School of Mathematics and Statistics, UNSW Sydney[1]

LTCI, Télécom ParisTech, Université Paris Saclay[2]

CREST, Ensai, Université Bretagne Loire[3]

Australasian Applied Statistics Conference 2018

A notion of depth for curve data

Pierre Lafaye de Micheaux[1], Pavlo Mozharovskyi[2] and Myriam Vimond[3]

lafaye@unsw.edu.au

Office 2050, The Red Centre, Centre Wing, Kensington

December 4, 2018

Outline of the talk

New Statistical Tools to Study Heritability of the Brain
(Great data ... new challenges)

Australian Statistical Conference in conjunction with
the Institute of Mathematical Statistics Annual Meeting

Sydney, July 10, 2014

with B. Liquet, P. Sachdev, A. Thalamuthu, and W. Wen.

Quality of brain fibres can impact quality of life

White matter (WM) comprises long myelinated **axonal fibres** generally regarded as passive routes connecting several grey matter regions to permit flow of information across them (brain networks).

- Elucidation of the **genes involved in WM integrity** may clarify the relationship between WM development and atrophy (e.g., Leukoaraiosis), or between WM integrity and age-related decline and disease (e.g., Alzheimer [Teipel et al., 2014]).
- This may help to suggest novel **preventative** (modification of environmental factors, if no genes are involved) or **treatment** (gene therapy) strategies for WM degeneration [Kanchibhotla et al., 2013].

OATS study

We will use the **O**ld **A**ustralian **T**win **S**tudy (OATS) [Sachdev et al., 2009] data set, that was built by members of the **C**entre for **H**ealthy **B**rain **A**geing (CHeBA), here in Sydney : `http://cheba.unsw.edu.au`.

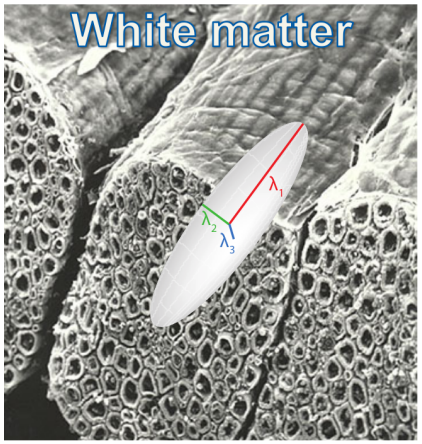
The OATS cohort was aged 65–88 at baseline (now has 3 waves of data over 4 years). The variables measured on the **twins** are : **Zygosity**, Age, Sex, Scanner information, **MRI measures**, **genetic information**, etc.

We want to rely the **genetic information** to some **brain charactetistics**.

New hot field of NeuroImaging Genetics !

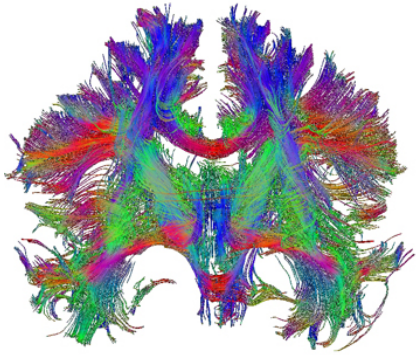
Let us first start by introducing neuroimaging concepts !

Diffusion MRI or Diffusion Tensor Imaging (DTI)



Water molecular diffusion in white matter in the brain is not free due to obstacles (fibres = neural axons). Water will diffuse more rapidly in the direction aligned with the internal structure, and more slowly as it moves perpendicular to the preferred direction.

In the diffusion tensor model, the (random vector of) water molecules' displacement (diffusion) $\mathbf{x} \in \mathbb{R}^3$ at voxel k (with center $\boldsymbol{\mu}_k$) follows a $\mathcal{N}_3(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ law. The convention is to call $D = \boldsymbol{\Sigma}/2$ the **diffusion tensor**, which is estimated at each voxel in the image from the available MR images.

