \lambda \left\{ I\{|\beta| \leq \lambda\} + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I\{|\beta| > \lambda\} \right\},$$

where $a > 2$ and $p_{\lambda,a}(0) = 0$. The SCAD penalty has several interesting theoretical properties, see Fan and Li (2001). Taking the loss function in (1.6) to be the quadratic function, we get the LS-SCAD estimator. The LS-SCAD estimator produces sparse models, but its computation is more involved than that of the LS-Lasso.

The theoretical properties of penalized least squares estimators have been extensively studied in the past years. Of special note is the so called *oracle property* defined in Fan and Li (2001): An estimator is said to have the oracle property if the estimated coefficients corresponding to zero coefficients of the true regression parameter are set to zero with probability tending to one, while at the same time the coefficients corresponding to non-zero coefficients of the true regression parameter are estimated with the asymptotic efficiency we would have if we had known the correct model in advance.

Knight and Fu (2000) derive the asymptotic distribution of LS-Bridge estimators and prove that for $q < 1$ these estimators can have the oracle property. They also show that for $q = 1$, the asymptotic distributions of the coordinates of the LS-Lasso corresponding to zero coefficients of $\boldsymbol{\beta}_0$ can put positive probability at zero. The LS-Lasso estimator does not in general have the oracle property; see Zou (2006) and Buhlmann and van de Geer (2011) for details. Moreover, the LS-Lasso estimator has a bias problem: it can excessively shrink large coefficients.

To remedy this issue, Zou (2006) introduced the adaptive LS-Lasso, where adaptive weights are used for penalizing different coefficients, and showed that the adaptive LS-Lasso can have the oracle property. Let $\hat{\boldsymbol{\beta}}_{ini}$ be an initial estimate (such as the LS-Lasso) and take $\varsigma > 0$. Then, the adaptive LS-Lasso is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{ini,j}|^{\varsigma}}.$$

Note that for coefficients corresponding to large coefficients of $\hat{\boldsymbol{\beta}}_{ini}$, the adaptive LS-Lasso employs a small penalty; this ameliorates the bias issues associated with the ℓ_1 penalty. As Zou (2006) points out, adaptive LS-Lasso estimators can be computed using any of the algorithms available to compute LS-Lasso estimators.

In Fan and Li (2001) the authors prove that there exists a *local minimum* of the objective function used to define the LS-SCAD estimator that has the oracle property.

Penalized least-squares estimators are not robust and may be highly inefficient under heavy tailed errors. In an attempt to remedy this issue, penalized M-estimators defined using a convex loss function have been proposed. For example, in Wang et al. (2007) the authors propose to take the absolute value loss, $\rho(x) = |x|$, and an adaptive Lasso type penalty. They show that their proposed estimator can have the oracle property. Li et al. (2011) study estimators defined using Huber's loss function or the absolute value loss and the SCAD penalty. Estimators based on ranks have also been proposed, see for example Johnson and Peng (2008) and Leng (2010). Zou and Yuan (2008) proposed the adaptive Lasso Penalized Composite Quantile Regression estimator. All of the aforementioned estimators aim at robustness towards

outliers in the response variable and/or heavy-tailed errors. Unfortunately, they are not robust with respect to contaminations in the predictor variables.

Khan et al. (2007) propose a robust version of the LARS procedure. However, since this procedure is not based on the minimization of a clearly defined objective function, a theoretical analysis of its properties is difficult. Wang and Li (2009) proposed a weighted Wilcoxon-type smoothly clipped absolute deviation (WW-SCAD) estimator. They showed that this estimator has the oracle property. Since the weights they use are fixed, the WW-SCAD can be highly unstable in the presence of high leverage outliers. Alfons et al. (2013) proposed the Sparse-LTS estimator, a least trimmed squares estimator with an ℓ_1 penalty. In a simulation study, Alfons et al. (2013) show that the Sparse-LTS can be robust with respect to contamination in both the response and predictor variables. The Sparse-LTS estimator can be calculated for $p > n$. However, Alfons et al. (2013) do not provide any asymptotic theory for their estimator. Wang et al. (2013) proposed a penalized regression estimator based on an exponential squared loss function. They call their estimator ESL-Lasso. They prove that a *local minimum* of the objective function used to define their estimator can have the oracle property. On the other hand, the estimator they propose cannot be calculated in regression scenarios with $p > n$. In Ollerer et al. (2014) the authors study the influence functions of penalized regression M-estimators. Gijbels and Vrinssen (2015) proposed nonnegative garrote versions of several robust regression estimators, including S and MM-estimators. They do not provide any theory for these estimators and they cannot be calculated for $p > n$. Avella-Medina (2016) proposed robust penalized M-estimators for generalized linear and additive models and derived the influence function of penalized M-estimators under general assumptions. Maronna (2011) introduced S-Ridge and MM-Ridge estimators: ℓ_2-penalized S and MM-estimators. In extensive simulation studies he shows that these estimators can be robust in a variety of contamination scenarios. Recall however that ℓ_2-penalized regression estimators do not produce sparse models. Maronna (2011) does not provide any asymptotic theory for these estimators.

In the first part of this thesis, which consists of Chapter 2, we introduce MM-Bridge and adaptive MM-Bridge estimators: ℓ_q-penalized MM-estimators and MM-estimators with an adaptive ℓ_t penalty. We calculate the breakdown point of MM-Bridge estimators and obtain a lower bound on the breakdown point of adaptive MM-Bridge estimators. For the case of a fixed number of random predictor variables, we prove the strong consistency of MM-Bridge and adaptive MM-Bridge estimators under general conditions. We derive the asymptotic distribution of MM-Bridge estimators for all q and prove that for $q < 1$ they can have the oracle property. For the special case of $q = 1$ we show that the asymptotic distributions of the coordinates of the MM-Bridge estimator corresponding to null coefficients of the true regression parameter put positive probability at zero. We show that adaptive MM-Bridge estimators can have the oracle property for all $t \leq 1$. We propose an algorithm to compute both MM-Bridge estimators with $q = 1$, which we call MM-Lasso estimators, and adaptive MM-Bridge estimators with $t = 1$, which we call adaptive MM-Lasso estimators. Our

algorithm uses the S-Ridge estimator of Maronna (2011) as an initial estimator and iteratively solves a weighted-Lasso type problem. MM-Lasso and adaptive MM-Lasso estimators can be computed for $p > n$.

1.4 Asymptotics with a diverging number of parameters

Over the last forty years, there has been a paradigm shift in the asymptotic analysis of certain multivariate statistical problems. The growing number of statistical problems with a large number of parameters has motivated the study of the asymptotic properties of estimators of models with a number of parameters that diverge with the sample size. For the case of linear regression, consider a sequence of linear regression models

$$y_{i,n} = \mathbf{x}_{i,n}^T \boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n.$$

where $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$ are fixed vectors, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ is to be estimated and $u_{i,n}$ are i.i.d. random variables defined in a common probability space. Note that p_n may depend on n in a way such that $p_n \to \infty$ at a certain rate.

A brief and somewhat incomplete history of the study of the asymptotic properties of estimators for linear regression models with a diverging number of parameters in the $p \ll n$ regime goes as follows. To the best of our knowledge, the first analysis of this problem appears in Huber (1973). Huber (1973) studied the asymptotic properties of monotone regression M-estimators defined without using an estimate of scale. Motivated by problems in X-ray crystallography, Huber proposed to study the properties of these estimators when $p = p_n \to \infty$. He proved the asymptotic normality of linear contrasts of these estimators when $p^3/n \to 0$. This result was improved by Yohai and Maronna (1979), who, under essentially the same hypothesis as Huber (1973), proved the asymptotic normality of linear contrasts requiring only $p^{5/2}/n \to 0$ and also the $\sqrt{n/p}$-consistency assuming $p^2/n \to 0$. Yohai and Maronna (1979) also provided analogous results for the case of monotone M-estimators defined using an estimate of scale. Portnoy (1984) and Portnoy (1985) studied the asymptotic properties of the solutions of M-estimating equations, (1.2), without including an estimate of scale and where the loss function is not necessarily convex. For the case of convex loss function, and under some technical assumptions on the covariates, Portnoy proved the $\sqrt{n/p}$-consistency of the estimators and the asymptotic normality of linear contrasts, requiring that $(p \log p)/n \to 0$ and $(p \log n)^{3/2}/n \to 0$ respectively. For the case of a non-convex loss function, an analogous consistency result holds for *some* solution of the M-estimating equations. Mammen (1988) obtained asymptotic expansions for the solutions of (1.2), without including an estimate of scale and where the loss function is convex, assuming only $p^{3/2} \log n/n \to 0$. Analogous results hold for *some* solution of the M-estimating equations when the loss function is not necessarily convex and a scale is simultaneously estimated. Welsh (1989) obtained results under more relaxed assumptions on the regularity of the loss function ρ but under more stringent conditions on the rate of growth of p. Bai and Wu (1994) and Bai and Wu, part

II (1994) further improved the aforementioned results by relaxing the regularity conditions imposed on ρ or the rate of growth of p. For example, for the case of a sufficiently smooth and convex loss function, they proved the consistency and asymptotic normality of M-estimators assuming $p/n \to 0$ and $p^2/n \to 0$ respectively.

None of the aforementioned results are directly applicable to M-estimators defined using a bounded loss function or to high-breakdown point estimators such as S-estimators or MM-estimators. Davies (1990) proved the consistency of regression S-estimators assuming $(p \log n)/n \to 0$.

More recently, El Karoui et al. (2013), El Karoui (2013), Donoho and Montanari (2015 a),Donoho and Montanari (2015 b) and Nevo and Ritov (2016) have studied the asymptotic properties of monotone M-estimators in the $p/n \to m \in (0, 1)$ regime.

Fan and Peng (2004) studied the asymptotic properties of maximum likelihood estimators with a general penalty term and a diverging number of parameters. They proved that there exists a *local maximum* of the objective function used to define the estimators that is $\sqrt{n/p}$ consistent assuming $p^4/n \to 0$ and that this local maximum has the oracle property assuming $p^5/n \to 0$. Specialised to the case of linear regression with quadratic loss, this result was improved by Huang and Xie (2007), who proved that the *global minimum* of the objective function used to define the estimators has the oracle property assuming that p grows with n at a certain rate that depends, among other things, on the number of non-zero coefficients of the true regression vector. Huang et al. (2008) proved that for all $q > 0$, LS-Bridge estimators are $\sqrt{n/p}$-consistent and that for $q < 1$, they have the oracle property assuming p grows with n at a certain rate that depends, among other things, on the number of non-zero coefficients of the true regression vector. Zou and Zhang (2009) proved that adaptive LS-Lasso estimators are $\sqrt{n/p}$-consistent and have the oracle property assuming $(\log p/ \log n) \to \nu < 1$. Huang, Ma and Zhang (2008) proved that adaptive LS-Lasso estimators can have the oracle property even when $p \gg n$. Li et al. (2011) studied the asymptotic properties of penalized M-estimators defined using a convex loss function and proved that there exists a local minimum of the objective function used to define the estimators that is $\sqrt{n/p}$-consistent if $(p \log n)/n \to 0$ and that this local minimum has the oracle property when $p^2/n \to 0$.

An excellent review of the statistical properties of penalized regression estimators in the $p \gg n$ regime can be found in Buhlmann and van de Geer (2011). More recently, Loh (2015) studied the theoretical properties of penalized regression M-estimators in the $p \gg n$ regime. She showed that, under regularity assumptions, all stationary points of the optimization program used to define the estimators that are sufficiently close to $\boldsymbol{\beta}_{0,n}$ converge at the same rate as the LS-Lasso does in the case of sub-Gaussian errors. Furthermore, if the penalty function is an appropriately chosen non-convex function, such stationary points are equal to the oracle solution. These results, however, are not directly applicable to the estimators we study here.

In the second part of this thesis, which consists of Chapter 3, we study the asymptotic properties of more general versions of MM-Bridge and adaptive MM-Bridge estimators, more precisely, Bridge-type penalized M-estimators defined using a bounded loss function and an

estimate of scale, in linear models with a diverging number of parameters in the $p \ll n$ regime. We prove the consistency of MM-Bridge and adaptive MM-Bridge estimators under general conditions, assuming only $p/n \to 0$. Assuming the same hypothesis on the covariates that Portnoy (1984) assumes and that $(p \log n)/n \to 0$, we prove that MM-Bridge and adaptive MM-Bridge estimators are $\sqrt{n/p}$-consistent. These results include as a special case the consistency of M-estimators defined using a bounded loss function and an estimate of scale and hence prove the consistency of S and MM-estimators. We show that MM-Bridge estimators with $q < 1$ and adaptive MM-Bridge estimators with $t \leq 1$ actually are $\sqrt{n/k}$-consistent, where k is the number of non-zeros coefficients of the true regression parameter, and have the oracle property. The same technical arguments allow us to derive the asymptotic distribution of M-estimators defined using a bounded loss function and an estimate of scale, and hence that of S and MM-estimators; this result is stated without proof, since it is straightforward.

The rest of this thesis is organized as follows. In Chapter 2 we define the estimators we propose, we study their breakdown points, we study their asymptotic properties for the case of fixed p and random predictor variables and we propose an algorithm to compute them. In extensive simulations, we study the performance with regards to stability in the presence of high-leverage outliers, and prediction accuracy and variable selection properties for uncontaminated samples of the MM-Lasso and adaptive MM-Lasso estimators. We also apply our proposed estimators to a real high-dimensional data set. In Chapter 3 we study the asymptotic properties of Bridge-type penalized M-estimators defined using a bounded loss function and an estimate of scale, in linear models with a diverging number of parameters and fixed predictor variables. The proofs of all our results are given in Chapter 4, a technical appendix.

Chapter 2

Penalized MM-estimators

2.1 Framework

In this chapter, we consider a linear regression model with random predictor variables: we observe $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ $i = 1, ..., n$, i.i.d. $(p+1)$-dimensional vectors, where y_i is the response variable and $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of random predictor variables, satisfying

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + u_i \text{ for } i = 1, ..., n, \tag{2.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is to be estimated and u_i is independent of $\mathbf{x}_i$. Some of the coefficients of $\boldsymbol{\beta}_0$ may be zero, and thus the corresponding carriers do not provide relevant information to predict y. We do not know in advance the set of indices corresponding to coefficients that are zero, and it may be of interest to estimate it. For simplicity, we will assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}, \boldsymbol{\beta}_{0,II})$, where $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$, $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-k}$, all the coordinates of $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$ are non-zero and all the coordinates of $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-k}$ are zero.

Let F_0 be the distribution of the errors u_i, G_0 the distribution of the predictors $\mathbf{x}_i$ and H_0 the distribution of $(\mathbf{x}_i^{\mathrm{T}}, y_i)$. Then H_0 satisfies

$$H_0(\mathbf{x}, y) = G_0(\mathbf{x})F_0(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0). \tag{2.2}$$

Let $\mathbf{x}_I$ stand for the first k coordinates of $\mathbf{x}$.

2.2 S-Bridge, MM-Bridge and adaptive MM-Bridge estimators

Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, $\gamma_n \geq 0$, $r > 0$, a bounded ρ-function ρ_0 and $0 < b < 1$, we define the ℓ_r-penalized S-Bridge estimator following Maronna (2011) as

$$\hat{\boldsymbol{\beta}}_{PS} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} n\, s_n^2(\mathbf{r}(\boldsymbol{\beta})) + \gamma_n \|\boldsymbol{\beta}\|_r^r,$$

where $s_n()$ is the M-estimator of scale defined using ρ_0 and b. If the model contains an intercept, then it is not penalized.

It is easy to see that

$$\|\hat{\boldsymbol{\beta}}_{PS}\|_r^r \leq \|\hat{\boldsymbol{\beta}}_S\|_r^r, \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}_S$ is the S-estimator calculated using ρ_0 and b.

Given another bounded ρ-function ρ_1 that satisfies $\rho_1 \leq \rho_0$, $\lambda_n \geq 0$ and $q > 0$ we define the ℓ_q-penalized MM-Bridge estimator as

$$\hat{\boldsymbol{\beta}}_B = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\boldsymbol{\beta}\|_q^q, \tag{2.4}$$

where $\hat{\boldsymbol{\beta}}_1$ is a strongly consistent initial estimate of $\boldsymbol{\beta}_0$. Clearly, the robustness of the MM-Bridge estimator will depend heavily on the robustness of the initial estimate. If the model contains an intercept, then it is not penalized. For the case $q = 1$ we will call the resulting estimator MM-Lasso.

Note that our definition of an MM-Bridge estimator with $q = 2$ is not exactly the same as the definition of MM-Ridge estimators of Maronna (2011). For a given λ_n, the MM-Ridge of Maronna (2011) is equal to our MM-Bridge estimator calculated with $\lambda_n/s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))^2$ and $q = 2$. Nonetheless, our asymptotic results can be very easily adapted to cover the MM-Ridge estimators as defined by Maronna (2011). However, this is not the case for our results concerning the finite-sample breakdown point of MM-Bridge estimators. Note that, for any fixed $\lambda_n > 0$, by definition of $\hat{\boldsymbol{\beta}}_B$ we have that

$$\begin{aligned}
\lambda_n\|\hat{\boldsymbol{\beta}}_B\|_q^q &\leq \sum_{i=1}^n \rho_1\left(\frac{r_i(\hat{\boldsymbol{\beta}}_B)}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\hat{\boldsymbol{\beta}}_B\|_q^q \\
&\leq \sum_{i=1}^n \rho_1\left(\frac{r_i(\mathbf{0}_p)}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\mathbf{0}_p\|_q^q \\
&\leq n,
\end{aligned} \tag{2.5}$$

since $\rho_1 \leq 1$. Hence, $\|\hat{\boldsymbol{\beta}}_B\|_q^q \leq n/\lambda_n$. This immediately implies that for any fixed $\lambda_n > 0$, the breakdown point of $\hat{\boldsymbol{\beta}}_B$ is equal to 1. Using Maronna's parametrization, we get $\|\hat{\boldsymbol{\beta}}_B\|_q^q \leq (ns_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))^2)/\lambda_n$, which implies that the breakdown point of $\hat{\boldsymbol{\beta}}_B$ is great than or equal to the breakdown point of the residual scale $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))$, as Maronna (2011) points out.

Given $\varsigma > 0$, $t > 0$ and $\iota_n \geq 0$ we define the adaptive MM-Bridge estimator as

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j=1}^n \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \iota_n \sum_{j=1}^p \frac{|\beta_j|^t}{|\hat{\beta}_{2,j}|^\varsigma}, \tag{2.6}$$

where $\hat{\boldsymbol{\beta}}_2$ is a strongly consistent initial estimate of $\boldsymbol{\beta}_0$. Clearly if $\hat{\beta}_{2,j} = 0$ for some j, then $\hat{\beta}_{A,j} = 0$. If the model contains an intercept, then it is not penalized. For the case $t = 1$ we

will call the resulting estimator adaptive MM-Lasso. Note that for coefficients corresponding to large coefficients of $\hat{\boldsymbol{\beta}}_2$, the adaptive MM-Lasso employs a small penalty; this ameliorates the bias issues associated with the ℓ_1 penalty.

Wang et al. (2013) prove that their estimator can have the highest possible breakdown point among regression equivariant estimators, but it must be noted that their estimator is not regression equivariant. Alfons et al. (2013) show that the breakdown point of the Sparse-LTS estimator is $(n-h)/n$, where $n-h$ is the number of trimmed observations, and prove that the breakdown point of any ℓ_1-penalized M-estimator defined using a convex loss function is 0. In particular, the breakdown points of the LS-Lasso and the LAD-Lasso are 0. Note that it follows immediately from (2.3) that for any γ_n, the finite-sample breakdown point of $\hat{\boldsymbol{\beta}}_{PS}$ is at least as high as that of $\hat{\boldsymbol{\beta}}_S$. It follows from (2.5) that the breakdown point of $\hat{\boldsymbol{\beta}}_B$ is 1. We believe that this result hints at the possibility that the breakdown point may not be an adequate measure of robustness for penalized estimators. More generally, one could argue that the breakdown point is not an adequate measure of robustness for estimators that are not regression equivariant. See Davies and Gather (2006). Nonetheless, in Theorem 2.2.1 we prove that for any fixed $\iota_n > 0$, the breakdown point of $\hat{\boldsymbol{\beta}}_A$ is greater than or equal to the breakdown point of $\hat{\boldsymbol{\beta}}_2$.

Theorem 2.2.1. *If $\iota_n > 0$ is fixed, then $FBP(\hat{\boldsymbol{\beta}}_A) \geq FBP(\hat{\boldsymbol{\beta}}_2)$.*

Note that if $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_B$, then $FBP(\hat{\boldsymbol{\beta}}_A) = 1$ whenever $\lambda_n, \iota_n > 0$. In practice, γ_n, λ_n and ι_n may be chosen via some data-driven procedure such as cross-validation. In this case, the breakdown points of the resulting MM-Bridge and adaptive MM-Bridge estimators may be lower. The robustness of the resulting estimators will depend sorely on the robustness of the cross-validation scheme, and hence the use of robust residual scales as objective functions, instead of the classical root mean squared error, is crucial.

2.2.1 Asymptotics

We now describe the set-up to study the asymptotic properties of S-Bridge, MM-Bridge and adaptive MM-Bridge estimators. We will assume that

A1. a) ρ_0 and ρ_1 are twice continuously differentiable and eventually constant.
b) Let $\psi_1 = \rho_1'$. Then $\mathbb{E}\psi_1'(u/s_0) > 0$, where s_0 is as in (1.3).

A2. $\mathbb{P}\left(\mathbf{x}^T\boldsymbol{\beta} = 0\right) < 1 - b$ for all non-zero $\boldsymbol{\beta} \in \mathbb{R}^p$, where b was used to define the scale s_n.

A3. F_0 has a density, f_0, that is even, a monotone decreasing function of $|u|$ and a strictly decreasing function of $|u|$ in a neighbourhood of 0.

A family of ρ-functions that satisfies [A1]a) is Tukey's Bisquare family of functions. Condition [A2] is needed in the proofs of the consistency of the estimators. Note that condition [A3] does not require finite moments from F_0. Thus, extremely heavy tailed error distributions,

such as Cauchy's distribution, can be easily seen to satisfy [A3]. However, [A3] does impose a rather stringent symmetry assumption on the error distribution. This requirement greatly simplifies the asymptotic treatment of the estimators and is usual in robust statistics.

The following theorem proves the strong consistency of S-Bridge, MM-Bridge and adaptive MM-Bridge estimators whenever $\gamma_n = o(n)$, $\lambda_n = o(n)$ and $\iota_n = o(n)$ respectively.

Theorem 2.2.2. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies* (2.2)*. Assume [A1]-[A3] hold. Then*

(i) If $\gamma_n = o(n)$, $\hat{\boldsymbol{\beta}}_{PS} \overset{a.s.}{\to} \boldsymbol{\beta}_0$.

(ii) If $\lambda_n = o(n)$, $\hat{\boldsymbol{\beta}}_B \overset{a.s.}{\to} \boldsymbol{\beta}_0$.

(iii) If $\iota_n = o(n)$, $\hat{\boldsymbol{\beta}}_A \overset{a.s.}{\to} \boldsymbol{\beta}_0$.

In practice, we will use the S-Ridge estimator of Maronna (2011) as the initial estimate $\hat{\boldsymbol{\beta}}_1$ in (2.4) and (2.6). Note that according to Theorem 2.2.2 and the remarks above Theorem 2.2.1, the S-Ridge is a high breakdown point and consistent estimate of $\boldsymbol{\beta}_0$, as long as the penalty parameter satisfies $\gamma_n = o(n)$.

In order to obtain the rate of convergence of MM-Bridge and adaptive MM-Bridge estimators we will have to make the following additional assumption:

A4. G_0 has finite second moments and $\mathbf{V}_\mathbf{x} = \mathbb{E}\mathbf{x}\mathbf{x}^\mathrm{T}$ is non-singular.

In the next theorem, we prove the $\sqrt{n}$-consistency of MM-Bridge and adaptive MM-Bridge estimators.

Theorem 2.2.3. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies* (2.2)*. Assume [A1]-[A4] hold. Then*

(i) If $\lambda_n = O(\sqrt{n})$, then $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| = O_P(1/\sqrt{n})$.

(ii) If $\iota_n = O(\sqrt{n})$, then $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| = O_P(1/\sqrt{n})$.

Remark 1. *From now on, we will assume that the initial estimator used to define the penalty weights for the adaptive MM-Bridge estimator, $\hat{\boldsymbol{\beta}}_2$, is $\sqrt{n}$-consistent. For example, according to Theorem 2.2.3, we could take $\hat{\boldsymbol{\beta}}_2$ to be some MM-Bridge estimator calculated with $\lambda_n = O(\sqrt{n})$.*

Let $\hat{\boldsymbol{\beta}}_{A,I}$ stand for the first k coordinates of $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{A,II}$ for the remaining $p - k$. Let $\hat{\boldsymbol{\beta}}_{B,I}$ stand for the first k coordinates of $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{\beta}}_{B,II}$ for the remaining $p - k$. The following theorem shows that, as long as $\varsigma > t - 1$ and $t \leq 1$, adaptive MM-Bridge estimators have a sparsity property, and that if $q < 1$, then MM-Bridge estimators do so as well. In particular, taking $t = 1$, we prove that adaptive MM-Lasso estimators have a sparsity property.

Theorem 2.2.4. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies* (2.2)*. Assume [A1]-[A4] hold.*

(i) Suppose $q < 1$, $\lambda_n = O(\sqrt{n})$ and $\lambda_n / n^{q/2} \to \infty$. Then

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{B,II} = \mathbf{0}_{p-k}\right) \to 1.$$

.

(ii) Suppose $t \leq 1$, $\iota_n = O(\sqrt{n})$ and $\iota_n n^{(\varsigma - t)/2} \to \infty$. Then

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{A,II} = \mathbf{0}_{p-k}\right) \to 1.$$

Next we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{B,I}$ and $\hat{\boldsymbol{\beta}}_{A,I}$.

Theorem 2.2.5. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies* (2.2)*. Assume [A1]-[A4] hold.*

(i) Suppose $q < 1$, $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n / n^{q/2} \to \infty$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{B,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{d} N_k\left(\mathbf{0}, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \mathbf{V}_{\mathbf{x}_I}^{-1}\right).$$

(ii) Suppose $t \leq 1$, $\iota_n / \sqrt{n} \to 0$ and $\iota_n n^{(\varsigma - t)/2} \to \infty$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{d} N_k\left(\mathbf{0}, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \mathbf{V}_{\mathbf{x}_I}^{-1}\right).$$

Here s_0, $a(\psi)$ and $b(\psi)$ are as in (1.3)*,* (1.4) *and* (1.5)*, and $\mathbf{V}_{\mathbf{x}_I} = \mathbb{E}\mathbf{x}_I \mathbf{x}_I^T$.*

Theorem 2.2.4 together with Theorem 2.2.5 prove that $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ can have the *oracle property* as long as $\varsigma > t - 1$ and $t \leq 1$, and $q < 1$ respectively. That is: the estimated coefficients corresponding to null coordinates of the true regression parameter are set to zero with probability tending to 1, while at the same time the coefficients corresponding to non-null coordinates of the true regression parameter are estimated with the same asymptotic efficiency we would have had if we had applied an ordinary MM-estimator to the relevant predictor variables only.

In Theorem 2.2.6 we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_B$ for $q \geq 1$. Our theorem is analogous to Theorem 2 of Knight and Fu (2000).

Theorem 2.2.6. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies (2.2). Let $q \geq 1$. Assume [A1]-[A4] hold and $\lambda_n/\sqrt{n} \to \lambda_0$. Then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \xrightarrow{d} \arg\min(R),$$

where

$$R(\mathbf{z}) = -\mathbf{z}^T\mathbf{W} + \frac{1}{2s_0^2}b(\psi_1)\mathbf{z}^T\mathbf{V}_\mathbf{x}\mathbf{z} + \lambda_0 q\sum_{j=1}^{p} z_j sgn(\beta_{0,j})|\beta_{0,j}|^{q-1},$$

for $q > 1$,

$$R(\mathbf{z}) = -\mathbf{z}^T\mathbf{W} + \frac{1}{2s_0^2}b(\psi_1)\mathbf{z}^T\mathbf{V}_\mathbf{x}\mathbf{z} + \lambda_0\sum_{j=1}^{p}(z_j sgn(\beta_{0,j})I\{\beta_{0,j} \neq 0\} + |z_j|I\{\beta_{0,j} = 0\}),$$

for $q = 1$ and $\mathbf{W} \sim N_p(\mathbf{0}, (a(\psi_1)/s_0^2)\mathbf{V}_\mathbf{x})$. Here s_0, $a(\psi)$ and $b(\psi)$ are as in (1.3), (1.4) and (1.5)

Note that if $\lambda_0 = 0$, $\hat{\boldsymbol{\beta}}_B$ has the same asymptotic distribution as the corresponding ordinary MM-estimator. If $\lambda_0 > 0$ and $q = 1$, the asymptotic distributions of the coordinates of $\hat{\boldsymbol{\beta}}_B$ corresponding to null coefficients of $\boldsymbol{\beta}_0$ put positive probability at zero, the proof of this is essentially the same as the one that appears in pages 1361-1362 of Knight and Fu (2000). However, one can show that

$$\limsup \mathbb{P}\left(\hat{\boldsymbol{\beta}}_{B,II} = \mathbf{0}_{p-k}\right) \leq c < 1,$$

for some c. The proof is essentially the same as the proof of Proposition 1 of Zou (2006).

If $q > 1$ the amount of shrinkage of the estimated regression coefficients increases with the magnitude of the true regression coefficients. Hence, for "large" parameters, the bias introduced by MM-Bridge estimators with $q > 1$ may be unacceptably large, at least for the fixed p scenario. For the case $q = 2$ we can calculate the asymptotic distribution of the estimator explicitly. It follows easily from Theorem 2.2.6 that the asymptotic distribution of the MM-Ridge estimator is

$$N_p\left(-2\lambda_0\frac{s_0^2}{b(\psi_1)}\mathbf{V}_\mathbf{x}^{-1}\boldsymbol{\beta}_0, s_0^2\frac{a(\psi_1)}{b(\psi_1)^2}\mathbf{V}_\mathbf{x}^{-1}\right).$$

In the next theorem we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_B$ for $q < 1$ when $\lambda_n/n^{q/2} \to \lambda_0$.

Theorem 2.2.7. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies (2.2). Let $q < 1$. Assume [A1]-[A4] hold and $\lambda_n/n^{q/2} \to \lambda_0$. Then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \xrightarrow{d} \arg\min(R),$$

where

$$R(\mathbf{z}) = -\mathbf{z}^T\mathbf{W} + \frac{1}{2s_0^2}b(\psi_1)\mathbf{z}^T\mathbf{V}_\mathbf{x}\mathbf{z} + \lambda_0\sum_{j=1}^{p}|z_j|^q I\{\beta_{0,j}=0\},$$

and $\mathbf{W} \sim N_p\left(\mathbf{0}, (a(\psi_1)/s_0^2)\mathbf{V}_\mathbf{x}\right)$. *Here* s_0, $a(\psi)$ *and* $b(\psi)$ *are as in* (1.3), (1.4) *and* (1.5)

Using Theorem 2.2.7 one can show, see page 1361 of Knight and Fu (2000), that for $q < 1$ and $\lambda_0 > 0$, the asymptotic distributions of the coordinates of $\hat{\boldsymbol{\beta}}_B$ corresponding to null coefficients of $\boldsymbol{\beta}_0$ put positive probability at zero. Moreover, in this case the shrinkage only affects the coordinates of the estimators corresponding to null coefficients of $\boldsymbol{\beta}_0$, and hence no asymptotic bias is introduced.

2.2.2 Computation

In this section, we describe an algorithm to obtain approximate solutions of (2.4) for $q = 1$, i.e. MM-Lasso estimators. Through out this section we will assume that our model, (2.1), contains an intercept, and that the first coordinate of each $\mathbf{x}_i \in \mathbb{R}^{p+1}$ equals 1. Let $\mathbf{X}$ be the matrix with $\mathbf{x}_i$ as rows.

Prior to any calculations, all the columns of $\mathbf{X}$, except the first one, are centered and scaled using the median and the normalized median absolute deviation respectively. The response vector $\mathbf{y}$ is centered using the median. At the end, the final estimates are expressed in the original coordinates.

We take the S-Ridge estimator of Maronna (2011), which we note $\hat{\boldsymbol{\beta}}_{PS}$, as the initial estimate in (2.4). The penalty parameter for the S-Ridge estimator, γ_n, is chosen via robust 5-fold cross-validation, as described in Maronna (2011). Let $s_n = s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))$.

Let $w(u) = \psi_1(u)/u$, where ψ_1 is the derivative of ρ_1. Suppose λ_n is given and let $\hat{\boldsymbol{\beta}}_B$ be the MM-Lasso estimator. Let $\omega_i = w(r_i(\hat{\boldsymbol{\beta}}_B)/s_n)$. Let $\mathbf{W}$ be the diagonal matrix formed by $\sqrt{\omega_1}, ..., \sqrt{\omega_n}$. Let $\mathbf{y}^* = \mathbf{W}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{W}\mathbf{X}$. We approximate the solutions of (2.4) by the solutions of

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \frac{1}{2}\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\|^2 + s_n^2\lambda_n\sum_{j=2}^{p+1}|\beta_j|. \tag{2.7}$$

Note that the first column of $\mathbf{X}^*$ equals $\mathbf{k}^* = (\sqrt{\omega_1}, ..., \sqrt{\omega_n})$. For each $j = 2, ..., p+1$ let $\mathbf{x}^{*(j)}$ be the j-th column of $\mathbf{X}^*$ and let

$$\eta_j = \frac{\mathbf{k}^{*T}\mathbf{x}^{*(j)}}{\|\mathbf{k}^*\|^2}.$$

Then $\mathbf{x}^{*(j)}$ can be decomposed as the sum of two vectors: $\eta_j\mathbf{k}^*$, in the direction of $\mathbf{k}^*$, and $\mathbf{x}^{*\perp(j)} = \mathbf{x}^{*(j)} - \eta_j\mathbf{k}^*$, orthogonal to $\mathbf{k}^*$. Let $\mathbf{X}^{*\perp}$ be the matrix with columns $\mathbf{x}^{*\perp(2)}, ..., \mathbf{x}^{*\perp(p+1)}$.

It is easy to show that solutions of (2.7) satisfy

$$\mathbf{k}^*\mathbf{y}^* - \|\mathbf{k}^*\|^2(\beta_1 + \eta_2\beta_2 + ... + \eta_{p+1}\beta_{p+1}) = 0 \tag{2.8}$$

$$(\beta_2, \ldots, \beta_{p+1}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}^* - \mathbf{X}^{*\perp}\boldsymbol{\beta}\|^2 + s_n^2\lambda_n \sum_{j=1}^{p} |\beta_j| \tag{2.9}$$

The fact that $\mathbf{k}^*, \mathbf{y}^*$ and $\mathbf{X}^{*\perp T}$ actually depend on $\hat{\boldsymbol{\beta}}_B$ suggests an iterative procedure, as is usual in robust statistics. Starting from $\hat{\boldsymbol{\beta}}_{PS}$ we iteratively solve equation (2.9) using the LARS algorithm without including an intercept and then solve for the intercept in (2.8). Call $\boldsymbol{\beta}^{(i)}$ the estimate at the i-th iteration. Convergence is declared when

$$\frac{\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\|}{\|\boldsymbol{\beta}^{(i)}\|} \leq \delta,$$

where δ is some fixed tolerance parameter. In our simulations we took $\delta = 10^{-4}$.

Regarding the computation of adaptive MM-Lasso estimators, we note that solving (2.6) with $t = 1$ is equivalent to solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_1\left(\frac{y_i - \hat{\mathbf{x}}_i^T\boldsymbol{\beta}}{s_n}\right) + \iota_n \sum_{j=2}^{p+1} |\beta_j|,$$

where $\hat{\mathbf{x}}_{i,1} = 1$, $\hat{\mathbf{x}}_{i,j} = \mathbf{x}_{i,j}|\hat{\beta}_{2,j}|^\varsigma$ for $j = 2, ..., p+1$ and taking $\hat{\beta}_{A,1} = \hat{\beta}_1$ and $\hat{\beta}_{A,j} = \hat{\beta}_j|\hat{\beta}_{2,j}|^\varsigma$ for $j = 2, ..., p+1$. Hence, our procedure to compute MM-Lasso estimators can be used to compute adaptive MM-Lasso estimators, simply applying the routine to the data with weighted covariates. To compute our proposed adaptive MM-Lasso estimator, we take $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{PS}$, $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_B$ and $\varsigma = 1$.

In practice, we chose the ρ-functions used to compute the initial S-Ridge estimator, the MM-Lasso estimator and the adaptive MM-Lasso estimator of the form $\rho_0 = \rho_{c_0}^B$ and $\rho_1 = \rho_{c_1}^B$ where $c_1 \geq c_0$ and ρ_c^B is Tukey's Bisquare loss. The tuning constants c_0 and c_1 were chosen as in Maronna (2011).

The penalty parameter for $\hat{\boldsymbol{\beta}}_B$, λ_n, is chosen over a set of candidates via robust 5-fold cross validation, using a τ-scale of the residuals as the objective function. The τ-scale was introduced by Yohai and Zamar (1988) to measure in a robust and efficient way the largeness of the residuals in a regression model. The set of candidate lambdas is taken as 30 equally spaced points between 0 and λ_{max}, where λ_{max} is approximately the minimum λ such that all the coefficients of $\hat{\boldsymbol{\beta}}_B$ except the intercept are zero. To estimate λ_{max} we first robustly estimate the maximal correlation between $\mathbf{y}$ and the columns of $\mathbf{X}$ using bivariate winsorization as advocated by Khan et al. (2007). We use this estimate as an initial guess for λ_{max} and then improve it using a binary search. If $p > n$, then 0 is excluded from the candidate set. The penalty parameter for $\hat{\boldsymbol{\beta}}_A$, ι_n, is chosen using the same scheme used to choose λ_n.

The initial S-Ridge estimate is calculated using our own adaption of Maronna's MATLAB code to C++. To solve equation (2.9) we use the FastLasso() function from the `robustHD R` package (Alfons (2014)). We use the `foreach R` (Revolution Analytics and Weston (2013)) package for parallel computations when it comes to finding optimal penalty parameters via cross-validation. This provided a significant reduction in computing times in computers with several cores. Extensive parts of our computer code are written in C++ and interfaced with `R` using the `RcppArmadillo` package (Eddelbuettel and Sanderson (2014)). An `R` package that includes the functions to compute the estimators we propose is available at `http://esmucler.github.io/mmlasso/`.

2.3 Simulations

In this section, we compare the performance with regards to prediction accuracy and variable selection properties of

- The MM-Lasso estimator described in the previous section.
- The adaptive MM-Lasso (adaMM-Lasso) estimator described in the previous section.
- The MM Nonnegative Garrote (MM-NNG) estimator of Gijbels and Vrinssen (2015). The estimator was computed using `R` code provided by the authors.
- The ESL-Lasso estimator of Wang et al. (2013). The estimator was computed using MATLAB code provided by the authors.
- The Sparse-LTS. The penalty parameter for this estimator was chosen using a BIC-type criterion as advocated by the authors. The estimator was computed using the sparseLTS() function from the `robustHD R` package.
- The Wilcoxon-SCAD estimator (WW-SCAD). The estimator was computed using `R` code provided by the authors.
- The LAD-Lasso estimator. The penalty parameter was chosen using 5-fold cross validation using the median of the absolute value of the residuals as the objective function. We implemented this estimator in `R`.
- The LS-Lasso estimator (Lasso). The penalty parameter for this estimator was chosen using 5-fold cross validation using the sum of the squared residuals as the objective function. The estimator was computed using the lars() function from the `lars R` package.
- The adaptive LS-Lasso estimator (adaLasso). We used as an initial estimator an LS-Lasso estimator, calculated as above. Both the initial and the final penalty parameters were chosen using 5-fold cross validation using the sum of the squared residuals as the

objective function. The estimator was computed using the lars() function from the `lars` `R` package.

- The Maximum Likelihood Oracle estimator (Oracle), that is, the Maximum Likelihood estimator applied to the relevant carriers only. When the errors follow a normal distribution, this is the Least Squares estimators applied to the relevant carriers only. Note that in any case, this is not a feasible estimator, and is included for benchmarking purposes only.

- For the contaminated scenarios, we will also include the Oracle MM estimator: an MM-estimator, calculated with Tukey's bisquare function and tuned to have 85% normal efficiency, applied to the relevant carriers only. The estimator was computed using the lmRob() function from the `robust R` package. Once again, note that this is not a feasible estimator, and is included for benchmarking purposes only.

2.3.1 Scenarios

To evaluate the estimators we generate two independent samples of size n of the model $y = \mathbf{x}^T\boldsymbol{\beta}_0 + u$, plus an intercept that is equal to zero. The first sample, called the training sample, is used to fit the estimates and the second sample, called the testing sample, is used to evaluate the prediction accuracy of the estimates. We considered three possible distributions for the errors: a zero mean normal distribution, Student's t-distribution with three degrees of freedom ($t(3)$) and Student's t-distribution with one degree of freedom ($t(1)$). The first case corresponds to the classical scenario of normal errors, the second case has heavy-tailed errors and the third case has extremely heavy-tailed errors. For the first two cases we use the prediction root mean squared error (RMSE) to evaluate the prediction accuracy of the estimates. For the third case, since Student's t-distribution with one degree of freedom does not have a finite first moment, we use the median of the absolute value (MAD) of the prediction residuals as a measure of the the estimators prediction accuracy. We also evaluate the variable selection performance of the estimators by calculating the false negative ratio (FNR), that is, the fraction of coefficients erroneously set to zero, and the false positive ratio (FPR), the fraction of coefficients erroneously not set to zero.

We consider the following four scenarios for the sample size, the number of covariates, $\boldsymbol{\beta}_0$ and the distribution of the carriers.

1. We take $p = 8$, $n = 40$ and $\boldsymbol{\beta}_0$ given by: component 1 is 3, component 2 is 1.5, component 6 is 2 and the rest of the coordinates are set to zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 3$.

2. We take $p = 30$, $n = 100$ and $\boldsymbol{\beta}_0$ given by: components 1-5 are 2.5, components 6-10 are 1.5, components 11-15 are 0.5 and the rest are zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with

$\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.95$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 1.5$.

3. We take $p = 200$, $n = 100$ and $\boldsymbol{\beta}_0$ given by: components 1-5 are 2.5, components 6-10 are 1.5, components 11-15 are 0.5 and the rest are zero. The first 15 covariates $(x_1, ..., x_{15})$ and the remaining 185 covariates $(x_{16}, ..., x_{200})$ are independent. The first 15 covariates have a zero mean multivariate normal distribution. The pairwise correlation between the ith and jth components of $(x_1, ..., x_{15})$ is $\rho^{|i-j|}$ with $\rho = 0.95$ for $i, j = 1, ..., 15$. The final 185 covariates have a zero mean multivariate normal distribution. The pairwise correlation between the ith and jth components of $(x_{16}, ..., x_{200})$ is $\rho^{|i-j|}$ with $\rho = 0.95$ for $i, j = 16, ..., 200$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 1.5$.

4. We take $p = 250$, $n = 50$ and $\boldsymbol{\beta}_0$ given by: component 1 is 3, component 2 is 1.5, component 6 is 2 and the rest of the coordinates are set to zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 3$.

In Scenario 1 we have a moderately high p/n ratio. In Scenario 2 we have $p < n$ and high p/n ratio and in Scenarios 3 and 4 we have $p > n$. Scenario 1 was analyzed Tibshirani (1996) and Fan and Li (2001). Scenarios 2 and 3 were analysed in Huang et al. (2008). Since the MM-NNG, the ESL-Lasso and the Wilcoxon-SCAD can only be computed for $p < n$, we only compute them for Scenarios 1 and 2. The MM-NNG estimator could not be computed for Scenario 1, since the program kept repeatedly crashing.

To evaluate the robustness of the estimators for the case of high-leverage outliers, we introduce contaminations in all six scenarios, for the case of normal errors. Note that we only contaminate the training sample and not the testing sample. We take $m = [0.1n]$ and for $i = 1, .., m$ we set $y_i = 5y_0$ and $\mathbf{x}_i = (5, 0, \ldots, 0)$. We moved y_0 in an uniformly spaced grid between 0 and 3 with step 0.1 and then between 3 and 10 with step 1. To summarize the results for the contaminated scenarios we report for each estimator the maximum RMSE, FNR and FPR over all outlier sizes y_0.

The number of Montecarlo replications for the uncontaminated scenarios was $M = 500$. The number of Montecarlo replications for contaminated scenarios was reduced to $M = 100$, to keep computation times reasonably low.

2.3.2 Results

We now present the results of our simulation study. All results are rounded to two decimal places. Tables 2.1 and 2.2 show the results for Scenarios 1 through 4, without contamination.

Regarding the prediction accuracy of the estimators, for the case of normal errors, the MM-Lasso, the adaptive MM-Lasso and the LAD-Lasso have RMSEs of the same order as those of the Lasso and the adaptive Lasso, which show the best overall performance. In

Scenario 2, the RMSE of the MM-NNG is higher than those of the MM-Lasso and adaptive MM-Lasso but lower than that of the Sparse-LTS. The ESL-Lasso shows a good behaviour in Scenario 1, comparable to that of the Sparse-LTS, but the worst behaviour of all the estimators considered in Scenario 2. The Sparse-LTS shows a good behaviour in Scenarios 1 and 3, but its prediction errors are somewhat larger than those of the MM-Lasso and the adaptive MM-Lasso. In Scenarios 1 and 2 the WW-SCAD shows a good behaviour, with RMSEs similar to those of the MM-Lasso and adaptive MM-Lasso.

For the case of errors with t(3) or t(1) distribution, the MM-Lasso, the adaptive MM-Lasso and the LAD-Lasso show the best overall performance. We were surprised by the fact that for t(3) errors, the Lasso and the adaptive Lasso have a reasonably low RMSE when compared with the maximum likelihood oracle. As expected, the Lasso and the adaptive Lasso lose all predictive power when the errors have a t(1) distribution. In Scenario 2 the prediction errors of the MM-NNG are higher than those of the MM-Lasso and adaptive MM-Lasso but lower than those of the Sparse-LTS. The ESL-Lasso shows a good performance in Scenario 1, comparable to that of the Sparse-LTS, but the worst performance for Scenario 2. Except for Scenarios 2 and 4 the Sparse-LTS shows a reasonably good performance. In Scenarios 1 and 2, for the case of t(3) errors the WW-SCAD has RMSEs similar to those of the MM-Lasso and adaptive MM-Lasso, but for the case of t(1) errors its prediction errors are much larger.

Regarding the variable selection properties of the estimators, we note that the FPRs and the FNRs of the MM-Lasso are comparable to those of the Lasso, and the FPRs and FNRs of the adaptive MM-Lasso are comparable to those of the adaptive Lasso for the case of normal errors. The FPRs and the FNRs of the LAD-Lasso are similar to those of the MM-Lasso. For errors with t(3) or t(1) distribution, the MM-Lasso, the adaptive MM-Lasso and the LAD-Lasso generally show the best behaviour. The FPR of the adaptive MM-Lasso is lower than that of the MM-Lasso, but the price to pay for this improvement is an increase in the FNR. Note that for Scenarios 1 and 2 the Sparse-LTS has a rather high FPR, always greater than 0.5. In Scenario 2 the MM-NNG, the ESL-Lasso and the WW-SCAD have a rather high FNR for the three error distributions.

Table 2.1: Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenarios 1 and 2. RMSE, MAD, FNR and FPR, averaged over 500 replications are reported for each estimator.

Scenario	Normal			$t(3)$			$t(1)$		
	RMSE	FNR	FPR	RMSE	FNR	FPR	MAD	FNR	FPR
1									
MM-Lasso	3.42	0.04	0.52	1.77	0	0.52	1.36	0.01	0.50
adaMM-Lasso	3.43	0.09	0.27	1.75	0	0.20	1.32	0.02	0.21
ESL-Lasso	4.09	0.41	0.06	1.94	0.04	0.04	1.58	0.10	0.02
Sparse-LTS	3.92	0.03	0.82	1.91	0	0.85	1.44	0	0.69
WW-SCAD	3.46	0.25	0.10	1.75	0	0.11	1.97	0.37	0.04
LAD-Lasso	3.51	0.05	0.54	1.82	0	0.57	1.40	0	0.56
Lasso	3.33	0.02	0.43	1.84	0	0.46	9.9	0.38	0.28
adaLasso	3.28	0.06	0.26	1.82	0.01	0.29	10	0.46	0.19
Oracle	3.15	0	0	1.69	0	0	1.16	0	0
2									
MM-Lasso	1.69	0.13	0.21	1.75	0.10	0.27	1.28	0.15	0.17
adaMM-Lasso	1.77	0.26	0.09	1.80	0.21	0.09	1.39	0.29	0.06
MM-NNG	1.97	0.40	0.09	1.95	0.35	0.09	1.58	0.46	0.08
ESL-Lasso	9.73	0.71	0.14	9.97	0.68	0.15	6.92	0.71	0.12
Sparse-LTS	2.25	0	1	2.14	0	1	1.78	0.01	0.97
WW-SCAD	1.83	0.41	0.08	1.89	0.37	0.08	1.95	0.61	0.08
LAD-Lasso	1.73	0.13	0.25	1.78	0.10	0.23	1.32	0.15	0.23
Lasso	1.74	0.11	0.22	1.90	0.12	0.27	10.7	0.55	0.21
adaLasso	1.74	0.21	0.13	1.94	0.22	0.14	10.7	0.72	0.09
Oracle	1.63	0	0	1.73	0	0	1.27	0	0

Table 2.2: Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenarios 3 and 4. RMSE, MAD, FNR and FPR, averaged over 500 replications are reported for each estimator.

Scenario	Normal			$t(3)$			$t(1)$		
	RMSE	FNR	FPR	RMSE	FNR	FPR	MAD	FNR	FPR
3									
MM-Lasso	1.92	0.16	0.08	1.89	0.11	0.06	1.42	0.16	0.05
adaMM-Lasso	1.94	0.29	0.03	1.89	0.22	0.02	1.47	0.31	0.01
Sparse-LTS	1.89	0.13	0	1.98	0.11	0	1.47	0.15	0
LAD-Lasso	1.81	0.14	0.07	1.86	0.11	0.06	1.41	0.16	0.05
Lasso	1.88	0.11	0.21	2.12	0.12	0.22	6.33	0.57	0.10
adaLasso	2.06	0.18	0.13	2.29	0.19	0.14	6.75	0.70	0.10
Oracle	1.64	0	0	1.77	0	0	1.29	0	0
4									
MM-Lasso	4.05	0.12	0.07	1.99	0	0.06	2.05	0.08	0.05
adaMM-Lasso	3.99	0.18	0.03	1.80	0.01	0.01	1.79	0.12	0.02
Sparse-LTS	4.72	0.26	0.12	2.60	0.04	0.09	2.22	0.07	0.11
LAD-Lasso	4.02	0.09	0.10	2.03	0	0.10	1.96	0.06	0.08
Lasso	3.67	0.05	0.07	2.04	0.01	0.07	30.5	0.62	0.03
adaLasso	3.97	0.06	0.06	2.26	0.01	0.06	31.3	0.64	0.02
Oracle	3.13	0	0	1.67	0	0	1.12	0	0

In Tables 2.3 and 2.4 we show the results for Scenarios 1 through 4 under high-leverage contamination. The MM-Lasso and the adaptive MM-Lasso show the best overall behaviour. The ESL-Lasso shows a good behaviour in Scenario 1, but the worst behaviour by far in Scenario 2. The Sparse-LTS shows a good behaviour for Scenario 1 and the best behaviour for Scenario 3, but its maximum RMSEs are considerably larger than those of the MM-Lasso and the adaptive MM-Lasso for the rest of the scenarios. In Scenario 2 the MM-NNG shows a good performance. Note that for Scenarios 1 and 2 the maximum RMSEs of the WW-SCAD are considerably larger than those of the MM-Lasso and adaptive MM-Lasso. The RMSEs of the LAD-Lasso are also larger than those of the MM-Lasso and adaptive MM-Lasso. As expected, the maximum RMSEs of the Lasso and the adaptive Lasso are large in all cases. In Figure 2.1 we show the RMSEs of the estimators as a function of the outlier size for Scenario 1. The MM-Lasso has the overall best behaviour, followed closely by the adaptive MM-Lasso. Note that even though the maximum RMSE of the LAD-Lasso is in this case similar to that of the MM-Lasso, Figure 2.1 shows that the overall behaviour of the MM-Lasso is substantially better. Regarding the variable selection properties of the estimators, the MM-Lasso and the adaptive MM-Lasso show the best overall balance between a low FNR and a low FPR. Note that for Scenarios 1 and 2 the maximum FPR of the Sparse-LTS is very high.

Table 2.3: Results for Scenarios 1 and 2 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR	Max. FPR
1			
MM-Lasso	4.38	0.11	0.55
adaMM-Lasso	4.49	0.23	0.32
ESL-Lasso	4.87	0.60	0.19
Sparse-LTS	4.92	0.07	0.95
WW-SCAD	5.69	0.54	0.18
LAD-Lasso	4.61	0.41	0.52
Lasso	5.78	0.27	0.49
adaLasso	6.14	0.36	0.33
Oracle MM	3.71	0	0
2			
MM-Lasso	2.02	0.20	0.35
adaMM-Lasso	2.11	0.36	0.21
MM-NNG	2.33	0.48	0.24
ESL-Lasso	13.69	0.83	0.21
Sparse-LTS	3.18	0	1
WW-SCAD	3.36	0.55	0.21
LAD-Lasso	3.31	0.29	0.34
Lasso	3.05	0.25	0.26
adaLasso	3.24	0.41	0.15
Oracle MM	2.09	0	0

Table 2.4: Results for Scenarios 3 and 4 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR	Max. FPR
3			
MM-Lasso	2.48	0.21	0.15
adaMM-Lasso	2.72	0.37	0.05
Sparse-LTS	2.14	0.22	0
LAD-Lasso	3.63	0.39	0.08
Lasso	20.25	0.64	0.15
adaLasso	13.03	0.79	0.06
Oracle MM	2.09	0	0
4			
MM-Lasso	4.97	0.36	0.08
adaMM-Lasso	5.08	0.45	0.04
Sparse-LTS	5.40	0.47	0.11
LAD-Lasso	5.13	0.45	0.10
Lasso	6.04	0.42	0.07
adaLasso	7.89	0.45	0.06
Oracle MM	3.68	0	0

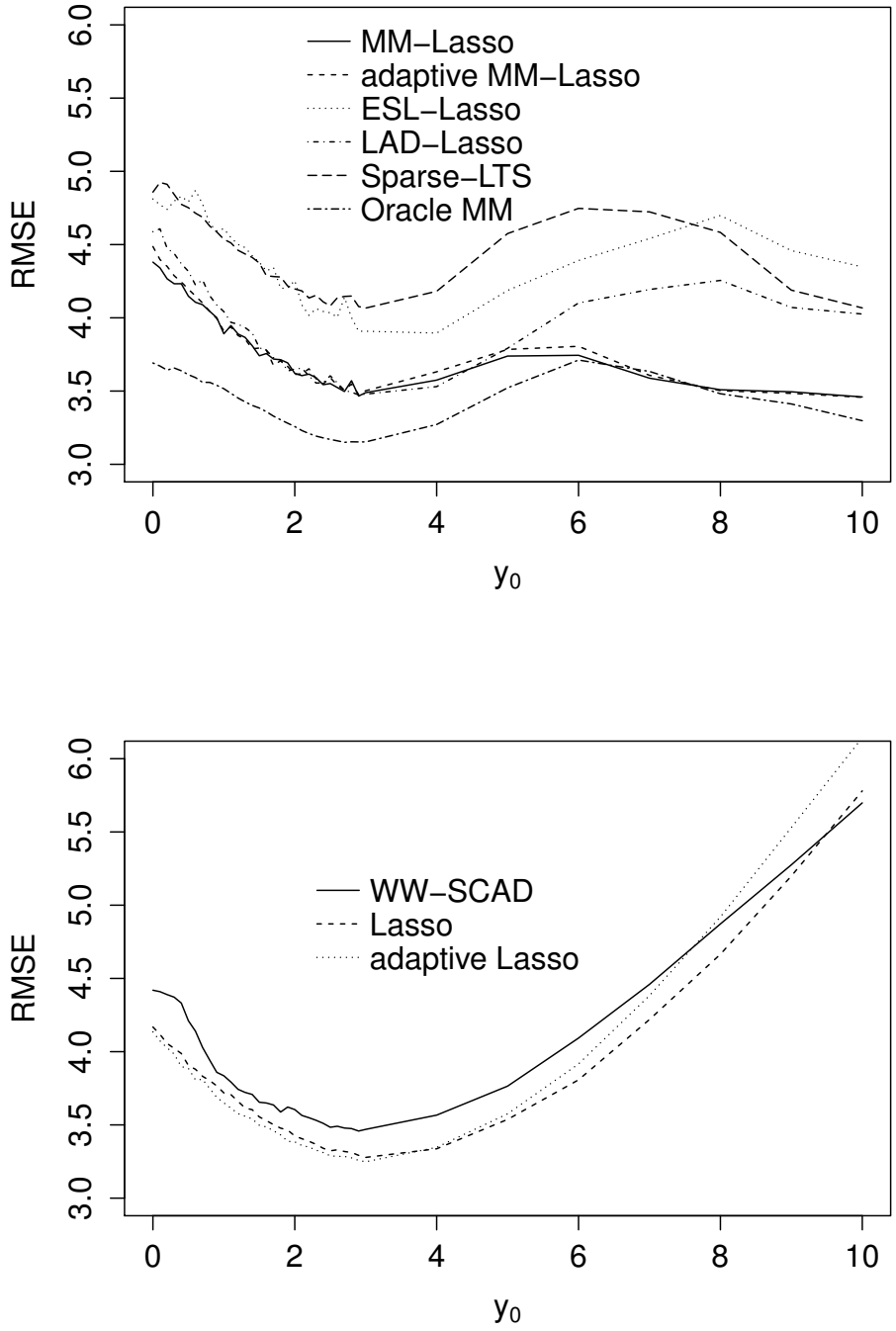

Figure 2.1: RMSEs as a function of outlier sizes for each of the estimators for the first scenario, with $p = 8$, $n = 40$, normal errors and 10% contamination. RMSEs are averaged over 100 replications.