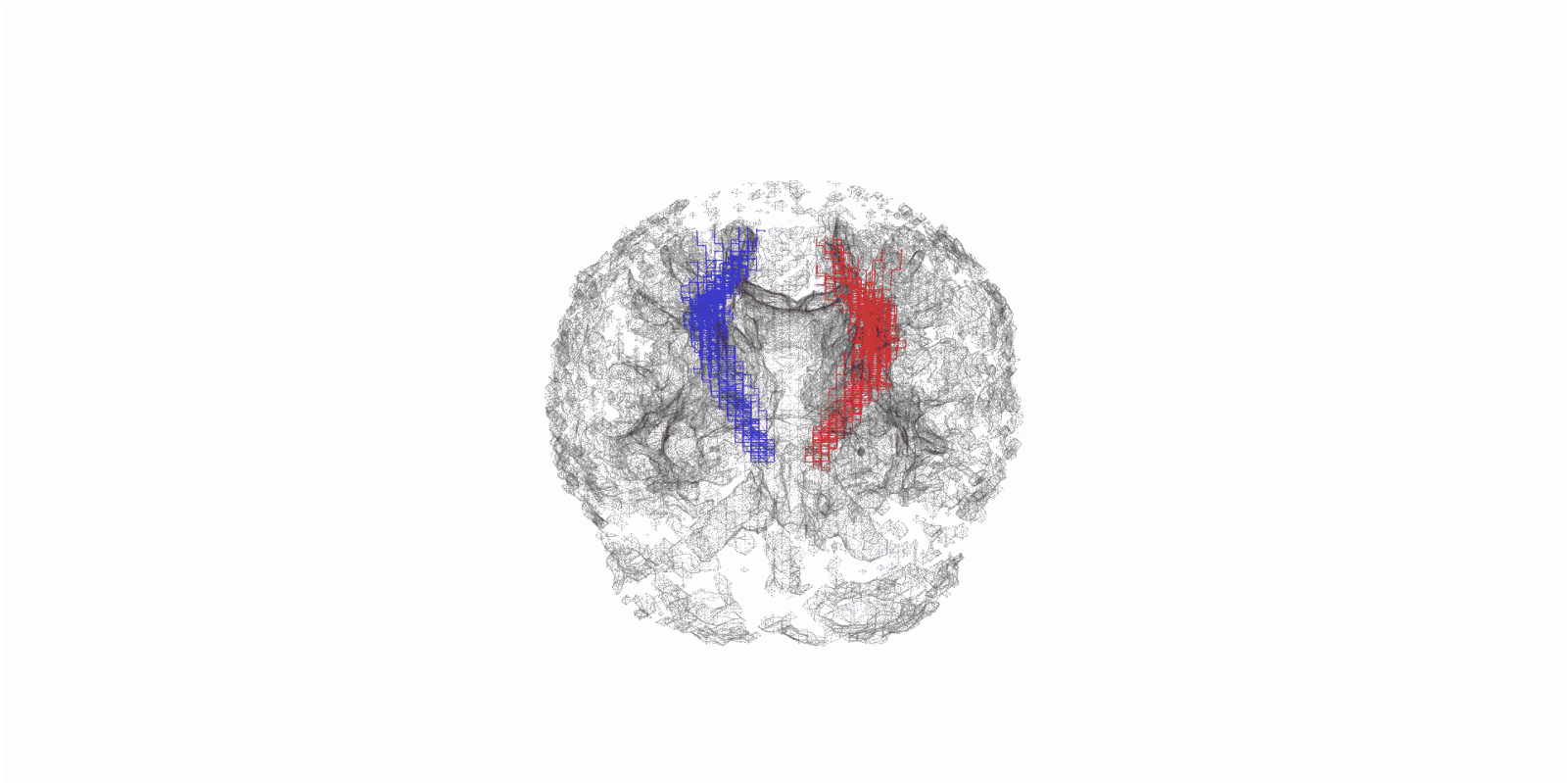


The principal direction of the diffusion tensor (first eigenvector of D) can be used to infer the **white-matter connectivity** of the brain (i.e., tractography = fibre tracking).

Studying the heritability of the CerebroSpinal Tract (CST)

Main fibre tract of the brain (from brainstem to motor cortex).

Visualization of fibres data set using our script `rgl-fibres.R`



What sort of modelling can we use for these data ?