Finally, we calculated the computing times of the estimators for the considered scenarios, for the case of normal errors and no contamination. Since the computing times for the adaptive MM-Lasso and the MM-Lasso, and for Lasso and the adaptive Lasso were very similar we only report the results for the adaptive MM-Lasso and the adaptive Lasso. Computing times were averaged over 100 replications and calculations were performed on a 3.07x4 GHz Intel Core i7 PC. Results are shown in Table 2.5. It is clear that the adaptive Lasso is orders of magnitude faster than the other estimators. The Sparse-LTS is generally faster than the adaptive MM-Lasso, except for Scenario 4, where the adaptive MM-Lasso is almost three times faster.

Table 2.5: Computing times in seconds for the estimators, averaged over 100 replications.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
adaMM-Lasso	3.46	7.58	44.59	9.05
MM-NNG	-	22.04	-	-
ESL-Lasso	0.30	0.93	-	-
Sparse-LTS	0.78	1.77	31.12	27.35
WW-SCAD	0.27	9.73	-	-
LAD-Lasso	2.22	2.37	10.31	10.25
adaLasso	0.05	0.14	0.99	0.30

2.4 A real high-dimensional data set

In this section, we analyse a data set corresponding to electron-probe X-ray microanalysis of archaeological glass vessels, where each of $n = 180$ glass vessels is represented by a spectrum on 1920 frequencies. For each vessel the contents of thirteen chemical compounds are registered. This data set appears in Janssens et al. (1998), and was previously analysed in Maronna (2011). We fit a linear model where the response variable is the content of the $13th$ chemical compound (PbO) and the carriers are the 1920 frequencies measures on each glass vessel. Since for frequencies below 15 and above 500 the values of x_{ij} are almost null and show very little variability, we keep frequencies 15 to 500, so that we have $p = 486$. We apply the MM-Lasso, the adaptive MM-Lasso, the Sparse-LTS, the LAD-Lasso, the Lasso and the adaptive Lasso estimators to the data.

The MM-Lasso selects seven variables. The adaptive MM-Lasso drops three of the variables selected by the MM-Lasso. The Sparse-LTS selects three variables. The LAD-Lasso selects five variables. The Lasso selects seventy one variables, the adaptive Lasso selects forty nine.

To asses the prediction accuracy of the estimators, we used 5-fold cross-validation. The criterion used was a τ-scale of the residuals, calculated as in Maronna and Zamar (2002). Results are shown in Table 2.6. The adaptive MM-Lasso and the Lasso show the best behaviour, followed by the LAD-Lasso, the Lasso, the adaptive Lasso and the Sparse-LTS, in that order.

	τ-scale
MM-Lasso	0.086
adaMM-Lasso	0.083
Sparse-LTS	0.329
LAD-Lasso	0.094
Lasso	0.131
adaLasso	0.138

Table 2.6: Cross-validated τ-scale of the residuals of each of the estimators for the electron-probe X-ray microanalysis data.

2.5 Resumen del Capítulo 2

En este capítulo proponemos estimadores de regresión robustos para modelos ralos y de alta dimensión. Estos están basados en agregar una penalidad tipo Bridge a los MM-estimatores de Yohai (1987).

En la Sección 2.2 damos las deficiones necesarias.

Dada una muestra $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, $\gamma_n \geq 0$, $r > 0$, una ρ-función acotada ρ_0 y $0 < b < 1$, definimos el estimador S-Bridge como

$$\hat{\boldsymbol{\beta}}_{PS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} n\, s_n^2(\mathbf{r}(\boldsymbol{\beta})) + \gamma_n \|\boldsymbol{\beta}\|_r^r,$$

donde $s_n()$ es el M-estimador de escala definido utilizando ρ_0 y b. Usaremos un estimador S-Bridge con $r = 2$ (S-Ridge, Maronna (2011)) como estimador inicial para computar los estimadores que proponemos.

Dada otra ρ-función ρ_1 que cumple $\rho_1 \leq \rho_0$, $\lambda_n \geq 0$ y $q > 0$ definimos el estimador MM-Bridge como

$$\hat{\boldsymbol{\beta}}_B = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))} \right) + \lambda_n \|\boldsymbol{\beta}\|_q^q,$$

donde $\hat{\boldsymbol{\beta}}_1$ es un estimador fuertemente consistente de $\boldsymbol{\beta}_0$. Cuando $q = 1$, llamaremos al estimador resultante MM-Lasso.

Por último, dados $\varsigma > 0$, $t > 0$ y $\iota_n \geq 0$ definimos el estimador MM-Bridge adaptivo como

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j=1}^{n} \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))} \right) + \iota_n \sum_{j=1}^{p} \frac{|\beta_j|^t}{|\hat{\beta}_{2,j}|^\varsigma},$$

donde $\hat{\boldsymbol{\beta}}_2$ es un estimador fuertemente consistente de $\boldsymbol{\beta}_0$. Cuando $t = 1$, llamaremos al estimador resultante MM-Lasso adaptivo.

Mostramos que el punto de ruptura del MM-Bridge es 1 para todo $\lambda_n > 0$ fijo y que el punto de ruptura del MM-Bridge adaptivo es mayor o igual que el de $\hat{\boldsymbol{\beta}}_2$ para todo $\iota_n > 0$ fijo. Sin embargo, como estos estimadores no son equivariantes por transformaciones de regresión creemos que estos resultados son un tanto engañosos y que posiblemente el punto de ruptura no sea una medida adecuada de robustez en este caso.

En la subsección 2.2.1 estudiamos las propiedades asintóticas de los estimadores en modelos de regresión lineal con un número fijo de variables predictivas aleatorias. En el Teorema 2.2.2 probamos su consistencia fuerte y en el Teorema 2.2.3 probamos que convergen con tasa $\sqrt{n}$. Los Teoremas 2.2.4 y 2.2.5 muestran que los estimadores MM-Bridge con $q < 1$ y los estimadores MM-Bridge adaptivos con $t \leq 1$ pueden tener la propiedad oráculo. En los Teoremas 2.2.6 y 2.2.7 derivamos al distribución asintótica de los estimadores MM-Bridge para todo $q > 0$.

En la Sección 2.2.2 proponemos un algoritmo para computar a los estimadores MM-Lasso y MM-Lasso adaptivo. El algoritmo usa como punto inicial el estimador S-Ridge de Maronna (2011) y usando el algoritmo LARS resuelve un problema de tipo Lasso con pesos iterativamente. Los párametros de penalización se eligen por validación cruzada robusta.

En la Sección 2.3 realizamos un extenso estudio de simulación. En diversos escenarios comparamos el rendimiento de los estimadores MM-Lasso y MM-Lasso adaptivo con el de varios competidores. Los resultados indican que el MM-Lasso y el MM-Lasso adaptivo dan el mejor balance entre estabilidad en la presencia de observaciones atípicas y precisión en las predicciones para muestras que siguen el modelo supuesto.

Por último en la Sección 2.4 aplicamos varios de los estimadores comparados en la Sección 2.3 a un conjunto de datos reales que surge de un problema quimeométrico. Comparamos su rendimiento calculando un error de predicción robusto por validación cruzada. Los estimadores MM-Lasso y MM-Lasso adaptivo dan los mejores resultados.

Chapter 3

Asymptotics for penalized M-estimators defined using a bounded loss function in linear models with a diverging number of parameters

3.1 Definitions and assumptions

In this chapter, we study the asymptotic properties of slightly more general versions of the estimators introduced in Chapter 2, in linear models with a diverging number of parameters and fixed predictor variables. We begin by fixing the notation to be used. Furthermore, we will discuss the assumptions needed to prove our results and compare them with those previously considered in the literature.

We consider a sequence of regression models

$$y_{i,n} = \mathbf{x}_{i,n}^T \boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n \tag{3.1}$$

where $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ is to be estimated and $u_{i,n}$ are i.i.d. random variables defined in a common probability space with distribution function F_0. From now on, we will drop the n subscript from $y_{i,n}, \mathbf{x}_{i,n}$, $\boldsymbol{\beta}_{0,n}$, p_n and $u_{i,n}$. To unburden the notation, we will assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}, \boldsymbol{\beta}_{0,II})$, where $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$, $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-k}$, all the coordinates of $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^k$ are non-zero and all the coordinates of $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-k}$ are zero. Note that k may also depend on n, that is, $k = k_n$.

For the sake of generality we will make some minor modifications to the definitions and notation of Chapter 2.

Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, $\gamma_n \geq 0$, $r > 0$, a bounded ρ-function ρ_0 and $0 < b < 1$, the ℓ_r-penalized S-Bridge estimator is defined as

$$\hat{\boldsymbol{\beta}}_{PS} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} n\, s_n^S(\mathbf{r}(\boldsymbol{\beta}))^2 + \gamma_n \|\boldsymbol{\beta}\|_r^r,$$

where $s_n^S(\mathbf{r}(\boldsymbol{\beta}))$ is the M-estimate of scale of the residuals defined using ρ_0 and b.

We will work with a slightly more general definition of MM-Bridge and adaptive MM-Bridge estimators than the one given in Chapter 2. Given a bounded ρ-function ρ_1, $\lambda_n \geq 0$ and $q > 0$ the ℓ_q-penalized MM-Bridge estimator is defined as

$$\hat{\boldsymbol{\beta}}_B = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) + \lambda_n \|\boldsymbol{\beta}\|_q^q,$$

where $s_n > 0$ is some positive random variable defined in the same probability space as the random errors u_i.

Given $\varsigma > 0$, $t > 0$ and $\iota_n \geq 0$ the adaptive MM-Bridge estimator is defined as

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) + \iota_n \sum_{j=1}^{p} \frac{|\beta_j|^t}{|\hat{\beta}_{ini,j}|^\varsigma},$$

where $\hat{\boldsymbol{\beta}}_{ini}$ is a consistent initial estimate of $\boldsymbol{\beta}_0$, i.e. $\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

Throughout this chapter we will assume that s_n, the random variable used to standardize the residuals in the definitions of MM-Bridge and adaptive MM-Bridge estimators, converges in probability to some value $s_0 > 0$. The constant s_0 need not necessarily be the same as the one in (1.3). For example, according to Lemma 3.2.1 and the comments following it, one may take $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))$ or $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$, where $\hat{\boldsymbol{\beta}}_S$ is the corresponding ordinary, i.e. non-penalized, S-estimator.

Having lifted the requirement on the form of the estimate of scale used in the definition of the MM-Bridge and adaptive MM-Bridge estimators, see Section 2.2, $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ are now only penalized M-estimators defined using a bounded loss function and an estimate of scale. Consider a regression M-estimator defined using a bounded loss function and an estimate of scale. This is an MM-Bridge estimator with $\lambda_n \equiv 0$. Hence, Theorems 3.2.2, 3.2.3 and 3.2.6 prove its consistency and its asymptotic normality. In particular, taking $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$, $\rho_1 = \rho_0$ for the case of S-estimators and ρ_1 that satisfies $\rho_1 \leq \rho_0$ for the case of MM-estimators, Theorems 3.2.2, 3.2.3 and 3.2.6 prove the consistency and asymptotic normality of ordinary S and MM-estimators.

We introduce some notation. Let $\mathbf{x}_I$ be the first k coordinates of $\mathbf{x}$, and $\mathbf{x}_{II}$ be the last $p-k$ coordinates of $\mathbf{x}$. Let $\rho_{1,n}$ and $\rho_{2,n}$ stand for the smallest and largest eigenvalues of $\boldsymbol{\Sigma}_n = 1/n \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$. Let $\boldsymbol{\Sigma}_{1,n} = 1/n \sum_{i=1}^{n} \mathbf{x}_{i,I} \mathbf{x}_{i,I}^T$. Note that the largest eigenvalue of $\boldsymbol{\Sigma}_{1,n}$ is smaller than or equal to $\rho_{2,n}$, while the smallest eigenvalue of $\boldsymbol{\Sigma}_{1,n}$ is larger than or equal to $\rho_{1,n}$. Throughout this chapter we will assume that $\boldsymbol{\Sigma}_n$ is non-singular for all n.

For $0 < \alpha < 1$, let

$$\eta_n(\alpha) = \min_{\mathcal{A}\subset\{1,\ldots,n\},\#\mathcal{A}=[n\alpha]} \min_{\|\theta\|=1} \max_{i\in\mathcal{A}} |\mathbf{x}_i^T \theta|. \tag{3.2}$$

The function $\eta_n(\alpha)$ was introduced in Davies (1990). It measures in some sense the worst possible conditioning of any subset of size $[n\alpha]$ of the carriers. To prove the consistency of the estimators, we will need to assume that for some $0 < \alpha < 1$, $\liminf \eta_n(\alpha) > 0$.

Suppose that $\mathbf{x}_i$, $i = 1, \ldots, n$ are independent and identically distributed random vectors in $\mathbb{R}^p$ such that there exists η_1, η_2 with $0 < \eta_1, \eta_2 < 1$ such that, for all n

$$\sup_{\|\boldsymbol{\theta}\|=1} \mathbb{P}\left(|\mathbf{x}^T\boldsymbol{\theta}| < \eta_1\right) < 1 - \eta_2.$$

This holds for example if $\mathbf{x}_i \sim N_p(\mathbf{0}, \mathbf{M}_n)$ and there exists some $\kappa > 0$ such that the smallest eigenvalue of $\mathbf{M}_n$ is bounded below by κ for all n. It can be shown, using maximal inequalities such as those of Theorem 4.2.1, that if $p/n \to 0$

$$\sup_{\|\boldsymbol{\theta}\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} I\left\{|\mathbf{x}_i^T\boldsymbol{\theta}| < \eta_1\right\} - \mathbb{P}\left(|\mathbf{x}^T\boldsymbol{\theta}| < \eta_1\right) \right| \xrightarrow{P} 0.$$

Hence, with arbitrarily high probability, for large enough n,

$$\sup_{\|\boldsymbol{\theta}\|=1} \frac{1}{n} \sum_{i=1}^{n} I\left\{|\mathbf{x}_i^T\boldsymbol{\theta}| < \eta_1\right\} < \sup_{\|\boldsymbol{\theta}\|=1} \mathbb{P}\left(|\mathbf{x}^T\boldsymbol{\theta}| < \eta_1\right) + \eta_2/2 < 1 - \eta_2/2.$$

In this case, for any α such that $1 - \eta_2/2 < \alpha < 1$, for large enough n it follows that for all $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$ and all subsets $\mathcal{A}$ of $\{1, \ldots, n\}$ with $\#\mathcal{A} = [n\alpha]$ there exists $i \in \mathcal{A}$ such that $|\mathbf{x}_i^T\boldsymbol{\theta}| \geq \eta_1$, which implies $\eta_n(\alpha) \geq \eta_1$.

For $\mathcal{A} \subset \{1, ..., n\}, \#\mathcal{A} = [n\alpha]$ let

$$\Sigma(\mathcal{A}) = \frac{1}{[n\alpha]} \sum_{i \in \mathcal{A}} \mathbf{x}_i \mathbf{x}_i^T.$$

Let $\rho_{1,n}(\mathcal{A})$ be the smallest eigenvalue of $\Sigma(\mathcal{A})$. Take $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$. Then

$$\boldsymbol{\theta}^T \Sigma(\mathcal{A}) \boldsymbol{\theta} \leq \max_{i \in \mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2.$$

Hence

$$\rho_{1,n}(\mathcal{A}) \leq \min_{\|\boldsymbol{\theta}\|=1} \max_{i \in \mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2$$

which implies that

$$\min_{\mathcal{A} \subset \{1,\ldots,n\}, \#\mathcal{A}=[n\alpha]} \rho_{1,n}(\mathcal{A}) \leq \eta_n(\alpha)^2.$$

Hence, $\liminf \eta_n(\alpha) > 0$ holds if the smallest eigenvalues of the covariance matrices formed from any subsample of size $[n\alpha]$ are uniformly bounded away from zero. The following Lemma, an adaptation of Lemma 3 of Davies (1990), gives necesary conditions for $\liminf \eta_n(\alpha) > 0$ to hold.

Lemma 3.1.1. *Assume there exists a constant $M > 0$ such that $1/n \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \leq pM$ for all n. Then if $\liminf \eta_n(\alpha) > 0$ for some $0 < \alpha < 1$ there exists positive numbers η_1, η_2 and n_0 such that*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T I\{\|\mathbf{x}_i\| < \eta_1 \sqrt{p}\} - \eta_2 \mathbf{I}_p$$

is positive definite for all $n \geq n_0$.

See also Examples 1, 2 and 3 of Davies (1990).

For $\mathbf{z} \in \mathbb{R}^p$ and $c > 0$ let

$$I(\mathbf{z}, c) = \left\{ i = 1, \ldots, n : \left| \mathbf{x}_i^T \mathbf{z} \right| \leq c \right\},$$

let $\mathcal{B}(\delta)$ be the ball in $\mathbb{R}^p$ centered at zero with radius δ and let $\mathcal{S}^*$ be the sphere centered at zero with radius 1.

We will need the following assumptions:

R0. ρ_0 is a bounded ρ-function and, for some $m > 0$, $\rho_0(u) = 1$ if $|u| \geq m$.

R1. ρ_1 is a continuously differentiable, bounded ρ-function. Let ψ_1 be the derivative of ρ_1. Then $\psi_1(t)$ and $t\psi_1(t)$ are bounded.

R2. ρ_1 is a three times continuously differentiable, bounded ρ-function. Let ψ_1 be the derivative of ρ_1. Then $\psi_1(t), \psi_1'(t), \psi_1''(t), t\psi_1(t), t\psi_1'(t)$ and $t\psi_1''(t)$ are bounded. Also, $\mathbb{E}\psi_1'(u/s_0) > 0$, where s_0 is the limit in probability of s_n.

F0. F_0 has a density, f_0, that is even, a monotone decreasing function of $|u|$ and a strictly decreasing function of $|u|$ in a neighbourhood of 0.

B0. $\gamma_n \|\boldsymbol{\beta}_0\|_r^r / n \to 0$.

B1. $\lambda_n \|\boldsymbol{\beta}_0\|_q^q / n \to 0$.

B2. $\iota_n \|\boldsymbol{\beta}_0\|_t^t / n \to 0$.

B3. There exists $b_1, b_2 > 0$ such that for all n

$$b_1 \leq \min_{1 \leq j \leq k} |\beta_{0,j}| \leq \max_{1 \leq j \leq k} |\beta_{0,j}| \leq b_2.$$

X0. $p < [n(1-b)]$ for all n.

X1. a) There exists a constant $M > 0$ such that $1/n \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \leq pM$ and $1/n \sum_{i=1}^{n} \|\mathbf{x}_{i,I}\|^2 \leq kM$ for all n.

b) There exists a constant $B > 0$ such that $\max_{i \leq n} \|\mathbf{x}_i\| \leq Bn$ for all n.

X2. $\tau = \sup_n \rho_{2,n} < \infty$.

X3. For some $0 < \alpha < 1$, $\liminf \eta_n(\alpha) > 0$.

X4. For any $c > 0$ there are constants $a > 0$, $\delta > 0$ and $C > 0$ such that for all $\boldsymbol{\beta} \in \mathcal{B}(\delta)$, $\mathbf{z} \in \mathcal{S}^*$, and n

$$\sum_{i \in J} \left(\mathbf{x}_i^T \mathbf{z}\right)^2 \geq an,$$

where $J = I(\boldsymbol{\beta}, c) \cap I(\mathbf{z}, C)$.

X5. For any $c > 0$ and $\varepsilon > 0$ there are constants $\delta^{'} > 0$ and $C > 0$ such that for all $\boldsymbol{\beta} \in \mathcal{B}(\delta^{'})$, $\mathbf{z} \in \mathcal{S}^*$, and n

$$\sum_{i \notin J} \left(\mathbf{x}_i^T \mathbf{z}\right)^2 \leq \varepsilon n,$$

where $J = I(\boldsymbol{\beta}, c) \cap I(\mathbf{z}, C)$.

X6. $\lambda_n \sqrt{k/n} \to 0$

X7. $\iota_n \sqrt{k/n} \to 0$

X8. $\lambda_n n^{-q/2} / p^{1-q/2} \to \infty$

X9. $\iota_n n^{(\varsigma - t)/2} p^{((t-\varsigma)/2)-1} \to \infty$

X10. $\max_{i \leq n} \|\mathbf{x}_{i,I}\|^2 = o(n/k^2)$.

Conditions [R0] and [R1] are satisfied by, for example, Tukey's Bisquare loss function. Condition [R2] is a strengthening of condition [R1]. It is satisfied by, for example, $\rho(x) = 1 - \exp(-x^2)$, used in Wang et al. (2013) to define the ESL-Lasso estimator, and $\rho(x) = 1 - (1 - x^2)^4 I\{|x| \leq 1\}$, which is similar to Tukey's Bisquare loss.

Condition [F0] is condition [A3] of Chapter 2.

Conditions [B0], [B1] and [B2] are typical in the asymptotic analysis of penalized regression estimators. Note that they are trivially satisfied by ordinary MM and S estimators. [B3] is needed to obtain the rate of consistency of the estimators and appears in Huang et al. (2008). Assuming [B3], [B0], [B1] and [B2] simply require that $\gamma_n = o(n/k)$, $\lambda_n = o(n/k)$ and $\iota_n = o(n/k)$ respectively.

Condition [X0] is needed in the proof of the consistency of the scale estimate provided by the S-Bridge estimator. To prove the consistency of the regression estimators we will need $p/n \to 0$. To obtain the rate of consistency of the estimators we will need $(p \log n)/n \to 0$. Note that $(p \log n)/n \to 0$ is no stronger than $(p \log p)/n \to 0$, paraphrasing Portnoy (1984): if $p \leq \sqrt{n}$, $(p \log n)/n \leq (\log n)/\sqrt{n} \to 0$; while if $p \geq \sqrt{n}$, $(p \log n)/n \leq (2p \log p)/n$.

[X1] a) holds when the covariates are standardized. [X1] b) appears in Portnoy (1984) and holds, for example, if all the covariates are bounded and $p/n^2 \to 0$. On the other hand, suppose that $\mathbf{x}_i$, $i = 1, \ldots, n$ are independent and identically distributed random vectors in $\mathbb{R}^p$ such that for some C, $\mathbb{E}x_{i,j}^2 \leq C$ for all i, j and n. Then, [X1] holds in probability if $p/n \to 0$; see Section 4 of Portnoy (1984). [X2] appears in, for example, Portnoy (1985), Welsh (1989), Zou and Zhang (2009) and Li et al. (2011). See also Bai and Wu (1994). Note that if [X1] and [X3] hold, by Lemma 3.1.1 we have that $\inf_n \rho_{1,n} > 0$. See the comments following (3.2) for a more thorough discussion of [X3].

[X4] and [X5] were introduced in Portnoy (1984) where they appear as X1 and X2. He shows that these conditions hold in probability if the covariates are sampled from an appropriate distribution in $\mathbb{R}^p$, such as a scale mixture of standard multivariate normals, and $(p \log n)/n \to 0$. If [R2], [F0], [X1], [X4], [X5] and $(p \log n)/n \to 0$ hold then Lemma 3.1 of Portnoy (1984) holds. In Lemma 4.2.5 we prove a simple modification of Portnoy's lemma that is needed in the proof of the rate of convergence of the estimators. The aforementioned lemma shows that, very loosely speaking, the objective functions used to defined the MM-Bridge and adaptive MM-Bridge estimators are convex in a neighbourhood of the true regression parameter with probability tending to one. [X4] and [X5] are not needed if p is fixed, in this case they can be replaced by, for example, $\max_{i \leq n} \|\mathbf{x}_i\| = O(1)$.

[X6] and [X8], and [X7] and [X9] are needed in the proof of the oracle property of MM-Bridge and adaptive MM-Bridge estimators respectively. They appear in Huang et al. (2008). If k is fixed and $\lambda_n = n^{(1-\delta)/2}$ for some $0 < \delta < 1$ then [X6] and [X8] hold if $p^{2-q}/n^{1-\delta-q} \to 0$. If k is fixed, $\varsigma = 1$, $t = 1$ and $\iota_n = n^{(1-\delta)/2}$ for some $0 < \delta < 1$ then [X7] and [X9] hold if $p^2/n^{1-\delta} \to 0$. Note that the combination of [X6] and [X8] excludes the case $q = 1$. [X10] is needed in the proof of the oracle property. It holds, for example, if the first k covariates are bounded and $k^3/n \to 0$.

3.2 Results

In this section, we state all our results. Proofs are given in Chapter 4.

First, we prove the consistency of $s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))$.

Lemma 3.2.1. *Assume [R0], [F0], [B0] and [X0] hold. Assume that $p/n \to 0$. Assume also that f_0 is strictly decreasing on the non negative real numbers. Then, $s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS})) \xrightarrow{P} s(F_0)$, where $s(F_0)$ is the solution of $\mathbb{E}\rho_0(u/s) = b$.*

It is worth noting that in Theorem 3 of Davies (1990), the author proves the consistency of regression S-estimators and the corresponding scale estimates assuming $(p \log n)/n \to 0$ ($(k_n \log n)/n \to 0$ in his notation). This condition can be weakened to $p/n \to 0$. To do so simply replace any appeals in his proof of his Lemma 2 by appeals to our Lemma 4.2.2.

Theorem 3.2.2. *Assume [R1] and [F0] hold. Assume that $p/n \to 0$. Then, for any $0 < \alpha < 1$*

(i) If [B1] holds, $\eta_n(\alpha)\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

(ii) If [B2] and [B3] hold, $\eta_n(\alpha)\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

Note that Theorem 3.2.2 together with [X3] entails that $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{\beta}}_A$ are consistent. In the following theorem, we derive their rate of convergence.

Theorem 3.2.3. *Assume [R2], [F0], [B1]-[B3] and [X1]-[X5] hold. Assume* $(p \log n)/n \to 0$. *Then*

(i) $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| = O_P(\sqrt{p/n} + \lambda_n\sqrt{k}/n)$

(ii) $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| = O_P(\sqrt{p/n} + \iota_n\sqrt{k}/n)$

Note that under the assumptions of Theorem 3.2.3, if in addition [X6] and [X7] hold, we have that $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| = O_P(\sqrt{p/n})$ and $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| = O_P(\sqrt{p/n})$. For the case $\lambda_n \equiv 0$ hypothesis [B3] is not needed. In this case, it follows from Theorem 3.2.3 that S and MM-estimators are $\sqrt{n/p}$-consistent. If we further assume that $\max_{i\leq n} \|\mathbf{x}_i\|^2 = o(n/p)$, we can apply Theorem 2 of Mammen (1988) to obtain asymptotic expansions for S-estimators.

Remark 2. *We will henceforth assume that the initial estimator used to define the weights for the adaptive MM-Bridge,* $\hat{\boldsymbol{\beta}}_{ini}$, *is* $\sqrt{n/p}$*-consistent. According to Theorem 3.2.3, we could take, for example, an MM-Lasso estimator.*

Let $\hat{\boldsymbol{\beta}}_{A,I}$ stand for the first k coordinates of $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{A,II}$ for the remaining $p-k$. Let $\hat{\boldsymbol{\beta}}_{B,I}$ stand for the first k coordinates of $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{\beta}}_{B,II}$ for the remaining $p-k$. In the following theorem, we prove a sparsity property of MM-Bridge estimators with $q < 1$ and adaptive MM-Bridge estimators with $t \leq 1$.

Theorem 3.2.4. *Assume [R2], [F0], [B1]-[B3] and [X1]-[X7] hold. Assume* $(p \log n)/n \to 0$. *Then*

(i) If [X8] holds and $q < 1$ *then* $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{B,II} = \mathbf{0}_{p-k}\right) \to 1$.

(ii) If [X9] holds and $t \leq 1$ *then* $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{A,II} = \mathbf{0}_{p-k}\right) \to 1$.

Remark 3. *Under the hypothesis of Theorem 3.2.4. we have that with probability tending to one* $\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0 = (\hat{\boldsymbol{\beta}}_{B,I} - \boldsymbol{\beta}_{0,I}, \mathbf{0}_{p-k})$. *Using this fact, essentially the same proof of Theorem 3.2.3, using Lemma 4.2.4(ii) instead of (i), replacing* $\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0$ *by* $\hat{\boldsymbol{\beta}}_{B,I} - \boldsymbol{\beta}_{0,I}$, $\mathbf{x}$ *by* $\mathbf{x}_I$ *and* p *by* k *where appropriate, shows that actually* $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| = O_P(\sqrt{k/n})$. *The same holds for* $\hat{\boldsymbol{\beta}}_A$.

Next we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{B,I}$ and $\hat{\boldsymbol{\beta}}_{A,I}$.

Theorem 3.2.5. *Assume [R2], [F0], [B1]-[B3] and [X1]-[X10] hold. Assume* $(p \log n)/n \to 0$. *Let* $\mathbf{a}_n$ *be a vector in* $\mathbb{R}^k$ *satisfying* $\|\mathbf{a}_n\| = 1$. *Let* $r_n^2 = \mathbf{a}_n^T \boldsymbol{\Sigma}_{1,n}^{-1} \mathbf{a}_n$. *Then*

(i) If $q < 1$

$$\sqrt{n} r_n^{-1} \mathbf{a}_n^T \left(\hat{\boldsymbol{\beta}}_{B,I} - \boldsymbol{\beta}_{0,I} \right) \xrightarrow{d} N \left(0, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \right).$$

(ii) If $t \leq 1$

$$\sqrt{n} r_n^{-1} \mathbf{a}_n^T \left(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I} \right) \xrightarrow{d} N \left(0, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \right).$$

Here $a(\psi_1) = \mathbb{E}\psi_1^2 (u/s_0)$ *and* $b(\psi_1) = \mathbb{E}\psi_1' (u/s_0)$.

Consider a regression M-estimator defined using a bounded loss function and an estimate of scale. This is an MM-Bridge estimator with $\lambda_n \equiv 0$. Note that, assuming [R2], the objective function used to define it is everywhere differentiable. Hence, a proof entirely analogous to that of Theorem 3.2.5, making the obvious adjustments to Lemmas 4.2.6 and 4.2.7, allows us to derive the asymptotic distribution of ordinary M-estimators defined using a bounded loss function and an estimate of scale. In particular, we derive the asymptotic distribution of S and MM-estimators. We state the result without proof.

Theorem 3.2.6. *Assume [R2], [F0] and [X1]-[X5] hold. Assume* $(p \log n)/n \to 0$. *Assume* $\max_{i \leq n} \|\mathbf{x}_i\|^2 = o(n/p^2)$ *and* $\lambda_n \equiv 0$. *Let* $\mathbf{a}_n$ *be a vector in* $\mathbb{R}^p$ *satisfying* $\|\mathbf{a}_n\| = 1$. *Let* $r_n^2 = \mathbf{a}_n^T \boldsymbol{\Sigma}_n^{-1} \mathbf{a}_n$. *Then*

$$\sqrt{n} r_n^{-1} \mathbf{a}_n^T \left(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0 \right) \xrightarrow{d} N \left(0, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2} \right).$$

Here $a(\psi_1) = \mathbb{E}\psi_1^2 (u/s_0)$ *and* $b(\psi_1) = \mathbb{E}\psi_1' (u/s_0)$.

Note that $\max_{i \leq n} \|\mathbf{x}_i\|^2 = o(n/p^2)$ holds, for example, if the covariates are bounded and $p^3/n \to 0$. This is the rate of growth of p allowed by the asymptotic normality result of Huber (1973).

Theorem 3.2.4 together with Theorem 3.2.5 prove that $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ with $t \leq 1$ and $q < 1$ respectively can have the oracle property in the following sense: the estimated coefficients corresponding to null coordinates of the true regression parameter are set to zero with probability tending to 1, while at the same time any linear contrast of the coefficients corresponding to non-null coordinates of the true regression parameter is estimated with the same asymptotic efficiency we would have if we had applied the corresponding ordinary M-estimator to the relevant predictor variables only.

3.3 Resumen del Capítulo 3

En este capítulo estudiamos las propiedades asintóticas de versiones más generales de los estimadores introducidos en el Capítulo 2, ahora en modelos lineales con un número de variables predictoras fijas que diverge. En la Sección 3.1 definimos los estimadores y listamos las hipótesis sobre la naturaleza de los datos y la regularidad de las funciones de pérdida utilizadas para definir los estimadores que serán necesarias para obtener nuestros resultados. Además, comparamos estas hipótesis con aquellas que hayan sido utilizadas por otro autores en trabajos sobre el problema de M-estimadores de regresión en modelos lineales con un número de parámetros que diverge. En la Sección 3.2 enunciamos los resultados. Más precisamente, probamos la consistencia de los estimadores en los Teoremas 3.2.2 y 3.2.3. En los Teoremas 3.2.4 y 3.2.5 probamos que si la función de penalización es convenientemente elegida, estos estimadores tienen la propiedad oráculo.

El marco en el que se trabaja en este capítulo es el siguiente. Consideramos una sucesión de modelos de regresión

$$y_{i,n} = \mathbf{x}_{i,n}^T \boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n,$$

donde $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ es un vector a estimar y $u_{i,n}$ son variables aleatorias independientes e idénticamente distribuidas definidas en un mismo espacio de probabilidad, con función de distribución F_0. Para simplificar la notación, a partir de ahora no incluiremos el subíndice n en $y_{i,n}, \mathbf{x}_{i,n}$, $\boldsymbol{\beta}_{0,n}$ and $u_{i,n}$.

Dada una muestra $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, ..., n$, $\gamma_n \geq 0$, $r > 0$, una ρ-función acotada ρ_0 y $0 < b < 1$, el estimador de regresión S-Bridge está definido por

$$\hat{\boldsymbol{\beta}}_{PS} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} n\, s_n^S(\mathbf{r}(\boldsymbol{\beta}))^2 + \gamma_n \|\boldsymbol{\beta}\|_r^r,$$

donde $s_n^S(\mathbf{r}(\boldsymbol{\beta}))$ es el M-estimador de escala residual definido utilizando ρ_0 y b.

Trabajaremos con una definición de estimadores MM-Bridge y MM-Bridge adaptivos más general que la dada en el Capítulo 2. Dada una ρ-función acotada ρ_1, $\lambda_n \geq 0$ y $q > 0$ el estimador de regresión MM-Bridge está definido por

$$\hat{\boldsymbol{\beta}}_B = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n} \right) + \lambda_n \|\boldsymbol{\beta}\|_q^q,$$

donde s_n es una variable aleatoria positiva definida en el mismo espacio de probabilidad que los errores u_i.

Dados $\varsigma > 0$, $t > 0$ y $\iota_n \geq 0$ el estimador de regresión MM-Bridge adaptivo está definido por

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n} \right) + \iota_n \sum_{j=1}^p \frac{|\beta_j|^t}{|\hat{\beta}_{ini,j}|^\varsigma},$$

donde $\hat{\boldsymbol{\beta}}_{ini}$ es un estimador inicial consistente de $\boldsymbol{\beta}_0$, i.e. $\|\hat{\boldsymbol{\beta}}_{ini} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

En este capítulo, asumimos que s_n, la variable aleatoria utilizada para estandarizar los residuos en las definiciones de los estimadores MM-Bridge y MM-Bridge adaptivos, converge en probabilidad a cierto valor $s_0 > 0$. Este s_0 no tiene porque ser necesariamente igual al que aparece en (1.3). Por ejemplo, de acuerdo al Lema 3.2.1, podríamos tomar $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))$ o $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$, donde $\hat{\boldsymbol{\beta}}_S$ es el correspondiente S-estimador no penalizado.

Habiendo retirado el requerimiento sobre la forma de el estimador de escala utilizado en la definición de los estimadores MM-Bridge y MM-Bridge adaptivos, $\hat{\boldsymbol{\beta}}_A$ y $\hat{\boldsymbol{\beta}}_B$ son solo M-estimadores de regresión penalizados, definidos utilizando una función de pérdida acotada y un estimador de escala. Consideremos un M-estimador de regresión definido utilizando una función de pérdida acotada y un estimador de escala. Esto es un estimador MM-Bridge con $\lambda_n \equiv 0$ y por lo tanto los Teoremas 3.2.2, 3.2.3 y 3.2.6 prueban su consistencia y su normalidad asintótica. En particular, tomando $s_n = s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$, $\rho_1 = \rho_0$ para el caso de los S-estimadores y ρ_1 que cumple $\rho_1 \leq \rho_0$ para el caso de los MM-estimadores, los Teoremas 3.2.2, 3.2.3 y 3.2.6 prueban la consistencia y normalidad asintótica de los S-estimadores y MM-estimadores no penalizados.

Chapter 4

Technical Appendix

4.1 Proofs for Chapter 2

Proof of Theorem 2.2.1. Let $m = nFBP(\hat{\boldsymbol{\beta}}_2)$. Take $C \subset \{1, 2, ..., n\}$ such that $\#C \leq m$ and a sequence $(\mathbf{x}_{Ni}^T, y_{Ni})_{N\in\mathbb{N}}$, such that $(\mathbf{x}_{N,i}^{\mathrm{T}}, y_{N,i}) = (\mathbf{x}_i^T, y_i)$ for $i \notin C$ and all $N \in \mathbb{N}$. Let $\hat{\boldsymbol{\beta}}_A^N$, $\hat{\boldsymbol{\beta}}_2^N$ and $\hat{\boldsymbol{\beta}}_1^N$ denote the estimators $\hat{\boldsymbol{\beta}}_A$, $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_1$ computed in $(\mathbf{x}_{Ni}^T, y_{Ni})$. Note that since $\#C \leq m$, $\hat{\boldsymbol{\beta}}_2^N$ is bounded. Since there are a finite number of sets included in $\{1, ..., n\}$, to prove the theorem it will be enough to show that $(\hat{\boldsymbol{\beta}}_A^N)_N$ is bounded. Suppose that this is not so, then eventually passing to a subsequence we can assume that for some j_0, $|\hat{\beta}_{A,j_0}^N| \to \infty$ when $N \to \infty$. Hence, there exists N_0, such that for $N \geq N_0$, $\hat{\beta}_{2,j_0}^N \neq 0$. It follows that $|\hat{\beta}_{A,j_0}^N|^t / |\hat{\beta}_{2,j_0}^N|^\varsigma \to \infty$.

Since ρ_1 is bounded, for sufficiently large N we have that

$$\sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\hat{\boldsymbol{\beta}}_A^N)}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1^N))} \right) + \iota_n \sum_{i=1}^{p} \frac{|\hat{\beta}_{A,j}^N|^t}{|\hat{\beta}_{2,j}^N|^\varsigma} > \sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\mathbf{0})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1^N))} \right) + \iota_n \sum_{i=1}^{p} \frac{|0|^t}{|\hat{\beta}_{2,j}^N|^\varsigma}$$

which contradicts the definition of $\hat{\boldsymbol{\beta}}_A^N$. □

Define for $\boldsymbol{\beta} \in \mathbb{R}^p$ $s(\boldsymbol{\beta})$ by

$$\mathbb{E}\rho_0 \left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{s(\boldsymbol{\beta})} \right) = b,$$

and let

$$g(\boldsymbol{\beta}) = \mathbb{E}\rho_1 \left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}}{s_0} \right).$$

Note that $s(\boldsymbol{\beta}_0) = s_0$ by definition. It can be readily verified that $s(\boldsymbol{\beta})$ is continuous and positive. Lemma 4.1 of Yohai and Zamar (1986) shows that $s(\boldsymbol{\beta})$ has a unique minimum at

$\boldsymbol{\beta} = \boldsymbol{\beta_0}$, and hence proves the Fisher consistency of S-estimators. The same lemma shows that $g(\boldsymbol{\beta})$ has a unique minimum at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

The following Lemma, which appears in Yohai and Zamar (1986) as Lemma 4.5, is a key result.

Lemma 4.1.1. *Let $(\mathbf{x}_i^T, y_i)$, $i = 1, ..., n$, be i.i.d observations with distribution H_0, which satisfies* (2.2)*. Assume [A1]-[A3] hold. Let $K \subseteq \mathbb{R}^p$ be a compact set. Then*

$$\sup_{\boldsymbol{\beta} \in K} |s_n(\mathbf{r}(\boldsymbol{\beta})) - s(\boldsymbol{\beta})| \overset{a.s.}{\to} 0$$

To ease notation, throughout this section we will note $s_n = s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))$, where $\hat{\boldsymbol{\beta}}_1$ is as in (2.4).

Proof of Theorem 2.2.2. We first prove (i). Let

$$Z_n^1(\boldsymbol{\beta}) = s_n^2(\mathbf{r}(\boldsymbol{\beta})) + \frac{\gamma_n}{n} \|\boldsymbol{\beta}\|_r^r,$$

so that $\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Z_n^1(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}_{PS}$. To prove (i), it suffices to show that

$$\hat{\boldsymbol{\beta}}_{PS} \text{ is bounded with probability 1} \tag{4.1}$$

and that given a compact set K, we have that

$$\sup_{\boldsymbol{\beta} \in K} |Z_n^1(\boldsymbol{\beta}) - s^2(\boldsymbol{\beta})| \overset{a.s.}{\to} 0. \tag{4.2}$$

Theorem 4.1 of Yohai and Zamar (1986) shows that $\hat{\boldsymbol{\beta}}_S$ converges almost surely to $\boldsymbol{\beta}_0$ and so (4.1) follows from (2.3). Theorem 4.1 of Yohai and Zamar (1988) is stated for τ-estimators, which include as a special case S-estimators. The theorem requires the hypotesis $2\rho_0 - x\rho_0'(x) \geq 0$, but this is not needed for the case of S-estimators. The theorem also requires that $\mathbb{P}(\mathbf{x}^T\boldsymbol{\beta} = 0) < 0.5$ for all non-zero $\boldsymbol{\beta} \in \mathbb{R}^p$, but this is only because in Yohai and Zamar (1988) the constant b used to define $s_n()$ is assumed to be equal to 0.5. In the general case, one should assume [A2], as we do.

Note that the second term in Z_n^1 converges uniformly to zero over compact sets, and hence Lemma 4.1.1 and the continuity of $s(\boldsymbol{\beta})$ show that (4.2) holds. Thus Theorem 2.2.2(i) is proven.

Next, we prove (iii). The proof of (ii) is essentially the same, and is thus omitted.

Note that by definition of $\hat{\boldsymbol{\beta}}_A$

$$\frac{1}{n}\sum_{i=1}^n \rho_1\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_A}{s_n}\right) \leq \frac{1}{n}\sum_{i=1}^n \rho_1\left(\frac{u_i}{s_n}\right) + \frac{\iota_n}{n}\sum_{j=1}^p \frac{|\beta_{0,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma} \tag{4.3}$$

The second term on the right hand side of (4.3) is

$$\leq \frac{\iota_n}{n} \sum_{j=1}^{k} \frac{|\beta_{0,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma} \to 0 \; a.s. \, ,$$

since $\hat{\boldsymbol{\beta}}_2$ is strongly consistent by assumption.

One can easily show that the family of functions

$$\mathcal{H} = \left\{ h_{s,\mathbf{b}} = \rho_1 \left(\frac{y - \mathbf{x}^T \mathbf{b}}{s} \right) : \; \mathbf{b} \in \mathbb{R}^p, \; s > 0 \right\},$$

is VC with a constant envelope. See the proof of Lemma 4.2.2. It follows that $\mathcal{H}$ is a Glivenko-Cantelli class of functions, i.e.

$$\sup_{\mathbf{b} \in \mathbb{R}^p, \, s>0} | \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s} \right) - \mathbb{E} \rho_1 \left(\frac{y - \mathbf{x}^T \mathbf{b}}{s} \right) | \stackrel{a.s.}{\to} 0. \tag{4.4}$$

Since $\hat{\boldsymbol{\beta}}_1$ is strongly consistent by assumption, by Lemma 4.1.1, $s_n \stackrel{a.s.}{\to} s_0$. Thus, it follows from (4.4) and the Bounded Convergence Theorem that the right hand side of (4.3) converges almost surely to

$$b^* = \mathbb{E} \rho_1 \left(\frac{u}{s_0} \right).$$

Hence,

$$\limsup \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_A}{s_n} \right) \leq b^* \text{ a.s..} \tag{4.5}$$

Note that $b^* \leq b$.

Fix $\varepsilon > 0$. We will show that

$$\liminf \inf_{\varepsilon \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|} \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{s_n} \right) > b^* \text{ a.s..}$$

Note that

$$\begin{aligned}
&\inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{s_n}\right)\\
&\ge \inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\left(\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{s_n}\right)-\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n}\right)\right)\\
&+\inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\left(\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n}\right)-\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)\right)+\inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)\\
&\ge -\sup_{\boldsymbol{\beta}\in\mathbb{R}^p,s>0}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{s}\right)-\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s}\right)\right|\\
&-\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\left|\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n}\right)-\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)\right|+\inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)\\
&= I+II+III.
\end{aligned}$$

Here the expectation in $\mathbb{E}\rho_1\left((y-\mathbf{x}^T\boldsymbol{\beta})/s_n\right)$ is taken only with respect to y and $\mathbf{x}$. By (4.4), term I converges almost surely to zero. Let $\phi_1(t)=\psi_1(t)t$. By [A1] ϕ_1 is bounded. Applying the Mean Value Theorem, we get that for all $\mathbf{x},\boldsymbol{\beta}\in\mathbb{R}^p$, $y\in\mathbb{R}$

$$\begin{aligned}
\left|\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n}\right)-\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)\right| &\le \left|\psi_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n^*}\right)\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_n^*}\right)\right|\left|\frac{s_n-s_0}{s_n^*}\right|\\
&\le \|\phi_1\|_\infty\left|\frac{s_n-s_0}{s_n^*}\right|,
\end{aligned}$$

where $|s_n^*-s_0|\le|s_n-s_0|$. Hence, term II converges almost surely to zero. We will show that $III>b^*$. Suppose $III\le b^*$. Take $\boldsymbol{\beta}_n$ such that $\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|\ge\varepsilon$ for all n and

$$\inf_{\varepsilon\le\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|}\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{s_0}\right)=\lim\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}_n}{s_0}\right).$$

If for some subsequence n_k, $\boldsymbol{\beta}_{n_k}$ converges, say to $\boldsymbol{\beta}^*$ with $\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0\|\ge\varepsilon$, it follows from the Bounded Convergence Theorem that

$$\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}^*}{s_0}\right)\le b^*.$$

This contradicts the fact that $\mathbb{E}\rho_1\left((y-\mathbf{x}^T\boldsymbol{\beta})/s_0\right)$ has a unique minimum at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. It must be that $\|\boldsymbol{\beta}_n\|\to\infty$. We can assume, eventually passing to a subsequence, that $\boldsymbol{\beta}_n^*=\boldsymbol{\beta}_n/\|\boldsymbol{\beta}_n\|\to\boldsymbol{\beta}^*$ with $\|\boldsymbol{\beta}^*\|=1$. It follows that

$$\begin{aligned}
b^*\ge\liminf\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}_n}{s_0}\right)&=\liminf\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}_n^*\|\boldsymbol{\beta}_n\|}{s_0}\right)\\
&\ge\liminf\mathbb{E}\rho_1\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}_n^*\|\boldsymbol{\beta}_n\|}{s_0}\right)I\{\mathbf{x}^T\boldsymbol{\beta}_n^*\ne 0\}.
\end{aligned}$$

By [A1] and Fatou's Lemma, the right hand side of the last inequality is at least

$$\mathbb{P}\left(\mathbf{x}^T\boldsymbol{\beta}^* \neq 0\right).$$

By [A2]

$$\mathbb{P}\left(\mathbf{x}^T\boldsymbol{\beta}^* \neq 0\right) > b.$$

Hence $b^* > b$, which is absurd. It follows that $III > b^*$. We have thus shown that for any $\varepsilon > 0$

$$\liminf \inf_{\varepsilon \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|} \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{s_n}\right) > b^* \text{ a.s..} \tag{4.6}$$

It follows from (4.5) and (4.6) that

$$\hat{\boldsymbol{\beta}}_A \overset{a.s.}{\to} \boldsymbol{\beta}_0.$$

□

Proof of Theorem 2.2.3. We prove (ii), the proof of (i) is essentially the same.

Let

$$Z_n^2(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) + \frac{\iota_n}{n}\sum_{j=1}^{p} \frac{|\beta_j|^t}{|\hat{\beta}_{2,j}|^\varsigma},$$

so that $\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} Z_n^2(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}_A$.

Note that

$$Z_n^2(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{u_i}{s_n}\right) + \frac{\iota_n}{n}\sum_{j=1}^{k} \frac{|\beta_{0,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma}.$$

A first order Taylor expansion shows that

$$\begin{aligned}
Z_n^2(\hat{\boldsymbol{\beta}}_A) &= \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{r_i(\hat{\boldsymbol{\beta}}_A)}{s_n}\right) + \frac{\iota_n}{n}\sum_{j=1}^{p} \frac{|\hat{\beta}_{A,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma} \\
&= \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{u_i}{s_n}\right) - \frac{(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)^{\mathrm{T}}}{n s_n}\sum_{i=1}^{n} \psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_i \\
&+ \frac{1}{2}\frac{1}{s_n^2}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_1'\left(\frac{u_i - \zeta_i\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0) \\
&+ \frac{\iota_n}{n}\sum_{j=1}^{p} \frac{|\hat{\beta}_{A,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma},
\end{aligned}$$

with $0 \leq \zeta_i \leq 1$. We will show that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_1'\left(\frac{u_i - \zeta_i \mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \stackrel{a.s.}{\to} \mathbb{E}\psi_1'\left(\frac{u}{s_0}\right)\mathbf{V}_{\mathbf{x}}.$$

By Lemma 4.2 of Yohai (1985)

$$\frac{1}{n}\sum_{i=1}^{n} \psi_1'\left(\frac{u_i}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \stackrel{a.s.}{\to} \mathbb{E}\psi_1'\left(\frac{u}{s_0}\right)\mathbf{V}_{\mathbf{x}}$$

and hence it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n} \left(\psi_1'\left(\frac{u_i - \zeta_i \mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)}{s_n}\right) - \psi_1'\left(\frac{u_i}{s_n}\right)\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \stackrel{a.s.}{\to} 0. \tag{4.7}$$

By the Bounded Convergence Theorem

$$r(\varepsilon) = \mathbb{E}\sup_{|s-s_0|\leq\varepsilon, |v|\leq\varepsilon}\left|\psi_1'\left(\frac{u-v}{s}\right) - \psi_1'\left(\frac{u}{s}\right)\right| \to 0$$

when $\varepsilon \to 0$. Fix $\eta > 0$. Take ε_0 such that $r(\varepsilon_0) < \eta^*$, where η^* will be chosen shortly. By [A4], there exists $M > 0$ such that $\mathbb{E}\|\mathbf{x}\|^2 I\{\|\mathbf{x}\| > M\} < \eta^*$. Since $\hat{\boldsymbol{\beta}}_A \stackrel{a.s.}{\to} \boldsymbol{\beta}_0$ and $s_n \stackrel{a.s.}{\to} s_0$, with probability one, for large enough n, $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| < \varepsilon_0/M$ and $|s_n - s_0| < \varepsilon_0$. By the Law of Large Numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{|s-s_0|\leq\varepsilon_0, |v|\leq\varepsilon_0}\left|\psi_1'\left(\frac{u_i-v}{s}\right) - \psi_1'\left(\frac{u_i}{s}\right)\right|\|\mathbf{x}_i\|^2 \stackrel{a.s.}{\to} r(\varepsilon_0)\mathbb{E}\|\mathbf{x}\|^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 I\{\|\mathbf{x}_i\| > M\} \stackrel{a.s.}{\to} \mathbb{E}\|\mathbf{x}\|^2 I\{\|\mathbf{x}\| > M\}.$$

Hence, with probability one, for sufficiently large n

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{|s-s_0|\leq\varepsilon_0, |v|\leq\varepsilon_0}\left|\psi_1'\left(\frac{u_i-v}{s}\right) - \psi_1'\left(\frac{u_i}{s}\right)\right|\|\mathbf{x}_i\|^2 < r(\varepsilon_0)\mathbb{E}\|\mathbf{x}\|^2 + \eta^*$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 I\{\|\mathbf{x}_i\| > M\} < \mathbb{E}\|\mathbf{x}\|^2 I\{\|\mathbf{x}\| > M\} + \eta^*.$$

Take $\eta^* < \max\{\eta/(4\mathbb{E}\|\mathbf{x}\|^2), \eta/(8\|\psi_1'\|_\infty), \eta/4\}$. Hence, for any $1 \le l, j \le p$, with probability one, for large enough n

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0)}{s_n}\right)-\psi_1'\left(\frac{u_i}{s_n}\right)\right||x_{i,j}x_{i,l}| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0)}{s_n}\right)-\psi_1'\left(\frac{u_i}{s_n}\right)\right|\|\mathbf{x}_i\|^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left|\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0)}{s_n}\right)-\psi_1'\left(\frac{u_i}{s_n}\right)\right|\|\mathbf{x}_i\|^2 I\{\|\mathbf{x}_i\|\le M\} \\
&\qquad +2\|\psi_1'\|_\infty\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 I\{\|\mathbf{x}_i\|>M\} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{|s-s_0|\le\varepsilon_0,|v|\le\varepsilon_0}\left|\psi_1'\left(\frac{u_i-v}{s}\right)-\psi_1'\left(\frac{u_i}{s}\right)\right|\|\mathbf{x}_i\|^2 \\
&\qquad +2\|\psi_1'\|_\infty\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|^2 I\{\|\mathbf{x}_i\|>M\} \\
&< r(\varepsilon_0)\mathbb{E}\|\mathbf{x}\|^2+\eta^*+2\|\psi_1'\|_\infty\mathbb{E}\|\mathbf{x}\|^2 I\{\|\mathbf{x}\|>M\}+\eta^* \\
&< \eta.
\end{aligned}$$

We have proven (4.7).

Then

$$\begin{aligned}
A_n &= \frac{1}{2}\frac{1}{s_n^2}(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0)^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)(\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0) \\
&\ge c_n\|\hat{\boldsymbol{\beta}}_A-\boldsymbol{\beta}_0\|^2,
\end{aligned}$$

where $c_n \xrightarrow{a.s.} c_0 > 0$.