The Halfspace Depth : Centrality of a Point

[Tukey, 1975] introduced the notion of *depth a point* w.r.t. a multivariate dataset, which can be extended to the depth of a point w.r.t. to a probability distribution.

Let Q be a distribution on $\mathbb{R}^d$.
The halfspace depth of $x \in \mathbb{R}^d$ with respect to Q is

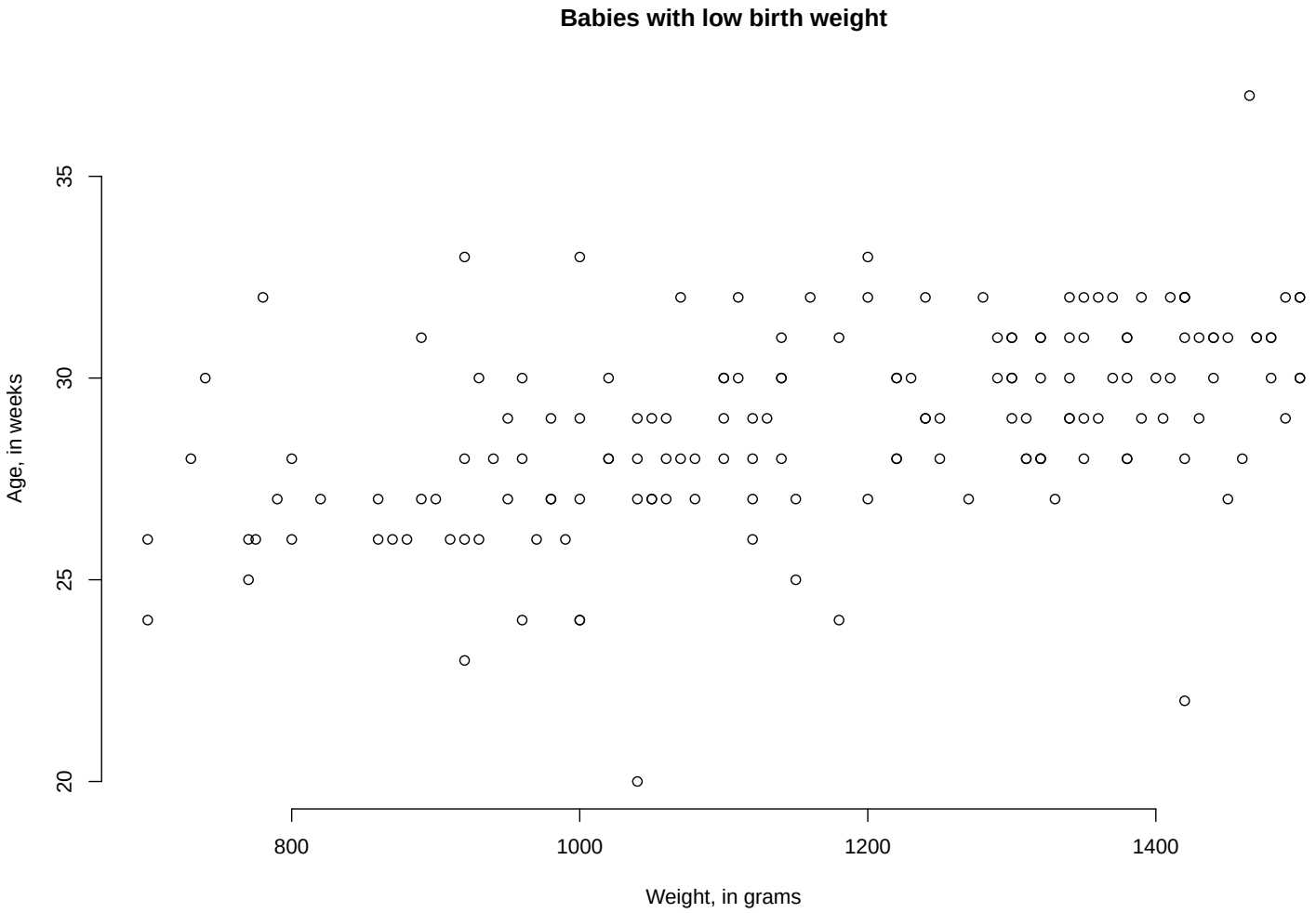
$$\begin{aligned} D(x|Q) &= \inf\{Q(H),\ x \in H \text{ closed halfspace}\} \\ &= \inf\{Q(H_{u,x}),\ u \in \mathcal{S}\} \end{aligned}$$