We also have that by Lemma 5.1 of Yohai (1985) and the Central Limit Theorem

$$B_n = \frac{1}{\sqrt{n}s_n}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_i = O_P(1).$$

Put

$$C_n = \frac{\iota_n}{n}\sum_{j=1}^{p}\frac{|\hat{\beta}_{A,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma}-\frac{|\beta_{0,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma}.$$

Then, since $\hat{\boldsymbol{\beta}}_A$ is strongly consistent for $\boldsymbol{\beta}_0$ and the first k coordinates of $\boldsymbol{\beta}_0$ are non-zero, for large enough n the first k coordinates of $\hat{\boldsymbol{\beta}}_A$ stay away from zero with arbitrarily high

probability. Applying the Mean Value Theorem we get that

$$C_n \geq \frac{\iota_n}{n}\sum_{j=1}^{k}\frac{|\hat{\beta}_{A,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma} - \frac{|\beta_{0,j}|^t}{|\hat{\beta}_{2,j}|^\varsigma} = \frac{\iota_n}{n}\sum_{j=1}^{k} t\ sgn(\theta_j)\frac{|\theta_j|^{t-1}}{|\hat{\beta}_{2,j}|^\varsigma}(\hat{\beta}_{A,j} - \beta_{0,j}),$$

for some θ_j such that $|\theta_j - \beta_{0,j}| \leq |\hat{\beta}_{A,j} - \beta_{0,j}|$. Since $\iota_n = O(\sqrt{n})$ and $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_2$ are consistent, we have that for some $M > 0$, for large enough n, with arbitrarily high probability

$$C_n \geq \frac{-M}{\sqrt{n}}\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\|.$$

Then

$$\begin{aligned} Z_n^2(\hat{\boldsymbol{\beta}}_A) - Z_n^2(\boldsymbol{\beta}_0) &= A_n - \frac{1}{\sqrt{n}}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)^{\mathrm{T}} B_n + C_n \\ &\geq c_n\|(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)\|^2 - \frac{1}{\sqrt{n}}\|(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)\|\|B_n\| + C_n \\ &= \frac{1}{\sqrt{n}}\|(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)\|(c_n\sqrt{n}\|(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)\| - \|B_n\| \\ &+ \sqrt{n}/\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\|C_n). \end{aligned}$$

Now, since $Z_n^2(\hat{\boldsymbol{\beta}}_A) - Z_n^2(\boldsymbol{\beta}_0) \leq 0$, we have that.

$$\sqrt{n}\|\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_0\| \leq \frac{\|B_n\| - \sqrt{n}/\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\|C_n}{c_n}.$$

But

$$\sqrt{n}/\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\|C_n \geq -M.$$

Hence, $\sqrt{n}\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| = O_P(1)$.

□

Proof of Theorem 2.2.4. We prove (ii). The proof of (i) is essentially the same.

We follow Lemma 2 of Huang et al. (2008). Since by Theorem 2.2.3 $\hat{\boldsymbol{\beta}}_A$ is $\sqrt{n}$-consistent, for a sufficiently large $C > 0$ and n, $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0\| \leq C/\sqrt{n}$ with arbitrarily high probability.

Let

$$\begin{aligned} V_n(\mathbf{u}_1, \mathbf{u}_2) = &\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I} + \mathbf{u}_1/\sqrt{n}, \boldsymbol{\beta}_{0,II} + \mathbf{u}_2/\sqrt{n})}{s_n}\right) \\ &+ \iota_n\left(\sum_{j=1}^{k}\frac{|\beta_{0,j} + u_{1,j}/\sqrt{n}|^t}{|\hat{\beta}_{2,j}|^\varsigma} + \sum_{j=k+1}^{p}\frac{|u_{2,j-k}/\sqrt{n}|^t}{|\hat{\beta}_{2,j}|^\varsigma}\right). \end{aligned}$$

Then for large enough n, with arbitrarily high probability, $(\hat{\boldsymbol{\beta}}_{A,I}, \hat{\boldsymbol{\beta}}_{A,II})$ is obtained by minimizing $V_n(\mathbf{u}_1, \mathbf{u}_2)$ over $\|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 \leq C^2$. We will show that if $\|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 \leq C^2$ and $\|\mathbf{u}_2\| > 0$ then, for large enough n, $V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}_{p-k}) > 0$ with arbitrarily high probability and the theorem will follow.

Let $(\mathbf{u}_1, \mathbf{u}_2)$ satisfy $\|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 \leq C^2$. It is easy to see that

$$V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}_{p-k}) =$$
$$\sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I} + \mathbf{u}_1/\sqrt{n}, \mathbf{u}_2/\sqrt{n})}{s_n}\right) - \rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I} + \mathbf{u}_1/\sqrt{n}, \mathbf{0}_{p-k})}{s_n}\right) +$$
$$\frac{\iota_n}{n^{t/2}} \sum_{j=k+1}^{p} \frac{|u_{2,j-s}|^t}{|\hat{\beta}_{2,j}|^\varsigma} =$$
$$(I) + (II).$$

Applying the Mean Value Theorem we get

$$(I) = (\mathbf{0}_s, \mathbf{u}_2)^T \frac{1}{\sqrt{n}} \frac{-1}{s_n} \sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\theta}_n^*)}{s_n}\right) \mathbf{x}_i,$$

where $\boldsymbol{\theta}_n^* = (\boldsymbol{\beta}_{0,I} + \mathbf{u}_1/\sqrt{n}, (1-\alpha_n)\mathbf{u}_2/\sqrt{n})$ for some $\alpha_n \in [0,1]$. Applying the Mean Value Theorem once more we get

$$(\mathbf{0}_s, \mathbf{u}_2)^T \frac{1}{\sqrt{n}} \frac{-1}{s_n} \sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\theta}_n^*)}{s_n}\right) \mathbf{x}_i = \frac{1}{\sqrt{n}} \frac{-1}{s_n} (\mathbf{0}_s, \mathbf{u}_2)^T \sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right) \mathbf{x}_i +$$
$$\frac{1}{\sqrt{n}} \frac{1}{s_n^2} (\mathbf{0}_s, \mathbf{u}_2)^T \sum_{i=1}^{n} \psi_1'\left(\frac{r_i(\boldsymbol{\theta}_n^{**})}{s_n}\right) \mathbf{x}_i \mathbf{x}_i^T (\mathbf{u}_1/\sqrt{n}, (1-\alpha_n)\mathbf{u}_2/\sqrt{n}) =$$
$$\frac{1}{\sqrt{n}} \frac{-1}{s_n} (\mathbf{0}_s, \mathbf{u}_2)^T \sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right) \mathbf{x}_i +$$
$$\frac{1}{n} \frac{1}{s_n^2} (\mathbf{0}_s, \mathbf{u}_2)^T \sum_{i=1}^{n} \psi_1'\left(\frac{r_i(\boldsymbol{\theta}_n^{**})}{s_n}\right) \mathbf{x}_i \mathbf{x}_i^T (\mathbf{u}_1, (1-\alpha_n)\mathbf{u}_2),$$

where $\|\boldsymbol{\theta}_n^{**} - \boldsymbol{\beta}_0\| \leq \|\boldsymbol{\theta}_n^* - \boldsymbol{\beta}_0\|$. By Lemma 4.1.1, Lemma 5.1 of Yohai (1985) and the Central Limit Theorem, the first term in the last equation is $\|\mathbf{u}_2\| O_P(1)$, uniformly in $\mathbf{u}_2$. The second

term satisfies

$$\begin{aligned}
&\left|\frac{1}{n}\frac{1}{s_n^2}(\mathbf{0}_s,\mathbf{u}_2)^T\sum_{i=1}^n\psi_1'\left(\frac{r_i(\boldsymbol{\theta}_n^{**})}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T(\mathbf{u}_1,(1-\alpha_n)\mathbf{u}_2)\right| \\
&\leq \frac{1}{n}\frac{\|\psi_1'\|_\infty}{s_n^2}\sum_{i=1}^n\|\mathbf{u}_2\|\|\mathbf{x}_i\|^2\|(\mathbf{u}_1,(1-\alpha_n)\mathbf{u}_2)\| \\
&\leq \|\mathbf{u}_2\|\frac{\|\psi_1'\|_\infty}{s_n^2}C\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i\|^2 \\
&= \|\mathbf{u}_2\|O_P(1),
\end{aligned}$$

by Lemma 4.1.1, the fact that by [A1] ψ_1' is bounded, [A4] and the Law of Large Numbers.

On the other hand

$$\frac{\iota_n}{n^{t/2}}\sum_{j=k+1}^p\frac{|u_{2,j-k}|^t}{|\hat\beta_{2,j}|^\varsigma}=\iota_n n^{(\varsigma-t)/2}\sum_{j=k+1}^p\frac{|u_{2,j-k}|^t}{|\sqrt{n}\hat\beta_{2,j}|^\varsigma}=\iota_n n^{(\varsigma-t)/2}\Omega_P(\|\mathbf{u}_2\|_t^t)$$

uniformly in $\mathbf{u}_2$, since $\hat{\boldsymbol{\beta}}_2$ is $\sqrt{n}$-consistent by assumption. Note also that $\|\mathbf{u}_2\|_t^t \geq \|\mathbf{u}_2\|^t$. Hence, for some $M_1, M_2 > 0$ and sufficiently large n, for all $(\mathbf{u}_1,\mathbf{u}_2)$ that satisfy $\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2 \leq C^2$, with arbitrarily high probability, we have that

$$\begin{aligned}
V_n(\mathbf{u}_1,\mathbf{u}_2)-V_n(\mathbf{u}_1,\mathbf{0}_{p-s}) &> -M_1\|\mathbf{u}_2\|+M_2\iota_n n^{(\varsigma-t)/2}\|\mathbf{u}_2\|^t \\
&= \|\mathbf{u}_2\|^t(-M_1\|\mathbf{u}_2\|^{1-t}+M_2\iota_n n^{(\varsigma-t)/2}) \\
&\geq \|\mathbf{u}_2\|^t(-M_1C^{1-t}+M_2\iota_n n^{(\varsigma-t)/2}). \qquad (4.8)
\end{aligned}$$

Finally, since by assumption $\iota_n n^{(\varsigma-t)/2}\to\infty$, we have that for sufficiently large n, for any non-zero $\mathbf{u}_2$, the right hand side of (4.8) is strictly positive .

□

Proof of Theorem 2.2.5. We prove (ii). The proof of (i) is essentially the same.

For $\boldsymbol{\theta}\in\mathbb{R}^k$ let $\mathbf{p}'(\boldsymbol{\theta})=t\sum_{j=1}^k sgn(\theta_j)|\theta_j|^{t-1}/|\hat\beta_{2,j}|^\varsigma\mathbf{e}_j$. Note that by Theorem 2.2.2, $\hat{\boldsymbol{\beta}}_A$ is strongly consistent for $\boldsymbol{\beta}_0$ and hence with probability 1 all the coordinates of $\hat{\boldsymbol{\beta}}_{A,I}$ stay away from zero for a sufficiently large n. Also, by Theorem 2.2.4, $\hat{\boldsymbol{\beta}}_{A,II}=\mathbf{0}_{p-k}$ with probability tending to one. Then for large enough n, with arbitrarily high probability the partial derivatives for the first k coordinates of Z_n^2 at $\hat{\boldsymbol{\beta}}_A$ exist, and hence

$$\begin{aligned}
\mathbf{0}_k &= \frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^n\psi_1\left(\frac{y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_A}{s_n}\right)\mathbf{x}_{i,I}+\frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I}) \\
&= \frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^n\psi_1\left(\frac{y_i-\mathbf{x}_{i,I}^T\hat{\boldsymbol{\beta}}_{A,I}}{s_n}\right)\mathbf{x}_{i,I}+\frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})+\mathbf{r}_n,
\end{aligned}$$

where $\mathbb{P}(\mathbf{r}_n = \mathbf{0}_k) \to 1$. Then the Mean Value Theorem gives

$$\mathbf{0}_k = \frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_{i,I} + \frac{1}{s_n^2}\mathbf{W}_n\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) + \frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I}) + \mathbf{r}_n,$$

where

$$\mathbf{W}_n = \frac{1}{n}\sum_{i=1}^{n}\psi_1'\left(\frac{u_i - \zeta_i\mathbf{x}_{i,I}^T(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I})}{s_n}\right)\mathbf{x}_{i,I}\mathbf{x}_{i,I}^T$$

and $0 \leq \zeta_i \leq 1$.

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) = s_n\mathbf{W}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_{i,I} - s_n^2\frac{\iota_n}{\sqrt{n}}\mathbf{W}_n^{-1}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I}) - s_n^2\mathbf{W}_n^{-1}\mathbf{r}_n.$$

By Lemma 4.1.1, $s_n \xrightarrow{a.s.} s_0$. By Lemma 5.1 of Yohai (1985) and the Central Limit Theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_{i,I} \xrightarrow{d} N_k\left(\mathbf{0}, a(\psi_1)\mathbf{V}_{\mathbf{x}_I}\right).$$

Proceeding as in the proof of Theorem 2.2.3, by Lemma 4.2 of Yohai (1985) and Lemma 4.1.1

$$\mathbf{W}_n \xrightarrow{a.s.} b(\psi_1)\mathbf{V}_{\mathbf{x}_I}.$$

Since $\iota_n/\sqrt{n} \to 0$, the theorem follows from Sluztky's Theorem.

□

Proof of Theorem 2.2.6. We define for $\mathbf{z} \in \mathbb{R}^p$

$$R_n(\mathbf{z}) = \sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0 + \mathbf{z}/\sqrt{n})}{s_n}\right) - \rho_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right) + \lambda_n\Big(\sum_{j=1}^{p}|\beta_{0,j} + z_j/\sqrt{n}|^q - |\beta_{0,j}|^q\Big),$$

so that $\arg\min(R_n) = \sqrt{n}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)$. We will show that R_n converges weakly to R in in the space of locally bounded functions with the topology of uniform convergence over compact sets. To do so, we will verify conditions (i) and (ii) of Theorem 2.3 of Kim and Pollard (1990).

We first prove condition (i): finite-dimensional convergence of R_n to R. A first order Taylor expansion shows that

$$\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0 + \mathbf{z}/\sqrt{n})}{s_n}\right) - \rho_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right) = \tag{4.9}$$

$$-\mathbf{z}^T\frac{1}{s_n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i + \frac{1}{2}\frac{1}{s_n^2}\mathbf{z}^T\frac{1}{n}\sum_{i=1}^{n}\psi_1'\left(\frac{u_i - \zeta_i\mathbf{x}_i^T\mathbf{z}/\sqrt{n}}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T\mathbf{z},$$

with $0 \leq \zeta_i \leq 1$.

It can be easily verified that for $q > 1$

$$\lambda_n(\sum_{j=1}^{p} |\beta_{0,j} + z_j/\sqrt{n}|^q - |\beta_{0,j}|^q) \to \lambda_0 q \sum_{j=1}^{p} z_j sgn(\beta_{0,j})|\beta_{0,j}|^{q-1} \tag{4.10}$$

uniformly over compact sets, whereas for $q = 1$

$$\lambda_n(\sum_{j=1}^{p} |\beta_{0,j} + z_j/\sqrt{n}| - |\beta_{0,j}|) \to \lambda_0 \sum_{j=1}^{p} (z_j sgn(\beta_{0,j}) I\{\beta_{0,j} \neq 0\} + |z_j| I\{\beta_{0,j} = 0\}), \tag{4.11}$$

uniformly over compact sets.

Then the finite-dimensional convergence follows from (4.9), (4.10), (4.11), Lemma 4.1.1, Lemmas 4.2 and 5.1 of Yohai (1985), Slutzky's Theorem and the Cramer-Wold device. See the proof of Theorem 2.2.5 for more details.

We now turn to proving condition (ii) of Theorem 2.3 of Kim and Pollard (1990), the stochastic equicontinuity of R_n. Fix ε, η and $M > 0$. Take $\mathbf{z}, \mathbf{z}'$ that satisfy $\|\mathbf{z}\| \leq M, \|\mathbf{z}'\| \leq M$.

A first order Taylor expansion shows that

$$\sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta}_0 + \mathbf{z}/\sqrt{n})}{s_n}\right) - \rho_1\left(\frac{r_i(\boldsymbol{\beta}_0 + \mathbf{z}'/\sqrt{n})}{s_n}\right) =$$

$$-(\mathbf{z} - \mathbf{z}')^T \frac{1}{s_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1\left(\frac{u_i - \mathbf{x}_i^T \mathbf{z}'/\sqrt{n}}{s_n}\right) \mathbf{x}_i +$$

$$\frac{1}{2} \frac{1}{s_n^2} (\mathbf{z} - \mathbf{z}')^T \frac{1}{n} \sum_{i=1}^{n} \psi_1'\left(\frac{u_i - \zeta_i \mathbf{x}_i^T \mathbf{z}/\sqrt{n} - (1-\zeta_i)\mathbf{x}_i^T \mathbf{z}'/\sqrt{n}}{s_n}\right) \mathbf{x}_i \mathbf{x}_i^T (\mathbf{z} - \mathbf{z}'),$$

with $0 \leq \zeta_i \leq 1$. Applying the Mean Value Theorem to the first term in the Taylor expansion we get

$$-(\mathbf{z} - \mathbf{z}')^T \frac{1}{s_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1\left(\frac{u_i - \mathbf{x}_i^T \mathbf{z}'/\sqrt{n})}{s_n}\right) \mathbf{x}_i = -(\mathbf{z} - \mathbf{z}')^T \frac{1}{s_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1\left(\frac{u_i}{s_n}\right) \mathbf{x}_i +$$

$$(\mathbf{z} - \mathbf{z}')^T \frac{1}{s_n^2} \frac{1}{n} \sum_{i=1}^{n} \psi_1'\left(\frac{u_i - \kappa_i \mathbf{x}_i^T \mathbf{z}'/\sqrt{n}}{s_n}\right) \mathbf{x}_i \mathbf{x}_i^T \mathbf{z}',$$

with $0 \leq \kappa_i \leq 1$. If $\|\mathbf{z} - \mathbf{z}'\| < \delta$, by Lemma 4.1.1, Lemma 5.1 of Yohai (1985) and the Central Limit Theorem, we have that

$$\left|(\mathbf{z} - \mathbf{z}')^T \frac{1}{s_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1\left(\frac{u_i}{s_n}\right) \mathbf{x}_i\right| \leq \delta O_P(1),$$

and by Lemma 4.1.1, the fact that by [A1] ψ_1' is bounded, [A4] and the Law of Large Numbers

$$\left|(\mathbf{z}-\mathbf{z}')^T\frac{1}{s_n^2}\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{u_i-\kappa_i\mathbf{x}_i^T\mathbf{z}'/\sqrt{n}}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T\mathbf{z}'\right| \le \frac{\|\psi_1'\|_\infty}{s_n^2}\frac{1}{n}\sum_{i=1}^n\left|(\mathbf{z}-\mathbf{z}')^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{z}'\right|$$
$$\le \delta M\frac{\|\psi_1'\|_\infty}{s_n^2}\frac{1}{n}\sum_{i=1}^n\|\mathbf{x}_i\|^2=\delta O_P(1)$$

and similarly

$$\left|\frac{1}{2}\frac{1}{s_n^2}(\mathbf{z}-\mathbf{z}')^T\frac{1}{n}\sum_{i=1}^n\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_i^T\mathbf{z}/\sqrt{n}-(1-\zeta_i)\mathbf{x}_i^T\mathbf{z}'/\sqrt{n}}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T(\mathbf{z}-\mathbf{z}')\right|\le\delta^2O_P(1).$$

Hence

$$|\sum_{i=1}^n\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0+\mathbf{z}/\sqrt{n})}{s_n}\right)-\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0+\mathbf{z}'/\sqrt{n})}{s_n}\right)|\le\delta O_P(1)+\delta^2O_P(1), \tag{4.12}$$

uniformly in $\mathbf{z},\mathbf{z}'$. Let $\mathbb{P}^*$ stand for outer probability. Then it follows from (4.12) that for sufficiently small δ

$$\limsup\mathbb{P}^*\left(\sup|\sum_{i=1}^n\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0+\mathbf{z}/\sqrt{n})}{s_n}\right)-\rho_1\left(\frac{r_i(\boldsymbol{\beta}_0+\mathbf{z}'/\sqrt{n})}{s_n}\right)|>\eta\right)<\varepsilon,$$

where the supremum runs over $\|\mathbf{z}\|\le M,\|\mathbf{z}'\|\le M,\|\mathbf{z}-\mathbf{z}'\|<\delta$. By (4.10) and (4.11), for all $q\ge1$ the penalty terms

$$\lambda_n(\sum_{j=1}^p|\beta_{0,j}+z_j/\sqrt{n}|^q-|\beta_{0,j}|^q)$$

converge uniformly over compact sets to a continuous function. Hence they are uniformly equicontinuous over compact sets. Condition (ii) is proven.

Since by Theorem 2.2.3, $\arg\min(R_n)=\sqrt{n}(\hat{\boldsymbol{\beta}}_B-\boldsymbol{\beta}_0)=O_P(1)$, the theorem follows from Theorem 2.7 of Kim and Pollard (1990). □

Proof of Theorem 2.2.7. After noting that

$$\lambda_n(\sum_{j=1}^p|\beta_{0,j}+z_j/\sqrt{n}|^q-|\beta_{0,j}|^q)\to\lambda_0\sum_{j=1}^p|z_j|^qI\{\beta_{0,j}=0\}$$

uniformly over compact sets, the proof follows along the same lines as the proof of Theorem 2.2.6. □

4.2 Proofs for Chapter 3

Proof of Lemma 3.1.1. Let $0 < \alpha < 1$ be such that $\liminf \eta_n(\alpha) > 0$. Note that for all $\eta > 0$, $\#\{i : \|\mathbf{x}_i\| \geq \eta\} \leq nMp\eta^{-2}$. Take $\eta = \sqrt{2Mp(1-\alpha)^{-1}}$. Let $\alpha_1 = (1/2)(1+\alpha)$, $\eta_1 = \sqrt{2M(1-\alpha)^{-1}}$ and $\mathcal{A} = \{i : \|\mathbf{x}_i\| < \eta_1\sqrt{p}\}$. Then $\#\mathcal{A} \geq n\alpha_1$, with $0 < \alpha < \alpha_1 < 1$.

Take $\boldsymbol{\theta}^*$ with $\|\boldsymbol{\theta}^*\| = 1$ such that

$$\sum_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}^*|^2 = \min_{\|\boldsymbol{\theta}\|=1} \sum_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2.$$

Let $\mathcal{G}$ be the set of $i \in \mathcal{A}$ giving rise to the smallest $[n\alpha]$ values of $|\mathbf{x}_i^T\boldsymbol{\theta}^*|$. Then, by definition of $\eta_n(\alpha)$, $\eta_n(\alpha) \leq \max_{i\in\mathcal{G}} |\mathbf{x}_i^T\boldsymbol{\theta}^*|$. Hence, $\eta_n(\alpha) \leq |\mathbf{x}_i^T\boldsymbol{\theta}^*|$ for all $i \in \mathcal{A} \setminus \mathcal{G}$. Thus

$$\min_{\|\boldsymbol{\theta}\|=1} \frac{1}{n}\sum_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2 = \frac{1}{n}\sum_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}^*|^2 \geq \frac{1}{n}\sum_{i\in\mathcal{A}\setminus\mathcal{G}} |\mathbf{x}_i^T\boldsymbol{\theta}^*|^2 \geq \frac{(n\alpha_1 - [n\alpha])\eta_n(\alpha)^2}{n} \geq (\alpha_1 - \alpha)\eta_n(\alpha)^2.$$

The lemma is proven. □

We will make extensive use of the tools from empirical processes theory that appear in Pollard (1989) and Van der Vaart and Wellner (1996). The results in Pollard (1989), in particular the maximal inequalities of Theorem 4.2, are stated for i.i.d random variables. Following the discussion in pages 1661-1662 of Davies (1990), in Theorem 4.2.1 we adapt Theorem 4.2 of Pollard (1989) to make it directly applicable to our scenario of interest.

We first introduce some notation. Let $\varepsilon > 0$. Let $\mathcal{H}$ be a class of funcions defined on $\mathbb{R}^d$ and let $\|.\|$ be a pseudo-norm on $\mathcal{H}$.

- The capacity number of $\mathcal{H}$, $D(\varepsilon, \mathcal{H}, \|.\|)$, is the largest N such that there exists $h_1, \ldots, h_N$ in $\mathcal{H}$ with $\|h_i - h_j\| > \varepsilon$ for all $i \neq j$. The capacity number is also called the packing number in the literature.

- The covering number of $\mathcal{H}$, $N(\varepsilon, \mathcal{H}, \|.\|)$, is the minimal number of open balls of radius ε needed to cover $\mathcal{H}$.

- Given two functions h, g a bracket $[h, g]$ is the set of all functions f such that $h \leq f \leq g$. An ε-bracket is a bracket $[h, g]$ such that $\|h - g\| < \varepsilon$. $N_{[]}(\varepsilon, \mathcal{H}, \|.\|)$ is the bracketing number of $\mathcal{H}$, that is, the minimum number of ε-brackets needed to cover $\mathcal{H}$.

- Given a metric space (T, d), the covering number of T, $N(\varepsilon, T, d)$, is the minimal number of open balls of radius ε needed to cover T.

It is easy to show that $D(\varepsilon, \mathcal{H}, \|.\|) \leq N(\varepsilon/2, \mathcal{H}, \|.\|) \leq N_{[]}(\varepsilon, \mathcal{H}, \|.\|)$. Given Q, a probability measure on $\mathbb{R}^d$ with finite support, let $\|.\|_{2,Q}$ be the $L^2(Q)$ pseudo-norm.

Theorem 4.2.1. *Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be fixed vectors in $\mathbb{R}^d$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be i.i.d. random vectors in $\mathbb{R}^m$. Let $\mathcal{H}$ be a class of functions defined in $\mathbb{R}^{m+d}$ and taking values in $\mathbb{R}$. Assume $\mathcal{H}$ has envelope H that satisfies $1/n \sum_{i=1}^{n} \mathbb{E}H^2(\mathbf{v}_i, \mathbf{z}_i) < \infty$ and that $\mathcal{H}$ contains the zero function. Furthermore, assume that there exists a decreasing function $D(\varepsilon)$ that satisfies $\int_0^1 (\log D(\varepsilon))^{1/2} d\varepsilon < \infty$, such that for all $0 < \varepsilon < 1$ and any probability measure on $\mathbb{R}^{m+d}$ with finite support Q with $\|H\|_{2,Q} > 0$,*

$$D(\varepsilon \|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q}) \leq D(\varepsilon).$$

Then

(i)

$$\mathbb{E} \sup_{\mathcal{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h(\mathbf{v}_i, \mathbf{z}_i) - \mathbb{E}h(\mathbf{v}_i, \mathbf{z}_i) \right) \right|$$
$$\leq M \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}H^2(\mathbf{v}_i, \mathbf{z}_i) \right)^{1/2} \left(\int_0^1 (\log D(\varepsilon))^{1/2} d\varepsilon \right),$$

(ii)

$$\mathbb{E} \sup_{\mathcal{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h(\mathbf{v}_i, \mathbf{z}_i) - \mathbb{E}h(\mathbf{v}_i, \mathbf{z}_i) \right) \right|^2$$
$$\leq M \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}H^2(\mathbf{v}_i, \mathbf{z}_i) \left(\int_0^1 (\log D(\varepsilon))^{1/2} d\varepsilon \right)^2,$$

where $M > 0$ is a fixed universal constant.

Proof. The proof of this theorem is only a small variation on the proof of Theorem 4.2 of Pollard (1989) and is included in this thesis for completeness sake only.

We prove (ii). Let $\mathbb{P}_n$ be the empirical probability measure that places mass $1/n$ at each of the points $(\mathbf{v}_i, \mathbf{z}_i)$ $i = 1, \ldots, n$. Let $\|.\|_{2,n}$ be the $L^2(\mathbb{P}_n)$ pseudo-norm. Let $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_n$ be i.i.d. random vectors independent of and with the same distribution as $\mathbf{v}_1, \ldots, \mathbf{v}_n$. With a slight abuse of notation denote $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ and let $\mathbb{E}_{\mathbf{v}}$ be the expectation conditional on $\mathbf{v}$. It follows that for all $i = 1, \ldots, n$, $h(\mathbf{v}_i, \mathbf{z}_i) = \mathbb{E}_{\mathbf{v}} h(\mathbf{v}_i, \mathbf{z}_i)$ and $\mathbb{E}h(\mathbf{v}_i, \mathbf{z}_i) = \mathbb{E}h(\tilde{\mathbf{v}}_i, \mathbf{z}_i) = \mathbb{E}_{\mathbf{v}} h(\tilde{\mathbf{v}}_i, \mathbf{z}_i)$. Then, for all $h \in \mathcal{H}$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h(\mathbf{v}_i, \mathbf{z}_i) - \mathbb{E}h(\mathbf{v}_i, \mathbf{z}_i) \right) = \mathbb{E}_{\mathbf{v}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h(\mathbf{v}_i, \mathbf{z}_i) - h(\tilde{\mathbf{v}}_i, \mathbf{z}_i) \right).$$

By Jensen's inequality

$$\left|\mathbb{E}_{\mathbf{v}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2 \leq \mathbb{E}_{\mathbf{v}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2.$$

Hence

$$\begin{aligned}\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-\mathbb{E}h(\mathbf{v}_i,\mathbf{z}_i)\right)\right|^2 &\leq \mathbb{E}\sup_{\mathcal{H}}\mathbb{E}_{\mathbf{v}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2\\ \leq \mathbb{E}\mathbb{E}_{\mathbf{v}}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2 &= \mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2.\end{aligned}$$

Let $g_1,\ldots,g_n$ be i.i.d random variables, independent of $\tilde{\mathbf{v}}_1,\ldots,\tilde{\mathbf{v}}_n$ and of $\mathbf{v}_1,\ldots,\mathbf{v}_n$ such that $g_i \sim N(0,1)$. Define $\sigma_i = g_i/|g_i|$ for $i=1,\ldots,n$. Then $\sigma_1,\ldots,\sigma_n$ are independent of $\tilde{\mathbf{v}}_1,\ldots,\tilde{\mathbf{v}}_n$ and of $\mathbf{v}_1,\ldots,\mathbf{v}_n$. Note that $\mathbb{P}(\sigma_i=1)=\mathbb{P}(\sigma_i=-1)=1/2$ and that σ_i is independent of $|g_i|$. Let $\boldsymbol{\sigma}=(\sigma_1,\ldots,\sigma_n)$. By the symmetry between $\tilde{\mathbf{v}}_i$ and $\mathbf{v}_i$ we have that

$$\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2 = \mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2.$$

Now

$$\sup_{\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right| \leq \sup_{\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i h(\mathbf{v}_i,\mathbf{z}_i)\right| + \sup_{\mathcal{H}}\left|\sum_{i=1}^{n}\sigma_i h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right|.$$

Hence

$$\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(h(\mathbf{v}_i,\mathbf{z}_i)-h(\tilde{\mathbf{v}}_i,\mathbf{z}_i)\right)\right|^2 \leq 4\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i h(\mathbf{v}_i,\mathbf{z}_i)\right|^2.$$

Let γ be the expectation of $|g_1|$ and let $\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}$ be the expectation conditional on $\mathbf{v}$ and $\boldsymbol{\sigma}$. Then for all $i=1,\ldots,n$, $\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}\sigma_i h(\mathbf{v}_i,\mathbf{z}_i) = \sigma_i h(\mathbf{v}_i,\mathbf{z}_i)$ and $\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}|g_i| = \gamma$. Hence, applying Jensen's

inequality

$$\begin{aligned}
\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i h(\mathbf{v}_i,\mathbf{z}_i)\right|^2 &= \mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n \sigma_i h(\mathbf{v}_i,\mathbf{z}_i)\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}|g_i|/\gamma\right|^2 \\
= \mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}\sum_{i=1}^n \sigma_i h(\mathbf{v}_i,\mathbf{z}_i)|g_i|/\gamma\right|^2 &= \mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i)/\gamma\right|^2 \\
\leq \mathbb{E}\sup_{\mathcal{H}}\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}\frac{1}{n}\left|\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i)/\gamma\right|^2 &\leq \mathbb{E}\mathbb{E}_{\mathbf{v},\boldsymbol{\sigma}}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i)/\gamma\right|^2 \\
&= \gamma^{-2}\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i)\right|^2.
\end{aligned}$$

In summary, we have shown that

$$\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(h(\mathbf{v}_i,\mathbf{z}_i)-\mathbb{E}h(\mathbf{v}_i,\mathbf{z}_i)\right)\right|^2 \leq 4\gamma^{-2}\mathbb{E}\sup_{\mathcal{H}}\frac{1}{n}\left|\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i)\right|^2. \tag{4.13}$$

Define for $h\in\mathcal{H}$

$$Z_n(h,\mathbf{v}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g_i h(\mathbf{v}_i,\mathbf{z}_i).$$

Then (4.13) can be written as

$$\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(h(\mathbf{v}_i,\mathbf{z}_i)-\mathbb{E}h(\mathbf{v},\mathbf{z}_i)\right)\right|^2 \leq 4\gamma^{-2}\mathbb{E}\sup_{\mathcal{H}}|Z_n(h,\mathbf{v})|^2. \tag{4.14}$$

Note that, conditionally on the $\mathbf{v}_i$, Z_n is a zero-mean Gaussian process with increments bounded by the $L^2(\mathbb{P}_n)$ pseudo-norm: for all $h_1,h_2\in\mathcal{H}$

$$\begin{aligned}
\mathbb{E}_{\mathbf{v}}|Z_n(h_1,\mathbf{v})-Z_n(h_2,\mathbf{v})|^2 &= \frac{1}{n}\mathbb{E}_{\mathbf{v}}\sum_{i,j}(h_1(\mathbf{v}_i,\mathbf{z}_i)-h_2(\mathbf{v}_i,\mathbf{z}_i))(h_1(\mathbf{v}_j,\mathbf{z}_j)-h_2(\mathbf{v}_j,\mathbf{z}_j))g_ig_j \\
&= \frac{1}{n}\sum_{i,j}(h_1(\mathbf{v}_i,\mathbf{z}_i)-h_2(\mathbf{v}_i,\mathbf{z}_i))(h_1(\mathbf{v}_j,\mathbf{z}_j)-h_2(\mathbf{v}_j,\mathbf{z}_j))\mathbb{E}_{\mathbf{v}}g_ig_j \\
&= \frac{1}{n}\sum_{i=1}^n(h_1(\mathbf{v}_i,\mathbf{z}_i)-h_2(\mathbf{v}_i,\mathbf{z}_i))^2 = \|h_1-h_2\|_{2,n}^2.
\end{aligned}$$

Also, for fixed $\mathbf{v}$, Z_n has continuous sample paths in the $L^2(\mathbb{P}_n)$ pseudo-norm: if $\|h-h_k\|_{2,n}\to 0$ when $k\to\infty$ then $h_k(\mathbf{v}_i,\mathbf{z}_i)\to h(\mathbf{v}_i,\mathbf{z}_i)$ for all $i=1,\ldots,n$ and hence $Z_n(h_k,\mathbf{v})\to Z_n(h,\mathbf{v})$

for each realization of the g_i. Therefore, we can apply Theorem 3.3 of Pollard (1989): there exists an universal constant $K > 0$ such that

$$\left(\mathbb{E}_{\mathbf{v}} \sup_{\mathcal{H}} |Z_n(h, \mathbf{v})|^2\right)^{1/2} \leq K \int_0^{\Delta(\mathbf{v})} (\log D(x, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, dx, \tag{4.15}$$

where $\Delta(\mathbf{v}) = \sup_{\mathcal{H}} \|h\|_{2,n}$.

If $\|H\|_{2,n} > 0$, since by assumption $D(\varepsilon\|H\|_{2,n}, \mathcal{H}, \|.\|_{2,n}) \leq D(\varepsilon)$ for all $0 < \varepsilon < 1$, we have that

$$\int_0^1 (\log D(\varepsilon\|H\|_{2,n}, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, d\varepsilon \leq \int_0^1 (\log D(\varepsilon))^{1/2} \, d\varepsilon < \infty.$$

Also, since $\Delta(\mathbf{v})/\|H\|_{2,n} \leq 1$

$$\int_0^{\Delta(\mathbf{v})/\|H\|_{2,n}} (\log D(\varepsilon\|H\|_{2,n}, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, d\varepsilon \leq \int_0^1 (\log D(\varepsilon\|H\|_{2,n}, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, d\varepsilon$$

The change of variables $x = \varepsilon\|H\|_{2,n}$ gives

$$\begin{aligned} \|H\|_{2,n} \int_0^{\Delta(\mathbf{v})/\|H\|_{2,n}} & (\log D(\varepsilon\|H\|_{2,n}, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, d\varepsilon \\ &= \int_0^{\Delta(\mathbf{v})} (\log D(x, \mathcal{H}, \|.\|_{2,n}))^{1/2} \, dx. \end{aligned}$$

Then

$$\left(\mathbb{E}_{\mathbf{v}} \sup_{\mathcal{H}} |Z_n(h, \mathbf{v})|^2\right)^{1/2} \leq K\|H\|_{2,n} \int_0^1 (\log D(\varepsilon))^{1/2} \, d\varepsilon. \tag{4.16}$$

On the other hand, if $\|H\|_{2,n} = 0$ then $\Delta(\mathbf{v}) = 0$ and this implies that the right hand side of (4.15) is zero. In this case (4.16) holds trivially.

We have thus shown that

$$\begin{aligned} \mathbb{E} \sup_{\mathcal{H}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (h(\mathbf{v}_i, \mathbf{z}_i) - \mathbb{E}h(\mathbf{v}_i, \mathbf{z}_i)) \right|^2 &\leq 4\gamma^{-2} \mathbb{E} \sup_{\mathcal{H}} |Z_n(h, \mathbf{v})|^2 \\ &= 4\gamma^{-2} \mathbb{E}\mathbb{E}_{\mathbf{v}} \sup_{\mathcal{H}} |Z_n(h, \mathbf{v})|^2 \\ &\leq 4\gamma^{-2} K^2 \mathbb{E}\|H\|_{2,n}^2 \left(\int_0^1 (\log D(\varepsilon))^{1/2} \, d\varepsilon \right)^2 \\ &= 4\gamma^{-2} K^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E}H^2(\mathbf{v}_i, \mathbf{z}_i) \left(\int_0^1 (\log D(\varepsilon))^{1/2} \, d\varepsilon \right)^2, \end{aligned}$$

which is what we wanted to prove. Part (i) can be proved by substituting $L^1(\mathbb{P}_n)$ norms by $L^2(\mathbb{P}_n)$ in the arguments leading to (4.14) and then applying Theorem 3.2 of Pollard (1989). □

The following lemma is a key result in the proof of the consistency of the estimators.

Lemma 4.2.2. *Assume ρ is a bounded ρ-function. Consider the class of functions*

$$\mathcal{H} = \left\{ h_{s,\mathbf{b}}(u,\mathbf{x}) = \rho\left(\frac{u-\mathbf{x}^T\mathbf{b}}{s}\right) : \mathbf{b} \in \mathbb{R}^p, s > 0 \right\}.$$

Then if $p/n \to 0$

$$\sup_{\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^{n} (h(u_i,\mathbf{x}_i) - \mathbb{E}h(u,\mathbf{x}_i)) \right| \xrightarrow{P} 0.$$

Proof. We will apply the maximal inequalities of Theorem 4.2.1 to $\mathcal{H} \cup \{0\}$.

Let

$$\mathcal{L} = \left\{ l_{s,\mathbf{b}}(u,\mathbf{x}) = \frac{u-\mathbf{x}^T\mathbf{b}}{s} : \mathbf{b} \in \mathbb{R}^p,\ s > 0 \right\}.$$

Then $\mathcal{L}$ is a subset of the vector space of all linear functions in $p+1$ variables. This vector space has dimension $p+1$. It follows from Lemma 2.6.15 of Van der Vaart and Wellner (1996) that $\mathcal{L}$ has VC-index at most $p+3$.

Note that $\rho = m^1 + m^2$, where $m^1(x) = \rho(x)I\{x \geq 0\}$ and $m^2(x) = \rho(x)I\{x < 0\}$. Note that m^1 is non-decreasing and m^2 is non-increasing. By Lemma 9.9 (viii) of Kosorok (2008), $m^1 \circ \mathcal{L}$ and $m^2 \circ \mathcal{L}$ have VC-index at most $p+3$. $m^1 \circ \mathcal{L}$ and $m^2 \circ \mathcal{L}$ have a constant envelope equal to 1.

Let Q be a probability measure on $\mathbb{R}^{p+1}$ with finite support. Fix $0 < \varepsilon < 1$. By Theorem 2.6.7 from Van der Vaart and Wellner (1996), for some universal constant K we have that for $i = 1, 2$

$$N(\varepsilon, m^i(\mathcal{L}), \|.\|_{2,Q}) \leq K(p+3)(16e)^{p+3}\varepsilon^{-2(p+2)}.$$

Note that $m^1 \circ \mathcal{L} + m^2 \circ \mathcal{L}$ has constant envelope equal to 2. It is easy to show that

$$\begin{aligned} N(2\varepsilon, m^1 \circ \mathcal{L} + m^2 \circ \mathcal{L}, \|.\|_{2,Q}) &\leq N(\varepsilon/2, m^1 \circ \mathcal{L}, \|.\|_{2,Q}) N(\varepsilon/2, m^2 \circ \mathcal{L}, \|.\|_{2,Q}) \\ &\leq (K(p+3)(16e)^{p+3}(\varepsilon/2)^{-2(p+2)})^2. \end{aligned}$$

Note that $\mathcal{H}$ has envelope $H(u,\mathbf{x}) = 2$ and that $\mathcal{H} \subset m^1 \circ \mathcal{L} + m^2 \circ \mathcal{L}$. Hence

$$N(\varepsilon\|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q}) \leq (K(p+3)(16e)^{p+3}(\varepsilon/2)^{-2(p+2)})^2.$$

Furthermore $\mathcal{H} \cup \{0\}$ also has envelope H. We can assume without loss of generality that $K > 1$. Hence,

$$\begin{aligned} N(\varepsilon\|H\|_{2,Q}, \mathcal{H} \cup \{0\}, \|.\|_{2,Q}) &\leq N(\varepsilon\|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q}) + 1 \\ &\leq (K(p+3)(16e)^{p+3}(\varepsilon/2)^{-2(p+2)})^2 + 1 \\ &\leq 2(K(p+3)(16e)^{p+3}(\varepsilon/2)^{-2(p+2)})^2 \end{aligned}$$

implies that

$$\begin{aligned} D(\varepsilon\|H\|_{2,Q}, \mathcal{H}\cup\{0\}, \|.\|_{2,Q}) &\leq N((\varepsilon/2)\|H\|_{2,Q}, \mathcal{H}\cup\{0\}, \|.\|_{2,Q}) \\ &\leq D(\varepsilon) \end{aligned}$$

where

$$D(\varepsilon) = 2(K(p+3)(16e)^{p+3}\,(\varepsilon/4)^{-2(p+2)})^2.$$

It follows from Theorem 4.2.1(i) that for some fixed $C_1 > 0$

$$\begin{aligned} \mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(u_i,\mathbf{x}_i)-\mathbb{E}h(u,\mathbf{x}_i)\right)\right| &\leq \mathbb{E}\sup_{\mathcal{H}\cup\{0\}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(u_i,\mathbf{x}_i)-\mathbb{E}h(u,\mathbf{x}_i)\right)\right| \\ &\leq C_1\int_0^1 \left(\log D(\varepsilon)\right)^{1/2}. \end{aligned}$$

Note that

$$\begin{aligned} \log D(\varepsilon) &= \log 2 + 2\log(K) + 2\log(p+3) + 2(p+3)\log(16e) + 4(p+2)\log\frac{4}{\varepsilon} \\ &\leq C_2 p\,(1-\log\varepsilon) \end{aligned}$$

for some fixed $C_2 > 0$. Hence

$$\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(u_i,\mathbf{x}_i)-\mathbb{E}h(u,\mathbf{x}_i)\right)\right| \leq \sqrt{p}C_1\sqrt{C_2}\int_0^1 (1-\log\varepsilon)^{1/2}\,d\varepsilon = \sqrt{p}C_3$$

where $C_3 > 0$ is fixed. Finally, the result follows from applying Markov's inequality and the fact that by assumption $p/n \to 0$. □

Proof of Lemma 3.2.1. Let $\hat{\boldsymbol{\beta}}_S$ be the corresponding S-estimator. Then

$$s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))^2 \leq s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))^2 \leq s_n^S(\mathbf{r}(\boldsymbol{\beta}_0))^2 + \frac{\gamma_n}{n}\|\boldsymbol{\beta}_0\|_r^r \tag{4.17}$$

By Theorem 3 of Davies (1990), replacing any appeals in the proof of that theorem to Lemma 2 of Davies (1990) by appeals to Lemma 4.2.2, we have that $s_n^S(\mathbf{r}(\hat{\boldsymbol{\beta}}_S)) \xrightarrow{P} s(F_0)$. Since by [B0], $\gamma_n\|\boldsymbol{\beta}_0\|_r^r/n \to 0$, it suffices to show that $s_n^S(\mathbf{r}(\boldsymbol{\beta}_0)) = s_n^S(\mathbf{u}) \xrightarrow{P} s(F_0)$.

Fix $\varepsilon > 0$. We can find $\delta > 0$ such that

$$\mathbb{E}\rho_0\left(\frac{u}{s(F_0)-\varepsilon}\right) \geq b+\delta \text{ and } \mathbb{E}\rho_0\left(\frac{u}{s(F_0)+\varepsilon}\right) \leq b-\delta.$$

Then by the Law of Large Numbers

$$\lim \frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{u_i}{s(F_0)-\varepsilon}\right) \geq b+\delta \; a.s. \text{ and } \lim \frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{u_i}{s(F_0)+\varepsilon}\right) \leq b-\delta \; a.s.$$

Since ρ_0 is monotone in $|u|$, with probability one, for large enough n, $s(F_0)-\varepsilon \leq s_n^S(\mathbf{u}) \leq s(F_0)+\varepsilon$. In particular, $s_n^S(\mathbf{u}) \xrightarrow{P} s(F_0)$. □

The following Lemma is a simple adaptation of Lemma 1 from Davies (1990). For $v \in \mathbb{R}$, $s \in \mathbb{R}$ let

$$R(v,s) = \mathbb{E}\rho_1\left(\frac{u-v}{s}\right).$$

Lemma 4.2.3. *Assume [R1] and [F0] hold. Then*

(i) $R : \mathbb{R}\times\mathbb{R}_+ \to [0,1]$ *is continuous.*

(ii) $R(0,s) \leq R(v,s)$ *for* $v \in \mathbb{R}$, $s>0$.

(iii) $R(0,s) < \inf_{|v|\geq\eta} R(v,s)$ *for all* $\eta>0$ *and* $s>0$.

Proof. (i) follows from the fact that ρ_1 is bounded and the Bounded Convergence Theorem.

Next we prove (ii). This is roughly Lemma 3.1 of Yohai (1985). Note that for any $v \neq 0$, the distribution function R_v of $|u-v|$ satisfies: $R_v(t) \leq R_0(t)$ for all $t>0$ and there exists $\delta>0$ such that $R_v(t) < R_0(t)$ for $0<t\leq\delta$. Since $\rho_1(u/s)$ is non decreasing in $|u|$ and strictly increasing in a neighbourhood of 0, it follows that for all s, $R(v,s)$ has a unique minimum at $v=0$.

Now we prove (iii). Suppose for some $\eta, s>0$, $R(0,s) \geq \inf_{|v|\geq\eta} R(v,s)$. Note that by [R1] and [F0], $R(0,s)<1$. Take v_n with $|v_n| \geq \eta$ such that $R(v_n,s) \to \inf_{|v|\geq\eta} R(v,s)$. Note that if for some subsequence v_{n_k}, $|v_{n_k}| \to \infty$, then by the Bounded Convergence Theorem $R(v_{n_k},s) \to 1$ and hence $R(0,s) \geq 1$, leading to a contradiction. Hence v_n must be bounded. We can assume, eventually passing to a subsequence, that $v_n \to v^*$, with $|v^*| \geq \eta$. Hence $R(v^*,s) = \inf_{|v|\geq\eta} R(v,s) \leq R(0,s)$. But by (ii), $R(v,s)$ has a unique minimum at $v=0$. Hence (iii) follows. □

Proof of Theorem 3.2.2. We will prove (i). Fix $0<\alpha<1$. Note that by definition of $\hat{\boldsymbol{\beta}}_B$

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{u_i-\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B-\boldsymbol{\beta}_0)}{s_n}\right) \leq \frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{u_i}{s_n}\right)+\frac{\lambda_n}{n}\|\boldsymbol{\beta}_0\|_q^q.$$

By Lemma 4.2.2 we have that

$$\sup_{\mathbf{b}\in\mathbb{R}^p, 0<s<2s_0}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\rho_1\left(\frac{u_i-\mathbf{x}_i^T\mathbf{b}}{s}\right)-R(\mathbf{x}_i^T\mathbf{b},s)\right)\right| \xrightarrow{P} 0. \tag{4.18}$$

Since by assumption $s_n \xrightarrow{P} s_0$, Lemma 4.2.3 (i) and [B1] imply that the right hand side of the last inequality converges in probability to

$$b^* = \mathbb{E}\rho_1\left(\frac{u}{s_0}\right). \tag{4.19}$$

By Lemma 4.2.3 (ii), $R(0,s) \leq R(v,s)$ for all $v \in \mathbb{R}$, $s \in \mathbb{R}$. Then

$$R(0,s_n) \leq \frac{1}{n}\sum_{i=1}^{n} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n). \tag{4.20}$$

By Lemma 4.2.3 (i)

$$R(0,s_n) \xrightarrow{P} b^*. \tag{4.21}$$

Then, it follows from (4.18), (4.19),(4.20) and (4.21) that

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{u_i - \mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)}{s_n}\right) \xrightarrow{P} b^*$$

and

$$\frac{1}{n}\sum_{i=1}^{n} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \xrightarrow{P} b^*. \tag{4.22}$$

By (4.22), given $\delta > 0$, with arbitrarily high probability, for large enough n we have that

$$\frac{1}{n}\sum_{i=1}^{n} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \leq b^* + \delta. \tag{4.23}$$

Let $\varepsilon > 0$, we will show that with arbitrarily high probability, for large enough n, $\eta_n(\alpha)\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \leq \varepsilon$. Let $A = \{i : |\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)| \geq \varepsilon\}$ and $N = \#A$. Then

$$\frac{1}{n}\sum_{i=1}^{n} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) = \frac{1}{n}\sum_{i\in A} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) + \frac{1}{n}\sum_{i\in A^c} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n).$$

Note that

$$\frac{1}{n}\sum_{i\in A^c} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \geq \frac{n-N}{n} R(0,s_n). \tag{4.24}$$

Also, if $|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)| \geq \varepsilon$ then $R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \geq \inf_{|v|\geq\varepsilon} R(v, s_n)$. Hence

$$R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \geq R(0,s_n) + \left(\inf_{|v|\geq\varepsilon} R(v,s_n) - R(0,s_n)\right).$$

We will show that with arbitrarily high probability, for large enough n and $i \in A$

$$R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \geq R(0, s_n) + \kappa, \tag{4.25}$$

for some $\kappa = \kappa(\varepsilon) > 0$.

First, we will show that

$$\sup_v |R(v, s_n) - R(v, s_0)| \xrightarrow{P} 0 \tag{4.26}$$

Fix $u, v \in \mathbb{R}$. Let $\phi_1(t) = \psi_1(t)t$. By [R1], ϕ_1 is bounded. Applying the Mean Value Theorem we get that, for some s_n^* such that $|s_n^* - s_0| \leq |s_n - s_0|$

$$\begin{aligned}\left|\rho_1\left(\frac{u-v}{s_n}\right) - \rho_1\left(\frac{u-v}{s_0}\right)\right| &\leq \left|\psi_1\left(\frac{u-v}{s_n^*}\right)\left(\frac{u-v}{s_n^*}\right)\right|\left|\frac{s_n - s_0}{s_n^*}\right| \\ &\leq \|\phi_1\|_\infty \left|\frac{s_n - s_0}{s_n^*}\right|. \end{aligned} \tag{4.27}$$

Fix some $\eta > 0$. Since $s_n \xrightarrow{P} s_0$, with arbitrarily high probability, for large enough n, the right hand side of (4.27) is smaller than η for all u, v. (4.26) is proven.

By Lemma 4.2.3 (iii), $\inf_{|v|\geq\varepsilon} R(v, s_0) > R(0, s_0)$. Let $\eta_1 = (\inf_{|v|\geq\varepsilon} R(v, s_0) - R(0, s_0))/4$. Fix $\eta_2 > 0$. Take n_0 such that for all $n \geq n_0$, $\sup_v |R(v, s_n) - R(v, s_0)| < \eta_1/2$ with probability greater than $1 - \eta_2$. For each $n_1 \geq n_0$, take v_{n_1} with $|v_{n_1}| \geq \varepsilon$ such that

$$\inf_{|v|\geq\varepsilon} R(v, s_{n_1}) \geq R(v_{n_1}, s_{n_1}) - \eta_1/2.$$

Note that v_{n_1} is random. It follows that with probability greater than $1 - \eta_2$, for all $n_1 \geq n_0$

$$\begin{aligned}\inf_{|v|\geq\varepsilon} R(v, s_0) - \inf_{|v|\geq\varepsilon} R(v, s_{n_1}) &\leq R(v_{n_1}, s_0) - R(v_{n_1}, s_{n_1}) + \eta_1/2 \\ &\leq \sup_v |R(v, s_{n_1}) - R(v, s_0)| + \eta_1/2 < \eta_1.\end{aligned}$$

Since $R(0, s_n) \xrightarrow{P} R(0, s_0)$, with arbitrarily high probability, for large enough n

$$\begin{aligned}&\inf_{|v|\geq\varepsilon} R(v, s_n) - R(0, s_n) \\ = \inf_{|v|\geq\varepsilon} R(v, s_n) - \inf_{|v|\geq\varepsilon} R(v, s_0) + \inf_{|v|\geq\varepsilon} R(v, s_0) - R(0, s_0) + R(0, s_0) - R(0, s_n) \\ &\geq 2\eta_1.\end{aligned}$$

We have proven (4.25) for $\kappa(\varepsilon) = \left(\inf_{|v|\geq\varepsilon} R(v, s_0) - R(0, s_0)\right)/2$. Hence with arbitrarily high probability, for large enough n

$$\frac{1}{n}\sum_{i\in A} R(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0), s_n) \geq \frac{N}{n}(R(0, s_n) + \kappa)$$

and thus by (4.23), (4.24) and (4.25) with arbitrarily high probability, for large n, we have that if $N \geq (1-\alpha)n$ then

$$R(0, s_n) \leq b^* + \delta - (1-\alpha)\kappa.$$

In summary, we have shown that

$$\{N \geq (1-\alpha)n\} \subseteq \{R(0, s_n) \leq b^* + \delta - (1-\alpha)\kappa\} \cup A_n, \tag{4.28}$$

where $\mathbb{P}(A_n) \to 0$. For any given ε, we can find a sufficiently small δ such that $\delta - (1-\alpha)\kappa < 0$. Then by (4.21) and (4.28), $\mathbb{P}(N \geq (1-\alpha)n) \to 0$. Hence, with arbitrarily high probability, for sufficiently large n, $n\alpha < n - N$. In this case, there must exist $\mathcal{A} \subset \{1, ..., n\}$ with $\#\mathcal{A} = [n\alpha]$ such that $\left|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)\right| < \varepsilon$ for all $i \in \mathcal{A}$ and this implies that

$$\eta_n(\alpha)\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \leq \min_{\mathcal{A}\subset\{1,...,n\}, \#\mathcal{A}=[n\alpha]} \max_{i\in\mathcal{A}} \left|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)\right| \leq \varepsilon,$$

which is what we wanted to prove.

For (ii), note that by [B3] and since by assumption $\hat{\boldsymbol{\beta}}_{ini}$ is consistent for $\boldsymbol{\beta}_0$, its first k coordinates are bounded away from zero with probability tending to one. Hence by [B2] we have that

$$\frac{\iota_n}{n} \sum_{j=1}^{k} \frac{|\beta_{0,j}|^t}{|\hat{\beta}_{ini,j}|^\varsigma} \xrightarrow{P} 0.$$

The rest of the proof follows along the same lines as the proof of (i). □

The following lemma is needed in the proof of Theorem 3.2.3.