where $H_{u,x} = \{y \in \mathbb{R}^d :\ y^\top u \geq x^\top u\}$, $\mathcal{S}$ is the unit sphere of $\mathbb{R}^d$.

Let $\mathbb{X}_m = (X_1, \ldots, X_m)$ be an i.i.d. sample of Q.
The halfspace depth of $x \in \mathbb{R}^d$ with respect to $\mathbb{X}_m$ is

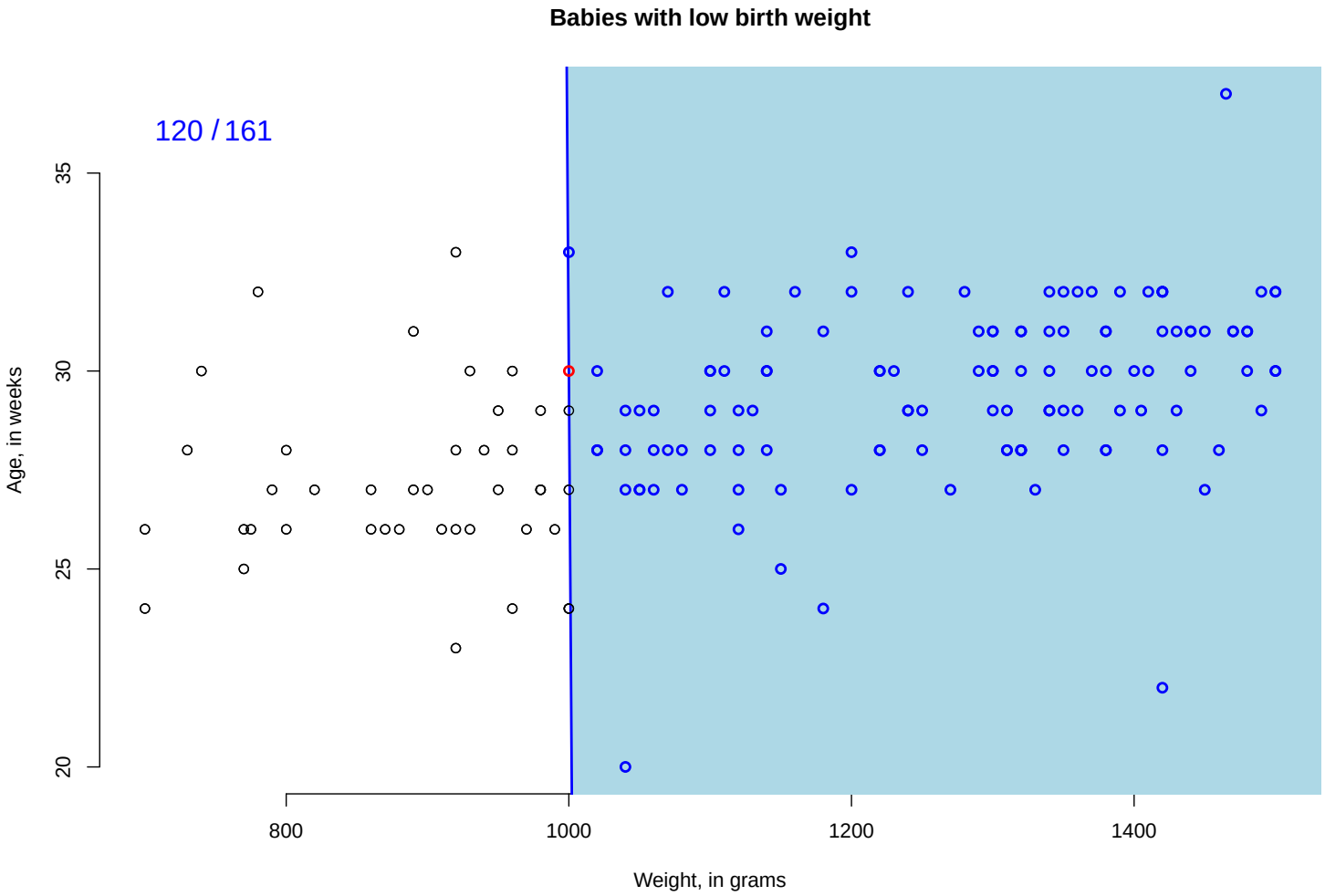
$$D(x|\mathbb{X}_m) = \inf\left\{\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{X_i^\top u \geq x^\top u},\ u \in \mathcal{S}\right\}.$$