Lemma 4.2.4. *Assume [R2], [F0] and [X1] a) hold. Let* $0 < a < b$.

(i) For $\mathbf{x} \in \mathbb{R}^p$, *consider the class of functions*

$$\mathcal{H} = \left\{ h_s(u, \mathbf{x}) = \psi_1\left(\frac{u}{s}\right)\mathbf{x} : s \in [a, b] \right\}.$$

Then for some fixed constant $A > 0$ *that depends only on* a, b, ψ_1 *and the constant that appears in [X1] a)*

$$\mathbb{E}\sup_{\mathcal{H}} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(u_i, \mathbf{x}_i) \right\| \leq A\sqrt{p}.$$

(ii) For $\mathbf{x} \in \mathbb{R}^k$, *consider the class of functions*

$$\mathcal{H} = \left\{ h_s(u, \mathbf{x}) = \psi_1\left(\frac{u}{s}\right)\mathbf{x} : s \in [a, b] \right\}.$$

Then for some fixed constant $B > 0$ *that depends only on* a, b, ψ_1 *and the constant that appears in [X1] a)*

$$\mathbb{E}\sup_{\mathcal{H}} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(u_i, \mathbf{x}_{i,I}) \right\| \leq B\sqrt{k}.$$

Proof. We prove (i). The proof of (ii) is entirely analogous. Let

$$\mathcal{G} = \left\{ g_s(u,x) = \psi_1\left(\frac{u}{s}\right) x : s \in [a,b] \right\}.$$

Fix $1 \le j \le p$. Note that $\mathbb{E}g(u, x_{i,j}) = 0$ for $g \in \mathcal{G}$ and $i = 1, \ldots, n$. Note also that $\mathcal{G}$ has envelope $G(u,x) = \|\psi_1\|_\infty |x|$ and that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}G^2(u_i, x_{i,j}) = \|\psi_1\|_\infty^2 \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 < \infty.$$

Let Q be a probability measure on $\mathbb{R}^2$ with finite support such that $\|G\|_{2,Q} > 0$. This implies that $\|x\|_{2,Q} > 0$

Let $\phi_1(t) = t\psi_1'(t)$. By [R2], ϕ_1 is bounded. Also, if $s_1, s_2 \in [a,b]$, then by the Mean Value Theorem

$$|g_{s_1}(u,x) - g_{s_2}(u,x)| \le \|\phi_1\|_\infty \frac{1}{a}|x|\,|s_1 - s_2|.$$

Then, by Theorem 2.7.11 of Van der Vaart and Wellner (1996), for all $\varepsilon > 0$ the bracketing number of $\mathcal{G}$ satisfies

$$N_{[]}(2\varepsilon\|\phi_1\|_\infty \frac{1}{a}\||x|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \le N(\varepsilon, [a,b], |\,.\,|). \tag{4.29}$$

Note that for some constant C_1 that depends only on a and b, for all $\varepsilon > 0$

$$N(\varepsilon, [a,b], |\,.\,|) \le \frac{C_1}{\varepsilon} + 1. \tag{4.30}$$

Fix $0 < \varepsilon < 1$. It follows from (4.29) and (4.30) that

$$\begin{aligned} N(\varepsilon\|G\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) &= N(\varepsilon\|\psi_1\|_\infty \||x|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \\ &\le N_{[]}(2\varepsilon\|\psi_1\|_\infty \||x|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \\ &\le N(\frac{a\varepsilon\|\psi_1\|_\infty}{\|\phi_1\|_\infty}, [a,b], |\,.\,|) \\ &\le \frac{C_1\|\phi_1\|_\infty}{a\varepsilon\|\psi_1\|_\infty} + 1 = \frac{C_2}{\varepsilon} + 1. \end{aligned}$$

Note that $\mathcal{G} \cup \{0\}$ has envelope G, $\mathbb{E}g(u, x_{i,j}) = 0$ for $g \in \mathcal{G} \cup \{0\}$ and $i = 1, \ldots, n$, and that

$$N(\varepsilon\|G\|_{2,Q}, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \le N(\varepsilon\|G\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) + 1 \le \frac{C_2}{\varepsilon} + 2.$$

Thus

$$D(\varepsilon\|G\|_{2,Q}, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \le N(\varepsilon\|G\|_{2,Q}/2, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \le \frac{2C_2}{\varepsilon} + 2.$$

Let $D(\varepsilon) = 2C_2/\varepsilon + 2$. Then by Theorem (4.2.1)(ii), for some fixed $C_3 > 0$

$$\mathbb{E}\sup_{\mathcal{G}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(u_i, x_{i,j})\right|^2 \leq \mathbb{E}\sup_{\mathcal{G}\cup\{0\}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(u_i, x_{i,j})\right|^2$$
$$\leq C_3\left(\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2\right)\left(\int_0^1 \left(\log\left(\frac{2C_2}{\varepsilon}+2\right)\right)^{1/2} d\varepsilon\right)^2. \qquad (4.31)$$

Note that (4.31) holds for all $1 \leq j \leq p$. Then, by (4.31) and [X1]a), for some fixed $C_4 > 0$

$$\mathbb{E}\sup_{\mathcal{H}}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(u_i, \mathbf{x}_i)\right\|^2 = \mathbb{E}\sup_{s\in[a,b]}\sum_{j=1}^{p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s}\right)x_{i,j}\right|^2$$
$$\leq \sum_{j=1}^{p}\mathbb{E}\sup_{s\in[a,b]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s}\right)x_{i,j}\right|^2$$
$$\leq \sum_{j=1}^{p} C_3\left(\frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2\right)\left(\int_0^1 \left(\log\left(\frac{2C_2}{\varepsilon}+2\right)\right)^{1/2} d\varepsilon\right)^2 \leq C_4 p.$$

The result now follows from applying Jensen's inequality. □

The following lemma, which is a very simple adaptation of Lemma 3.1 of Portnoy (1984), is needed to obtain the rate of consistency of the estimators. Define

$$H_i(\mathbf{x}_i^T\boldsymbol{\beta}) = \inf\left\{\psi_1'\left(\frac{u_i - v}{s_0}\right) : |v| \leq |\mathbf{x}_i^T\boldsymbol{\beta}|\right\},$$
$$H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) = \inf\left\{\psi_1'\left(\frac{u_i - v}{s_n}\right) : |v| \leq |\mathbf{x}_i^T\boldsymbol{\beta}|\right\}.$$

Lemma 4.2.5. *Assume [R2], [F0], [X1], [X2], [X4] and [X5] hold. Assume* $(p\log n)/n \to 0$. *Then there exists* $a^* > 0$ *and* $\delta > 0$ *such that*

$$\mathbb{P}\left(\inf\left\{\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) : \|\mathbf{z}\| = 1, \|\boldsymbol{\beta}\| \leq \delta\right\} \geq a^* n\right) \to 1. \qquad (4.32)$$

Proof. Note that

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2\left(H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta})\right).$$

Hence, for any $\delta > 0$

$$\inf_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) \geq \inf_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \inf_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 \left(H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta})\right).$$

By Lemma 3.1 of Portnoy (1984), (4.32) holds when H_i^n is replaced by H_i. Hence, for some $a^* > 0$ and $\delta > 0$, for sufficiently large n, with arbitrarily high probability

$$\inf_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i(\mathbf{x}_i^T\boldsymbol{\beta}) \geq a^*.$$

We will show that

$$\sup_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 \left(H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta})\right)\right| \xrightarrow{P} 0.$$

Fix $i \leq n$, $\mathbf{z}$ with $\|\mathbf{z}\| = 1$, and $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| \leq \delta$. We will bound $\left|H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta})\right|$. Assume $H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) \geq H_i(\mathbf{x}_i^T\boldsymbol{\beta})$. By [R2], $H_i(\mathbf{x}_i^T\boldsymbol{\beta}) = \psi_1'\left((u_i - v_i^*)/s_0\right)$ for some v_i^* with $|v_i^*| \leq |\mathbf{x}_i^T\boldsymbol{\beta}|$. Then

$$\begin{aligned}\left|H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta})\right| = H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta}) &= H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - \psi_1'\left(\frac{u_i - v_i^*}{s_0}\right) \\ &\leq \psi_1'\left(\frac{u_i - v_i^*}{s_n}\right) - \psi_1'\left(\frac{u_i - v_i^*}{s_0}\right) \\ &\leq \left|\psi_1'\left(\frac{u_i - v_i^*}{s_n}\right) - \psi_1'\left(\frac{u_i - v_i^*}{s_0}\right)\right|.\end{aligned}$$

Note that by [R2], $\phi(t) = \psi_1''(t)t$ is bounded. Applying the Mean Value Theorem we get that

$$\begin{aligned}\left|\psi_1'\left(\frac{u_i - v_i^*}{s_n}\right) - \psi_1'\left(\frac{u_i - v_i^*}{s_0}\right)\right| &= \left|\psi_1''\left(\frac{u_i - v_i^*}{s_{i,n}^*}\right)\left(\frac{u_i - v_i^*}{s_{i,n}^*}\right)\right|\left|\frac{s_n - s_0}{s_{i,n}^*}\right| \\ &\leq \|\phi\|_\infty \left|\frac{s_n - s_0}{s_{i,n}^*}\right|.\end{aligned}$$

where $s_{i,n}^*$ is such that $|s_{i,n}^* - s_0| \leq |s_n - s_0|$. Note that $s_{i,n}^*$ may depend on $\boldsymbol{\beta}$, say $s_{i,n}^* = s_{i,n}^*(\boldsymbol{\beta})$. The same type of argument can be used to show that an analogous bound holds when $H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) \leq H_i(\mathbf{x}_i^T\boldsymbol{\beta})$.

Note that since $s_n \xrightarrow{P} s_0$, we have that $\sup_{\|\boldsymbol{\beta}\|\leq\delta} \max_i |s^*_{i,n}(\boldsymbol{\beta}) - s_0| \leq |s_n - s_0| \xrightarrow{P} 0$. Then

$$\begin{aligned}
&\sup_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \left| \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{z})^2 \left(H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta}) \right) \right| \\
&\leq \sup_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{z})^2 \left| H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) - H_i(\mathbf{x}_i^T\boldsymbol{\beta}) \right| \\
&\leq \sup_{\|\mathbf{z}\|=1,\|\boldsymbol{\beta}\|\leq\delta} \|\phi\|_\infty \max_i \frac{1}{\left|s^*_{i,n}(\boldsymbol{\beta})\right|} |s_n - s_0| \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{z})^2 \\
&\leq \sup_{\|\boldsymbol{\beta}\|\leq\delta} \|\phi\|_\infty \max_i \frac{1}{|s^*_{i,n}(\boldsymbol{\beta})|} |s_n - s_0| \, \tau \xrightarrow{P} 0.
\end{aligned}$$

It follows that for sufficiently large n, with arbitrarily high probability,

$$\inf \left\{ \sum_{i=1}^{n} (\mathbf{x}_i^T\mathbf{z})^2 H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) : \|\mathbf{z}\| = 1, \|\boldsymbol{\beta}\| \leq \delta \right\} \geq n(a^* - a^*/2),$$

and so the lemma is proven. □

Proof of Theorem 3.2.3. We prove (i). The proof of (ii) is very similar and is omitted.

Let

$$Z_n^2(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\frac{r_i(\boldsymbol{\beta})}{s_n} \right) + \frac{\lambda_n}{n}\sum_{j=1}^{p} |\beta_j|^q,$$

so that $\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} Z_n^2(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}_B$.

A first order Taylor expansion shows that, for some $0 \leq \zeta_i \leq 1$,

$$\begin{aligned}
0 \geq Z_n^2(\hat{\boldsymbol{\beta}}_B) - Z_n^2(\boldsymbol{\beta}_0) &= -\frac{1}{ns_n}\sum_{i=1}^{n} \psi_1 \left(\frac{u_i}{s_n} \right) \mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \\
&+ \frac{1}{2}\frac{1}{s_n^2}\frac{1}{n}\sum_{i=1}^{n} \psi_1' \left(\frac{u_i - \zeta_i(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{x}_i}{s_n} \right) (\mathbf{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0))^2 + \frac{\lambda_n}{n}\|\hat{\boldsymbol{\beta}}_B\|_q^q - \frac{\lambda_n}{n}\|\boldsymbol{\beta}_0\|_q^q \\
&\geq -\frac{1}{ns_n}\sum_{i=1}^{n} \psi_1 \left(\frac{u_i}{s_n} \right) \mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \\
&+ \frac{1}{2}\frac{1}{s_n^2}\frac{1}{n}\sum_{i=1}^{n} \psi_1' \left(\frac{u_i - \zeta_i(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{x}_i}{s_n} \right) (\mathbf{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0))^2 + \frac{\lambda_n}{n}\sum_{j=1}^{k} |\hat{\beta}_{B,j}|^q - |\beta_{0,j}|^q \\
&= A_n + B_n + C_n.
\end{aligned}$$

Since by assumption $s_n \xrightarrow{P} s_0$, by Lemma 4.2.4(i) we have that

$$A_n = \|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| O_P\left(\sqrt{\frac{p}{n}}\right). \tag{4.33}$$

Let δ and a^* be as in Lemma 4.2.5. Since $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$, for sufficiently large n, with arbitrarily high probability we have that $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| < \delta$ and by Lemma 4.2.5

$$\begin{aligned}
B_n &\geq \frac{1}{2\, s_n^2}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0))^2 \inf\left\{\psi_1'\left(\frac{u_i - v}{s_n}\right) : |v| \leq |\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)|\right\} \\
&= \frac{1}{2\, s_n^2}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0))^2 H_i^n(\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)) \\
&\geq \frac{1}{2\, s_n^2}\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\|^2 \inf_{\|\mathbf{z}\|=1, \|\boldsymbol{\beta}\| \leq \delta} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{z})^2 H_i^n(\mathbf{x}_i^T\boldsymbol{\beta}) \\
&\geq \frac{a^*}{2\, s_n^2}\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\|^2.
\end{aligned} \tag{4.34}$$

Since $\hat{\boldsymbol{\beta}}_B$ is consistent, by [B3], its first k coordinates are bounded and bounded away from zero in probability. Applying the Mean Value Theorem we get

$$|C_n| \leq O_P\left(\frac{\lambda_n}{n}\right)\sum_{j=1}^{k}|\hat{\beta}_{B,j} - \beta_{0,j}| \leq \|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| O_P\left(\frac{\lambda_n\sqrt{k}}{n}\right). \tag{4.35}$$

Hence, it follows from (4.33), (4.34) and (4.35) that with arbitrarily high probability, for large enough n and some positive constants M_1 and M_2

$$\begin{aligned}
0 &\geq A_n + B_n + C_n \\
&\geq -M_1\sqrt{\frac{p}{n}}\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0)\| - M_2\frac{\lambda_n\sqrt{k}}{n}\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| + r_n\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\|^2,
\end{aligned}$$

where $r_n \xrightarrow{P} r_0 > 0$. Hence

$$\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \leq \frac{1}{r_n}\left(M_1\sqrt{\frac{p}{n}} + M_2\left(\frac{\lambda_n\sqrt{k}}{n}\right)\right),$$

which proves the theorem. □

Proof of Theorem 3.2.4. We prove (i). The proof of (ii) is very similar and is thus omitted. We follow Lemma 2 of Huang et al. (2008). Since by Theorem 3.2.3 $\hat{\boldsymbol{\beta}}_B$ is $\sqrt{n/p}$-consistent, for a sufficiently large $C > 0$ and n, $\|\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0\| \leq C\sqrt{p/n}$ with arbitrarily high probability.

Let

$$V_n(\mathbf{u}_1,\mathbf{u}_2)=\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I}+\mathbf{u}_1\sqrt{p/n},\boldsymbol{\beta}_{0,II}+\mathbf{u}_2\sqrt{p/n})}{s_n}\right)$$
$$+\lambda_n\left(\sum_{j=1}^{k}|\beta_{0,j}+u_{1,j}\sqrt{p/n}|^q+\sum_{j=k+1}^{p}|u_{2,j-k}\sqrt{p/n}|^q\right).$$

Then for large enough n, with arbitrarily high probability, $(\hat{\boldsymbol{\beta}}_{B,I},\hat{\boldsymbol{\beta}}_{B,II})$ is obtained by minimizing $V_n(\mathbf{u}_1,\mathbf{u}_2)$ over $\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2\leq C^2$. We will show that if $\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2\leq C^2$ and $\|\mathbf{u}_2\|>0$ then, for large enough n, $V_n(\mathbf{u}_1,\mathbf{u}_2)-V_n(\mathbf{u}_1,\mathbf{0}_{p-k})>0$ with arbitrarily high probability and the theorem will follow.

Let $(\mathbf{u}_1,\mathbf{u}_2)$ satisfy $\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2\leq C^2$. It is easy to see that

$$V_n(\mathbf{u}_1,\mathbf{u}_2)-V_n(\mathbf{u}_1,\mathbf{0}_{p-k})=$$
$$\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I}+\mathbf{u}_1\sqrt{p/n},\mathbf{u}_2\sqrt{p/n})}{s_n}\right)-\rho_1\left(\frac{r_i(\boldsymbol{\beta}_{0,I}+\mathbf{u}_1\sqrt{p/n},\mathbf{0}_{p-k})}{s_n}\right)+$$
$$\lambda_n\left(\frac{p}{n}\right)^{q/2}\sum_{j=k+1}^{p}|u_{2,j-k}|^q=$$
$$(I)+(II).$$

Applying the Mean Value Theorem we get

$$(I)=(\mathbf{0}_k,\mathbf{u}_2)^T\sqrt{\frac{p}{n}}\frac{(-1)}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{r_i(\boldsymbol{\theta}_n^*)}{s_n}\right)\mathbf{x}_i,$$

where $\boldsymbol{\theta}_n^*=(\boldsymbol{\beta}_{0,I}+\mathbf{u}_1\sqrt{p/n},(1-\alpha_n)\mathbf{u}_2\sqrt{p/n})$ for some $\alpha_n\in[0,1]$. Applying the Mean Value Theorem once more we get

$$(\mathbf{0}_k,\mathbf{u}_2)^T\sqrt{\frac{p}{n}}\frac{(-1)}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{r_i(\boldsymbol{\theta}_n^*)}{s_n}\right)\mathbf{x}_i=\sqrt{\frac{p}{n}}\frac{(-1)}{s_n}(\mathbf{0}_k,\mathbf{u}_2)^T\sum_{i=1}^{n}\psi_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i+$$
$$\sqrt{\frac{p}{n}}\frac{1}{s_n^2}(\mathbf{0}_k,\mathbf{u}_2)^T\sum_{i=1}^{n}\psi_1'\left(\frac{r_i(\boldsymbol{\theta}_n^{**})}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T(\mathbf{u}_1\sqrt{p/n},(1-\alpha_n)\mathbf{u}_2\sqrt{p/n})=$$
$$\sqrt{\frac{p}{n}}\frac{(-1)}{s_n}(\mathbf{0}_k,\mathbf{u}_2)^T\sum_{i=1}^{n}\psi_1\left(\frac{r_i(\boldsymbol{\beta}_0)}{s_n}\right)\mathbf{x}_i+$$
$$\frac{p}{n}\frac{1}{s_n^2}(\mathbf{0}_k,\mathbf{u}_2)^T\sum_{i=1}^{n}\psi_1'\left(\frac{r_i(\boldsymbol{\theta}_n^{**})}{s_n}\right)\mathbf{x}_i\mathbf{x}_i^T(\mathbf{u}_1,(1-\alpha_n)\mathbf{u}_2)=$$
$$A_n+B_n,$$

where $\|\boldsymbol{\theta}_n^{**} - \boldsymbol{\beta}_0\| \leq \|\boldsymbol{\theta}_n^{*} - \boldsymbol{\beta}_0\|$. By Lemma 4.2.4(i), $A_n = \|\mathbf{u}_2\| O_P(p)$ uniformly in $\mathbf{u}_2$. We will show that $B_n = (\|\mathbf{u}_2\|^2 + \|\mathbf{u}_2\|)\, O_P(p)$ uniformly in $\mathbf{u}_2$.

$$\begin{aligned}
|B_n| &\leq \frac{p}{s_n^2}\|\psi_1'\|_\infty \frac{1}{n}\sum_{i=1}^n \left|(\mathbf{0}_k, \mathbf{u}_2)^T \mathbf{x}_i \mathbf{x}_i^T (\mathbf{u}_1, \mathbf{0}_{p-k}) + (1-\alpha_n)((\mathbf{0}_k, \mathbf{u}_2)^T \mathbf{x}_i)^2\right| \\
&\leq \frac{p}{s_n^2}\|\psi_1'\|_\infty \left(\frac{1}{n}\sum_{i=1}^n \left|(\mathbf{0}_k, \mathbf{u}_2)^T \mathbf{x}_i \mathbf{x}_i^T (\mathbf{u}_1, \mathbf{0}_{p-k})\right| + \frac{1}{n}\sum_{i=1}^n \left|(\mathbf{0}_k, \mathbf{u}_2)^T \mathbf{x}_i\right|^2\right) \\
&= \frac{p}{s_n^2}\|\psi_1'\|_\infty \left(B_{n,1} + B_{n,2}\right).
\end{aligned}$$

Applying the Cauchy-Schwartz inequality and using [X2] we get

$$B_{n,1} \leq \left(\frac{1}{n}\sum_{i=1}^n \left|(\mathbf{0}_k, \mathbf{u}_2)^T \mathbf{x}_i\right|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \left|\mathbf{x}_i^T (\mathbf{u}_1, \mathbf{0}_{p-k})\right|^2\right)^{1/2} \leq \|\mathbf{u}_2\| \tau^{1/2} \|\mathbf{u}_1\| \tau^{1/2} \leq \|\mathbf{u}_2\| C\tau.$$

Also,

$$B_{n,2} \leq \tau \|\mathbf{u}_2\|^2.$$

Hence $B_n = (\|\mathbf{u}_2\|^2 + \|\mathbf{u}_2\|)\, O_P(p)$. On the other hand

$$\lambda_n \left(\frac{p}{n}\right)^{q/2} \sum_{j=k+1}^{p} |u_{2,j-k}|^q = \lambda_n \left(\frac{p}{n}\right)^{q/2} \|\mathbf{u}_2\|_q^q,$$

Note also that $\|\mathbf{u}_2\|_q^q \geq \|\mathbf{u}_2\|^q$. Hence, for some $M_1, M_2 > 0$ and sufficiently large n, for all $(\mathbf{u}_1, \mathbf{u}_2)$ that satisfy $\|\mathbf{u}_1\|^2 + \|\mathbf{u}_2\|^2 \leq C^2$, with arbitrarily high probability, we have that

$$\begin{aligned}
V_n(\mathbf{u}_1, \mathbf{u}_2) - V_n(\mathbf{u}_1, \mathbf{0}_{p-k}) &> -M_1 p\|\mathbf{u}_2\| - M_2 p\|\mathbf{u}_2\|^2 - M_2 p\|\mathbf{u}_2\| + \lambda_n \left(\frac{p}{n}\right)^{q/2} \|\mathbf{u}_2\|^q \\
&= \|\mathbf{u}_2\|^q p(-M_1\|\mathbf{u}_2\|^{1-q} - M_2\|\mathbf{u}_2\|^{2-q} - M_2\|\mathbf{u}_2\|^{1-q} + \lambda_n \frac{p^{q/2-1}}{n^{q/2}}) \\
&\geq \|\mathbf{u}_2\|^q p(-M_1 C^{1-q} - M_2 C^{2-q} - M_2 C^{1-q} + \lambda_n \frac{p^{q/2-1}}{n^{q/2}}). \quad (4.36)
\end{aligned}$$

Finally, since by [X8] $\lambda_n n^{-q/2}/p^{1-q/2} \to \infty$, we have that for sufficiently large n, for any non-zero $\mathbf{u}_2$, the right hand side of (4.36) is strictly positive.

□

The following lemmas are needed in the proof of Theorem 3.2.5.

Lemma 4.2.6. *Assume [R2], [F0], [X1], [X2], [X3] and [X10] hold. Let* $\mathbf{a}_n \in \mathbb{R}^k$, $\|\mathbf{a}_n\| = 1$. *Let* $r_n^2 = \mathbf{a}_n^T \boldsymbol{\Sigma}_{1,n}^{-1} \mathbf{a}_n$. *Then*

a)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi_1\left(\frac{u_i}{s_n}\right)-\psi_1\left(\frac{u_i}{s_0}\right)\right)\left(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\right)\xrightarrow{P}0.$$

b)

$$\frac{1}{r_n\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_0}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\xrightarrow{d}N\left(0,\mathbb{E}\psi_1^2\left(\frac{u}{s_0}\right)\right).$$

Proof. We first prove a). For $t\in[0,1]$ let

$$G_n(t)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{0.5s_0+ts_0}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}.$$

Since by assumption $s_n\xrightarrow{P}s_0$, it suffices to show that $(G_n)_n$ is a tight sequence in $C[0,1]$. By Theorem 12.3 of Billingsley (1968), it suffices to show that

(i) $G_n(0)$ is tight

(ii) There exists $\gamma\geq 0$, $\alpha>1$ and a nondecreasing, continuous function f on $[0,1]$, such that for any $0\leq t_1\leq t_2\leq 1$ and any $\lambda>0$ we have

$$\mathbb{P}\left(|G_n(t_2)-G_n(t_1)|\geq\lambda\right)\leq\frac{1}{\lambda^\gamma}\left(f(t_2)-f(t_1)\right)^\alpha \text{ for all } n.$$

We first prove (i). Let

$$h_n^2=\mathbb{E}\psi_1^2\left(\frac{u}{0.5s_0}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{a}_n.$$

By [X1], [X3] and Lemma 3.1.1, $\inf_n\rho_{1,n}>0$. This together with [X2] implies that h_n and $1/h_n$ are bounded. Note that since ψ_1 is odd and the errors have a symmetric distribution, $\mathbb{E}\psi_1\left(u/(0.5s_0)\right)=0$. Also,

$$\sum_{i=1}^{n}\mathbb{E}\left(\frac{1}{\sqrt{n}}\psi_1\left(\frac{u}{0.5s_0}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\right)^2=h_n^2.$$

Note that by [X10] $\max_{i\leq n}(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I})/(\sqrt{n}h_n)\to 0$. Then for any fixed $\varepsilon>0$,

$$\sum_{i=1}^{n}\mathbb{E}\left(\frac{1}{\sqrt{n}h_n}\psi_1\left(\frac{u}{0.5s_0}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\right)^2 I\left\{\left|\psi_1\left(\frac{u}{0.5s_0}\right)(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I})/(\sqrt{n}h_n)\right|>\varepsilon\right\}$$
$$\to 0.$$

Hence, by the Lindberg-Feller Theorem, $G_n(0)/h_n \xrightarrow{d} N(0,1)$ and (i) follows. Note that roughly the same argument proves b).

Now, we prove (ii). By Tchebyshev's inequality, it suffices to show that there exists $K > 0$ such that for all t_1, t_2 in $[0,1]$, $\mathbb{E}(G_n(t_1) - G_n(t_2))^2 \leq K(t_2 - t_1)^2$ for all n. Let

$$\Delta_i(t_1, t_2) = \psi_1\left(\frac{u_i}{0.5s_0 + t_1 s_0}\right) - \psi_1\left(\frac{u_i}{0.5s_0 + t_2 s_0}\right).$$

Note that $\mathbb{E}\Delta_i(t_1, t_2) = 0$ for all t_1, t_2 and i. Using the independence of $u_1, \ldots, u_n$, we get

$$\begin{aligned}
\mathbb{E}(G_n(t_1) - G_n(t_2))^2 &= \frac{1}{n}\sum_{i,j}\mathbb{E}\Delta_i(t_1,t_2)\Delta_j(t_1,t_2)(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I})(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{j,I}) \\
&= \mathbb{E}\Delta_1(t_1,t_2)^2\frac{1}{n}\sum_{i=1}^{n}(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I})^2 \\
&= \mathbb{E}\left(\psi_1\left(\frac{u}{0.5s_0 + t_1 s_0}\right) - \psi_1\left(\frac{u}{0.5s_0 + t_2 s_0}\right)\right)^2 \mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{a}_n. \qquad (4.37)
\end{aligned}$$

Let $\phi_1(t) = \psi_1'(t)t$. By [R2] ϕ_1 is bounded. Applying the Mean Value Theorem we get that

$$\begin{aligned}
&\left|\psi_1\left(\frac{u}{0.5s_0 + t_1 s_0}\right) - \psi_1\left(\frac{u}{0.5s_0 + t_2 s_0}\right)\right| \\
&= \left|\psi_1'\left(\frac{u}{0.5s_0 + t^* s_0}\right)\left(\frac{u}{0.5s_0 + t^* s_0}\right)\left(\frac{s_0}{0.5s_0 + t^* s_0}\right)(t_1 - t_2)\right| \\
&\leq 2\|\phi_1\|_\infty |(t_1 - t_2)|,
\end{aligned}$$

where t^* lies between t_1 and t_2.