Halfspace data depth

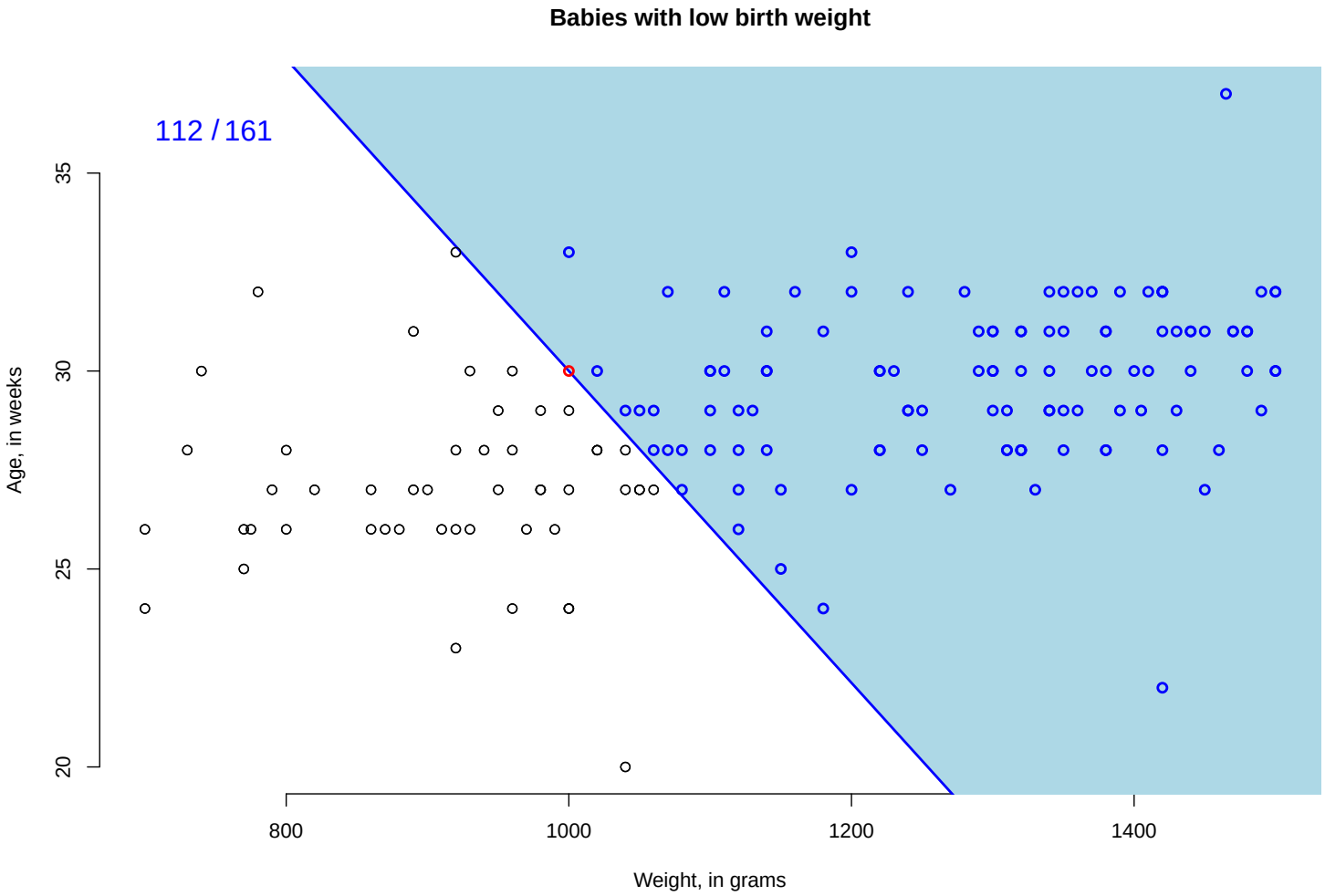


Halfspace data depth

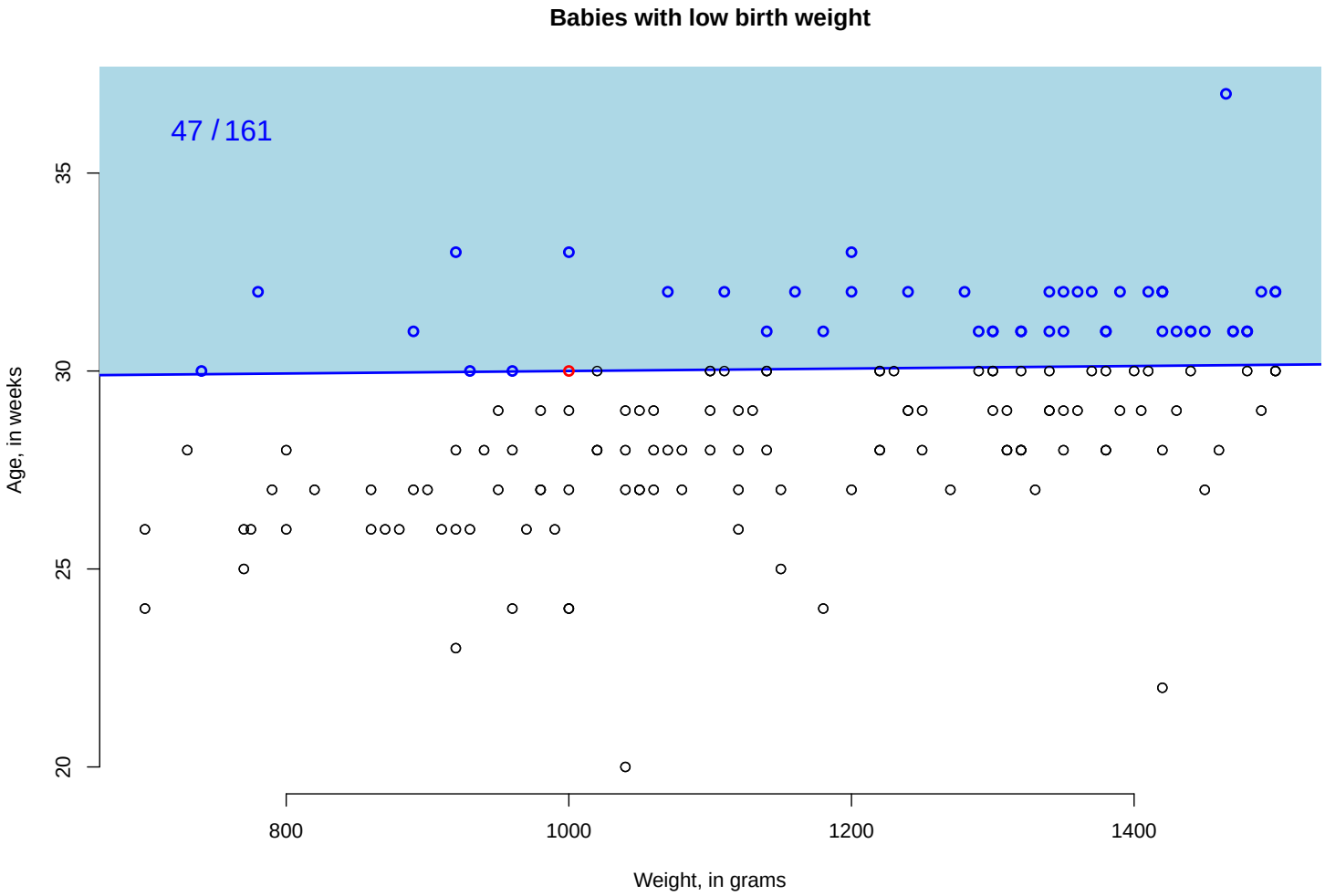
Halfspace data depth