Hence, for some fixed constant $C > 0$

$$\mathbb{E}\left(\psi_1\left(\frac{u}{0.5s_0 + t_1 s_0}\right) - \psi_1\left(\frac{u}{0.5s_0 + t_2 s_0}\right)\right)^2 \leq C(t_2 - t_1)^2.$$

Hence, since $\inf_n \rho_{1,n} > 0$, from (4.37) it follows that (ii) holds and thus the lemma is proven. □

Lemma 4.2.7. *Assume [R2], [F0] and [X1] a) hold. Let $0 < a < b$. Then for some fixed constant $A > 0$ that depends only on a, b, ψ_1' and the constant that appears in [X1] a)*

$$\mathbb{E}\sup_{s\in[a,b]}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi_1'\left(\frac{u_i}{s}\right) - \mathbb{E}\psi_1'\left(\frac{u}{s}\right)\right)\mathbf{x}_{i,I}\mathbf{x}_{i,I}^T\right\|_F \leq A\sqrt{k}\max_{i\leq n}\|\mathbf{x}_{i,I}\|,$$

where $\|.\|_F$ is the Frobenius norm.

Proof. Let

$$\mathcal{G} = \left\{ g_s(u, x_1, x_2) = \psi_1' \left(\frac{u}{s} \right) x_1 x_2 : s \in [a, b] \right\}.$$

Fix $1 \leq j,\ l \leq k$. Note that $\mathcal{G}$ has envelope $G(u, x_1, x_2) = \|\psi_1'\|_\infty |x_1 x_2|$ and that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} G^2(u_i, x_{i,j}, x_{i,l}) = \|\psi_1'\|_\infty^2 \frac{1}{n} \sum_{i=1}^{n} |x_{i,j} x_{i,l}|^2 < \infty.$$

Let Q be a probability measure on $\mathbb{R}^3$ with finite support such that $\|G\|_{2,Q} > 0$. This implies that $\||x_1 x_2|\|_{2,Q} > 0$.

Let $\phi_1(t) = t\psi_1''(t)$. By [R2], ϕ_1 is bounded. Also, if $s_1, s_2 \in [a, b]$, then by the Mean Value Theorem

$$|g_{s_1}(u, x_1, x_2) - g_{s_2}(u, x_1, x_2)| \leq \|\phi_1\|_\infty \frac{1}{a} |x_1 x_2| \, |s_1 - s_2| \, .$$

Then, by Theorem 2.7.11 of Van der Vaart and Wellner (1996), for all $\varepsilon > 0$ the bracketing number of $\mathcal{G}$ satisfies

$$N_{[\,]}(2\varepsilon \|\phi_1\|_\infty \frac{1}{a} \||x_1 x_2|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \leq N(\varepsilon, [a, b], |\,.\,|). \tag{4.38}$$

Note that for some constant C_1 that depends only on a and b, for all $\varepsilon > 0$

$$N(\varepsilon, [a, b], |\,.\,|) \leq \frac{C_1}{\varepsilon} + 1. \tag{4.39}$$

Fix $0 < \varepsilon < 1$. It follows from (4.38) and (4.39) that

$$\begin{aligned} N(\varepsilon \|G\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) &= N(\varepsilon \|\psi_1'\|_\infty \||x_1 x_2|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \leq N_{[\,]}(2\varepsilon \|\psi_1'\|_\infty \||x_1 x_2|\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) \\ &\leq N(\frac{a\varepsilon \|\psi_1'\|_\infty}{\|\phi_1\|_\infty}, [a, b], |\,.\,|) \leq \frac{C_1 \|\phi_1\|_\infty}{a\varepsilon \|\psi_1'\|_\infty} + 1 = \frac{C_2}{\varepsilon} + 1. \end{aligned}$$

Note that $\mathcal{G} \cup \{0\}$ has envelope G and that

$$N(\varepsilon \|G\|_{2,Q}, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \leq N(\varepsilon \|G\|_{2,Q}, \mathcal{G}, \|.\|_{2,Q}) + 1 \leq \frac{C_2}{\varepsilon} + 2,$$

which implies that

$$D(\varepsilon \|G\|_{2,Q}, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \leq N(\varepsilon \|G\|_{2,Q}/2, \mathcal{G} \cup \{0\}, \|.\|_{2,Q}) \leq \frac{2C_2}{\varepsilon} + 2.$$

Let $D(\varepsilon) = 2C_2/\varepsilon + 2$. By Theorem (4.2.1)(ii), for some fixed constants $C_3, C_4 > 0$

$$\begin{aligned}
&\mathbb{E} \sup_{\mathcal{G}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(u_i, x_{i,j}, x_{i,l}) - \mathbb{E} g(u, x_{i,j}, x_{i,l}) \right) \right|^2 \\
&\leq \mathbb{E} \sup_{\mathcal{G} \cup \{0\}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(g(u_i, x_{i,j}, x_{i,l}) - \mathbb{E} g(u, x_{i,j}, x_{i,l}) \right) \right|^2 \\
&\leq C_3 \left(\frac{1}{n} \sum_{i=1}^{n} (x_{i,l} x_{i,j})^2 \right) \left(\int_0^1 \left(\log \left(\frac{2C_2}{\varepsilon} + 2 \right) \right)^{1/2} d\varepsilon \right)^2 \\
&\leq C_4 \frac{1}{n} \sum_{i=1}^{n} (x_{i,l} x_{i,j})^2. \qquad (4.40)
\end{aligned}$$

Note that (4.40) holds for all $1 \leq j, l \leq k$. Then, by (4.40) and [X1] a), for some fixed $C_5 > 0$

$$\begin{aligned}
&\mathbb{E} \sup_{s \in [a,b]} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\psi_1' \left(\frac{u_i}{s} \right) - \mathbb{E} \psi_1' \left(\frac{u}{s} \right) \right) \mathbf{x}_{i,I} \mathbf{x}_{i,I}^T \right\|_F^2 \\
&= \mathbb{E} \sup_{s \in [a,b]} \sum_{j=1}^{k} \sum_{l=1}^{k} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\psi_1' \left(\frac{u_i}{s} \right) - \mathbb{E} \psi_1' \left(\frac{u}{s} \right) \right) x_{i,j} x_{i,l} \right|^2 \\
&\leq \sum_{j=1}^{k} \sum_{l=1}^{k} \mathbb{E} \sup_{s \in [a,b]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\psi_1' \left(\frac{u_i}{s} \right) - \mathbb{E} \psi_1' \left(\frac{u}{s} \right) \right) x_{i,j} x_{i,l} \right|^2 \\
&\leq \sum_{j=1}^{k} \sum_{l=1}^{k} C_4 \frac{1}{n} \sum_{i=1}^{n} x_{i,l}^2 x_{i,j}^2 \leq C_5 k \max_{i \leq n} \|\mathbf{x}_{i,I}\|^2
\end{aligned}$$

The result now follows from applying Jensen's inequality. □

Proof of Theorem 3.2.5. We prove (ii). The proof of (i) is entirely analogous. For $\boldsymbol{\theta} \in \mathbb{R}^k$ let $\mathbf{p}'(\boldsymbol{\theta}) = t \sum_{j=1}^{k} sgn(\theta_j) |\theta_j|^{t-1} / |\hat{\beta}_{ini,j}|^{\varsigma} \mathbf{e}_j$. Note that by Theorem 3.2.2, $\hat{\boldsymbol{\beta}}_A$ is consistent for $\boldsymbol{\beta}_0$ and hence, by [B3], with probability tending to one all the coordinates of $\hat{\boldsymbol{\beta}}_{A,I}$ stay away from zero for a sufficiently large n. Also, by Theorem 3.2.4, $\hat{\boldsymbol{\beta}}_{A,II} = \mathbf{0}_{p-k}$ with probability tending to one. Then for large enough n, with arbitrarily high probability the partial derivatives for the first k coordinates of the objective function used to define the estimator at $\hat{\boldsymbol{\beta}}_A$ exist, and

hence

$$\begin{aligned}\mathbf{0}_k &= \frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_A}{s_n}\right)\mathbf{x}_{i,I}+\frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})\\ &= \frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{y_i-\mathbf{x}_{i,I}^T\hat{\boldsymbol{\beta}}_{A,I}}{s_n}\right)\mathbf{x}_{i,I}+\frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})+\mathbf{b}_n,\end{aligned}$$

where $\mathbb{P}(\mathbf{b}_n=\mathbf{0}_k)\to 1$. Then the Mean Value Theorem gives

$$\mathbf{0}_k=\frac{1}{\sqrt{n}}\frac{-1}{s_n}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{x}_{i,I}+\frac{1}{s_n^2}\mathbf{W}_n\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I}-\boldsymbol{\beta}_{0,I})+\frac{\iota_n}{\sqrt{n}}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})+\mathbf{b}_n,$$

where

$$\mathbf{W}_n=\frac{1}{n}\sum_{i=1}^{n}\psi_1'\left(\frac{u_i-\zeta_i\mathbf{x}_{i,I}^T(\hat{\boldsymbol{\beta}}_{A,I}-\boldsymbol{\beta}_{0,I})}{s_n}\right)\mathbf{x}_{i,I}\mathbf{x}_{i,I}^T$$

and $0\le\zeta_i\le 1$.

Let

$$\begin{aligned}\mathbf{W}_n^1&=\frac{1}{n}\sum_{i=1}^{n}\psi_1'\left(\frac{u_i}{s_n}\right)\mathbf{x}_{i,I}\mathbf{x}_{i,I}^T,\\ \mathbf{W}_n^2&=\mathbb{E}\psi_1'\left(\frac{u}{s_n}\right)\boldsymbol{\Sigma}_{1,n},\end{aligned}$$

where the expectation in $\mathbb{E}\psi_1'(u/s_n)$ is taken only with respect to u.

Then

$$\begin{aligned}\sqrt{n}\mathbf{a}_n^T(\hat{\boldsymbol{\beta}}_{A,I}-\boldsymbol{\beta}_{0,I}) &= \frac{s_n}{\mathbb{E}\psi_1'(u/s_n)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\\ &-\frac{s_n^2}{\mathbb{E}\psi_1'(u/s_n)}\frac{\iota_n}{\sqrt{n}}\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})\\ &-\frac{s_n^2}{\mathbb{E}\psi_1'(u/s_n)}\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{b}_n\\ &-\frac{1}{\mathbb{E}\psi_1'(u/s_n)}\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}(\mathbf{W}_n-\mathbf{W}_n^1)\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I}-\boldsymbol{\beta}_{0,I})\\ &-\frac{1}{\mathbb{E}\psi_1'(u/s_n)}\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}(\mathbf{W}_n^1-\mathbf{W}_n^2)\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I}-\boldsymbol{\beta}_{0,I})\\ &=\frac{s_n}{\mathbb{E}\psi_1'(u/s_n)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I}\\ &+A_n+B_n+C_n+D_n.\end{aligned}$$

We will show that $A_n + B_n + C_n + D_n = o_P(1)$. Note that by [B3], and since $\hat{\boldsymbol{\beta}}_A$ and the initial estimator $\hat{\boldsymbol{\beta}}_{ini}$ are consistent, their first k coordinates are bounded and stay away from zero with probability tending to one. Hence, with probability tending to one

$$\begin{aligned}|\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})| \leq \left(\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{a}_n\right)^{1/2}\left(\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})\right)^{1/2} &\leq \frac{1}{\rho_{1,n}}\|\mathbf{p}'(\hat{\boldsymbol{\beta}}_{A,I})\| \\ &= \frac{t}{\rho_{1,n}}\left(\sum_{j=1}^{k}\frac{|\hat{\beta}_{A,j}|^{2(t-1)}}{|\hat{\beta}_{ini,j}|^{2\varsigma}}\right)^{1/2} \\ &\leq \frac{t}{\rho_{1,n}}\sqrt{k}\frac{(b_0/2)^{t-1}}{(b_0/2)^{\varsigma}}\end{aligned}$$

Note that by [R2] and the Bounded Convergence Theorem, $\mathbb{E}\psi_1'(u/s_n) \xrightarrow{P} \mathbb{E}\psi_1'(u/s_0)$. Note also that by [X1], [X3] and Lemma 3.1.1, $\inf_n \rho_{1,n} > 0$. Hence by [X7], $A_n = o_P(1)$.

Note that $\mathbb{P}(B_n = 0) \geq \mathbb{P}(\mathbf{b}_n = \mathbf{0}_k)$. Hence $\mathbb{P}(B_n = 0) \to 1$, so that $B_n = o_P(1)$.

For a matrix $\mathbf{W}$ let $\|\mathbf{W}\|$ be its spectral norm and let $\|\mathbf{W}\|_F$ be its Frobenius norm. Recall that for any $\mathbf{W}$, $\|\mathbf{W}\| \leq \|\mathbf{W}\|_F$. We will show that $\|\mathbf{W}_n - \mathbf{W}_n^1\| = o_P(1/\sqrt{k})$ and $\|\mathbf{W}_n^1 - \mathbf{W}_n^2\| = o_P(1/\sqrt{k})$. Take $\mathbf{b} \in \mathbb{R}^k$ with $\|\mathbf{b}\| = 1$. Then, applying the Mean Value Theorem, we get

$$\begin{aligned}&|\mathbf{b}^T(\mathbf{W}_n - \mathbf{W}_n^1)\mathbf{b}| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\psi_1'\left(\frac{u_i}{s_n}\right) - \psi_1'\left(\frac{u_i - \zeta_i\mathbf{x}_{i,I}^T(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I})}{s_n}\right)\right|(\mathbf{b}^T\mathbf{x}_{i,I})^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\frac{\|\psi_1''\|_\infty}{s_n}|\mathbf{x}_{i,I}^T(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I})|(\mathbf{b}^T\mathbf{x}_{i,I})^2 \leq \frac{\|\psi_1''\|_\infty}{s_n}\max\|\mathbf{x}_{i,I}\|\|\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}\|\tau.\end{aligned}$$

Since $\|\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}\| = O_P(\sqrt{k/n})$, taking supremum over $\mathbf{b}$, from [X10] it follows that

$$\|\mathbf{W}_n - \mathbf{W}_n^1\| = o_P\left(\frac{1}{\sqrt{k}}\right),$$

and hence we have that $C_n = o_P(1)$. By Lemma 4.2.7 and [X10]

$$\|\mathbf{W}_n^1 - \mathbf{W}_n^2\|_F = O_P\left(\frac{\sqrt{k}}{\sqrt{n}}\max\|\mathbf{x}_{i,I}\|\right) = o_P\left(\frac{1}{\sqrt{k}}\right)$$

and hence we have that $D_n = o_P(1)$.

We have thus shown that $A_n + B_n + C_n + D_n = o_P(1)$ and so it follows that

$$\sqrt{n}\mathbf{a}_n^T(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) = \frac{s_n}{\mathbb{E}\psi_1'(u/s_n)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I} + o_P(1).$$

Note that r_n and $1/r_n$ are bounded. By Lemma 4.2.6

$$r_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1\left(\frac{u_i}{s_n}\right)\mathbf{a}_n^T\boldsymbol{\Sigma}_{1,n}^{-1}\mathbf{x}_{i,I} \xrightarrow{d} N(0, a(\psi_1)).$$

It follows from Slutzky's Theorem that

$$\sqrt{n}r_n^{-1}\mathbf{a}_n^T(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{d} N\left(0, s_0^2\frac{a(\psi_1)}{b(\psi_1)^2}\right).$$

□

4.3 Resumen del Capítulo 4

En este capítulo se encuentran las pruebas de los resultados originales de esta tesis. En la Sección 4.1 probamos los resultados enunciados en el Capítulo 2. En la Sección 4.2 probamos los resultados enunciados en el Capítulo 3.

Las demostraciones de la Sección 4.1 están basadas mayormente en las técnicas de Yohai (1987), Kim and Pollard (1990) Knight and Fu (2000) y Huang et al. (2008). En particular, los Teoremas 2.2.6 y 2.2.7, inspirados en resultados análogos de Knight and Fu (2000), se basan fuertemente en los Teoremas 2.3 y 2.7 de Kim and Pollard (1990). El Teorema 2.3 de Kim and Pollard (1990) da condiciones suficientes para la convergencia débil de elementos aleatorios del espacio de funciones localmente acotadas con la topología de la convergencia sobre compactos. Aplicar este teorema a cierta sucesión de procesos requiere, esencialmente, verificar la convergencia finito-dimensional de la sucesión a cierto proceso límite y la equicontinuidad asintótica estocástica de la sucesión de procesos. Por otro lado, el Teorema 2.7 de Kim and Pollard (1990) es una especie de Teorema de la Aplicación Continua para el operador arg min. Este nos permite caracterizar la distribución límite del minimizador de un proceso como aquella del minimizador del limíte del proceso.

Las demostraciones de la Sección 4.2 hacen un uso extensivo de las herramientas de la teorí de procesos empíricos que aparecen en Pollard (1989) y Van der Vaart and Wellner (1996). Los resultados de Pollard (1989), en particular las desigualdades maximales de su Teorema 4.2, están pensadas para ser aplicadas en un contexto en el que se tienen vectores aleatorios independientes e idénticamente distribuidos. Siendo que en el Capítulo 3 consideramos modelos de regresión lineal con variables predictivas fijas, no es posible aplicar directamente las técnicas de Pollard (1989) para obtener desigualdades maximales que nos resulten útiles. Por esto, en el Teorema 4.2.1, generalizamos el Teorema 4.2 de Pollard (1989) para que sea aplicable a nuestra situación. A continuación, enunciamos el Teorema 4.2.1.

Teorema. *Sean* $\mathbf{z}_1, \ldots, \mathbf{z}_n$ *vectores fijos en* $\mathbb{R}^d$. *Sean* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *vectores aleatorios i.i.d. en* $\mathbb{R}^m$. *Sea* $\mathcal{H}$ *una clase de funciones definidas en* $\mathbb{R}^{m+d}$ *y tomando valores en* $\mathbb{R}$. *Supongamos que* $\mathcal{H}$ *tiene una envolvente* H *que cumple que* $1/n \sum_{i=1}^{n} \mathbb{E}H^2(\mathbf{v}_i, \mathbf{z}_i) < \infty$. *Supongamos que* $\mathcal{H}$ *contiene a la función cero. Supongamos que para cierta función decreciente* $D(\varepsilon)$ *que satisfcae* $\int_0^1 (\log D(\varepsilon))^{1/2} d\varepsilon < \infty$, *se cumple que para todo* $0 < \varepsilon < 1$ *y para cualquier medidad de probabilidad* Q *en* $\mathbb{R}^{m+d}$ *con soporte finito tal que* $\|H\|_{2,Q} > 0$

$$D(\varepsilon \|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q}) \leq D(\varepsilon),$$

donde $D(\varepsilon \|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q})$ *es el número de capacidad de* $\mathcal{H}$.

Entonces

(i)

$$\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-\mathbb{E}h(\mathbf{v},\mathbf{z}_i)\right)\right|$$
$$\leq M\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}H^2(\mathbf{v}_i,\mathbf{z}_i)\right)^{1/2}\left(\int_0^1\left(\log D(\varepsilon)\right)^{1/2}d\varepsilon\right),$$

(ii)

$$\mathbb{E}\sup_{\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(h(\mathbf{v}_i,\mathbf{z}_i)-\mathbb{E}h(\mathbf{v},\mathbf{z}_i)\right)\right|^2$$
$$\leq M\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}H^2(\mathbf{v}_i,\mathbf{z}_i)\left(\int_0^1\left(\log D(\varepsilon)\right)^{1/2}d\varepsilon\right)^2,$$

donde $M>0$ es una constante universal fija.

Como la dimensión del paŕametro vectorial a estimar en un modelo de regresión con un número de parámetros que diverge, p_n, depende del tamaño de muestra n, las cotas para la entropía de muchas de las clases de funciones consideradas en este capítulo también dependen en general de n. Es aquí donde surgen muchas de la restricciones para la tasa a la cual puede crecer p_n.

Chapter 5

Bibliography

Alfons, A. (2014). robustHD: Robust methods for high-dimensional data. R package version 0.5.0. `http://CRAN.R-project.org/package=robustHD`.

Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional data sets. *Annals of Applied Statistics* **7** 226-248.

Avella-Medina, M. (2016). Robust penalized M-estimators for generalized linear and additive models. Ph.D. thesis, University of Geneva.

Bai Z.D and Wu, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models. I. Scale-Dependent Case. *Journal of Multivariate Analyisis* **51** 211-239.

Bai Z.D and Wu, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models. I. Scale-Invariant Case. *Journal of Multivariate Analyisis* **51** 240-251.

Billingsley, A. (1968). *Convergence of Probability Measures*. Wiley, New York.

Buhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York.

Davies P.L. (1990). Asymptotics of S-estimators in the linear model. *Annals of Statistics* **18** 1651-1675.

Davies P.L. and Gather, U. (2005). Breakdown and groups. *Annals of Statistics* **34** 1577-579.

Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr., eds.) 157-184. Wadsworth, Belmont, Calif.

Donoho, D.L. and Montanari, A. (2015). High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*. Available online at `link.springer.com/content/pdf/10.1007/s00440-015-0675-z.pdf`.

Donoho, D.L. and Montanari, A. (2015). Variance Breakdown of Huber (M)-estimators: $n/p \to m \in (1, \infty)$. Available online at `http://arxiv.org/abs/1503.02106`.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis* **71** 1054-1063.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407-499.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results. Available online at `http://arxiv.org/abs/1311.2445`.

El Karoui, N., Bean, D., Bickel, P.J., Limb, C. and Yu. B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*. **110** 14557–14562.

Fan J. and Li R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **96** 1348-1360.

Fan J. and Peng H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32** 928-961.

Fasano M.V., Maronna R.A., Sued R.M. and Yohai V.J. (2012). Continuity and differentiability of regression M functionals. *Bernoulli* **18** 1289-1309.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some common chemometrics regression tools (with discussion). *Technometrics* **35** 109-148.

Gijbels, I. and Vrinssen, I. (2015). Robust nonnegative garrote variable selection in linear regression. *Computational Statistics and Data Analysis* **85** 1-22.

Hampel, F.R. (1975). Beyond location parameters: robust concepts and methods (with discussion). *Bull. ISI* **46** 375-391

Hastie, T. and Efron, B. (2013). lars: Least Angle Regression, Lasso and Forward Stagewise. R package version 1.2. `http://CRAN.R-project.org/package=lars`.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for Nonorthogonal problems. *Technometrics* **8** 27-51.

Hössjer, O. (1992). On the optimality of S-estimators. *Statistics and Probability Letters* **14** 413-419.

Huang, J. and Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. IMS Lecture Notes-Monograph Series. Asymptotics: Particles, Processes and Inverse Problems **55** 146-166.

Huang, J., Horowitz J.L. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36** 587-613.

Huang, J., Ma, S. and Zhang, C. (2008). Adaptive Lasso for sparse high-dimensional linear models. *Statistica Sinica* **18** 1603-1618.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35** 73-101.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics* **1** 799-821.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Janssens, K., Deraedt, I., Freddy, A. and Veekman, J. (1998). Composition of 15–17th Century Archeological Glass Vessels Excavated in Antwerp, Belgium. *Mikrochimica Acta* **15** 253–267.

Johnson, B. and Peng, L. (2008). Rank-based variable selection. *Journal of Nonparametric Statistics* **20** 241-252.

Khan, J.A, Van Aelst, S. and Zamar, R.H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association* **102** 1289-1299.

Kim, J. and Pollard, D. (1990). Cube Root Asymptotics. *Annals of Statistics* **18** 191-219.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics* **28** 1356-1378.

Konis, K., Maechler, M., Marazzi, A., Maronna, R., Martin, D.R., Rocke D., Salibian-Barrera, M., Wang, J., Yohai V.J., Zamar R. and Zivot E. (2014). robust: Robust Library. R package version 0.4-16. `http://CRAN.R-project.org/package=robust`.

Kosorok, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica* **20** 167-181.

Li, G., Peng, H. and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica* **21** 391-419.

Loh, P. (2015). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Available online at `http://arxiv.org/abs/1501.00312`.

Mammen E. (1988). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Annals of Statistics* **17** 382-400.

Maronna, R.A. (2011). Robust Ridge Regression for High-Dimensional Data. *Technometrics* **53** 44-53.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

Maronna, R.A. and Yohai, V.J. (2015). High finite-sample efficiency and robustness based on distance-constrained maximum likelihood. *Computational Statistics and Data Analysis* **83** 262-274.

Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of location and dispersion of high-dimensional datasets. *Technometrics* **44** 307–317.

Nevo, D. and Ritov, Y. (2016). On Bayesian robust regression with diverging number of predictors. Available online at `http://arxiv.org/abs/1507.02074`.

Ollerer, V., Croux, C. and Alfons, A. (2014). The influence function of penalized regression estimators. *Statistics: A Journal of Theoretical and Applied Statistics* **49** 741-765.

Pollard D. (1989). Asymptotics via empirical processes. *Statistical Science* **4** 341-366.

Portnoy, S. (1984). Asymptotic Behavior of M-Estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* **12** 1298-1309.

Portnoy, S. (1985). Asymptotic Behavior of M-Estimators of p regression parameters when p^2/n is large. II. Normal Approximation. *Annals of Statistics* **13** 1403-1417.

Revolution Analytics and Weston, S. (2013). foreach: Foreach looping construct for R. R package version 1.4.1. `http://CRAN.R-project.org/package=foreach`.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79** 871-880.

Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series* (J. Franke, W. Hardle and D. Martin, eds.). *Lecture Notes in Statistics* **26** 256-272. Springer, New York.

Salibian-Barrera, M. (2006). The asymptotics of MM-estimators for linear regression with fixed designs. *Metrika* **63** 283-294.

Smucler E. and Yohai, V.J. (2015). Highly Robust and Highly Finite Sample Efficient Estimators for the Linear Model. In *Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja* (Nordhausen, K. and Taskinen, S., eds.) 91-108. Springer, New York.

Smucler E. and Yohai, V.J. (2015). Robust and sparse estimators for linear regression models. Available online at `http://arxiv.org/pdf/1508.01967.pdf`.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B* **58** 267-288.

Van der Vaart, A. W. and Wellner , J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econ. Statis.* **25** 347-355.

Wang, L. and Li, R. (2009). Weighted Wilcoxon-type Smoothly Clipped Absolute Deviation Method. *Biometrics* **65** 564-571.

Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association* **108** 632-643.

Welsh, A.H. (1989). On M-processes and M-estimation. *Annals of Statistics* **17** 337-361.

Yohai, V.J. (1985). High Breakdown Point and High Efficiency Robust Estimates for Regression. Technical Report No.66, Department of Statistics, University of Washington, Seattle, Washington, USA. Available at `http://www.stat.washington.edu/research/reports/1985/tr066.pdf`.

Yohai, V.J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *Annals of Statistics* **15** 642-656.

Yohai, V.J. and Maronna, R.A. (1979). Asymptotic Behavior of M-Estimators for the Linear Model. *Annals Statistics* **7** 258-268.

Yohai, V.J. and Zamar, R.H. (1986). High breakdown point estimates of regression by means of the minimization of an efficient scale. Technical Report No.84, Department of Statistics, University of Washington, Seattle, Washington, USA. Available at `https://www.stat.washington.edu/research/reports/1986/tr084.pdf`.

Yohai, V.J. and Zamar, R.H. (1988). High breakdown point estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* **83** 406-413.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association* **101** 1418-1429.

Zou, H. and Yuan, M. (2006). Composite quantile regression and the oracle model selection theory. *Annals of Statistics* **36** 1108-1126.

Zou, H. and Zhang, H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* **37** 1733-1751.