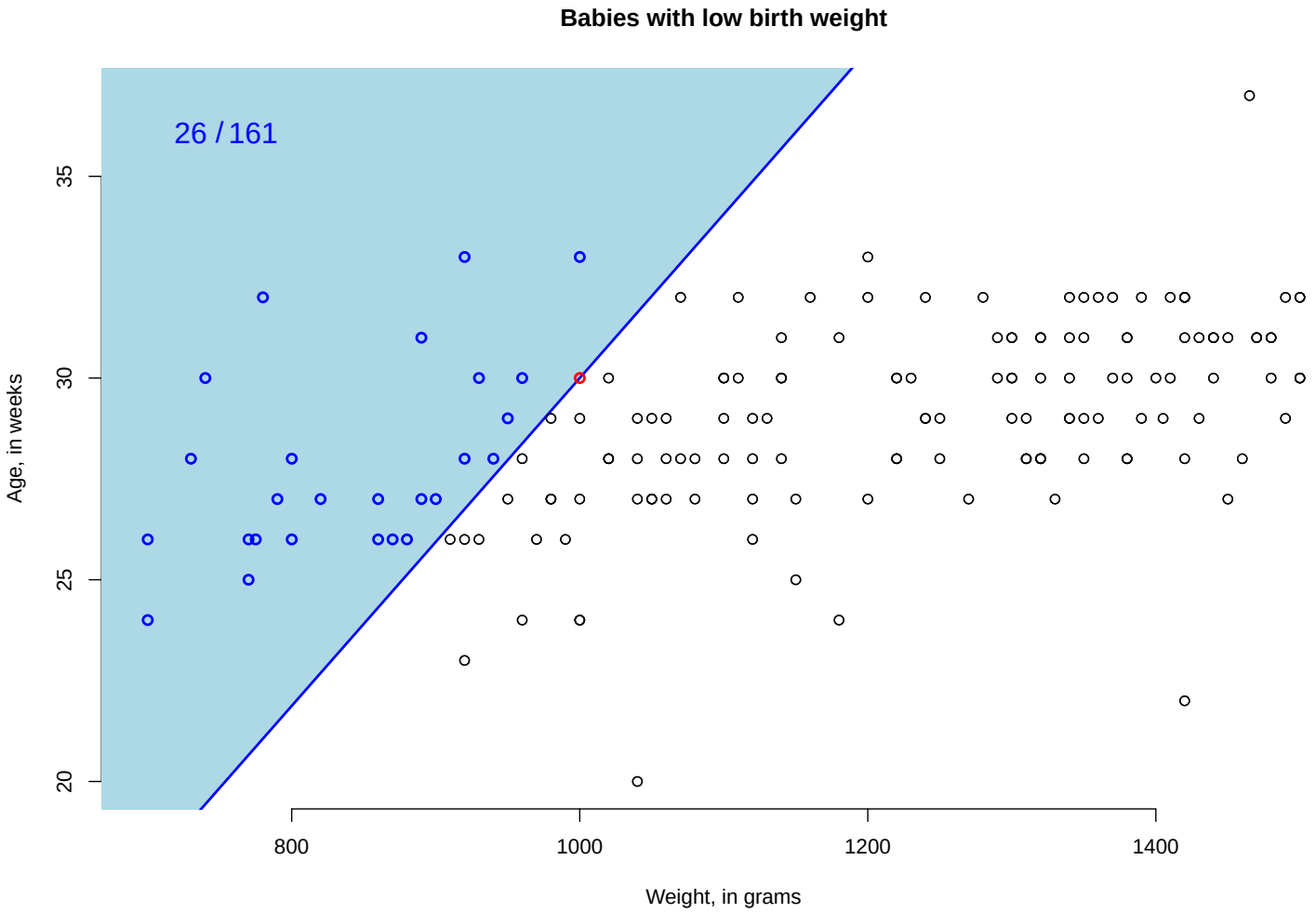
Halfspace data depth



Halfspace data depth



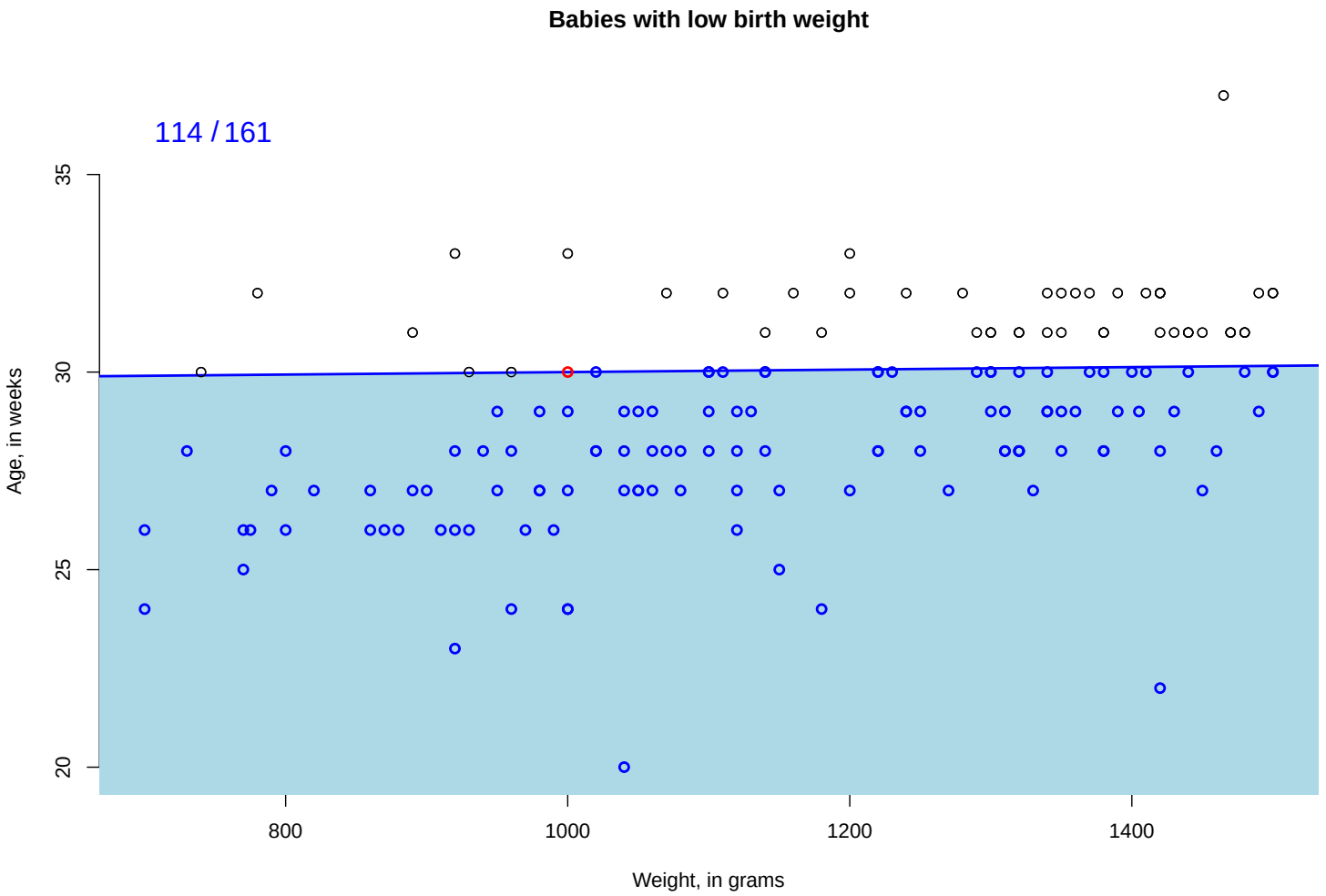
Halfspace data depth



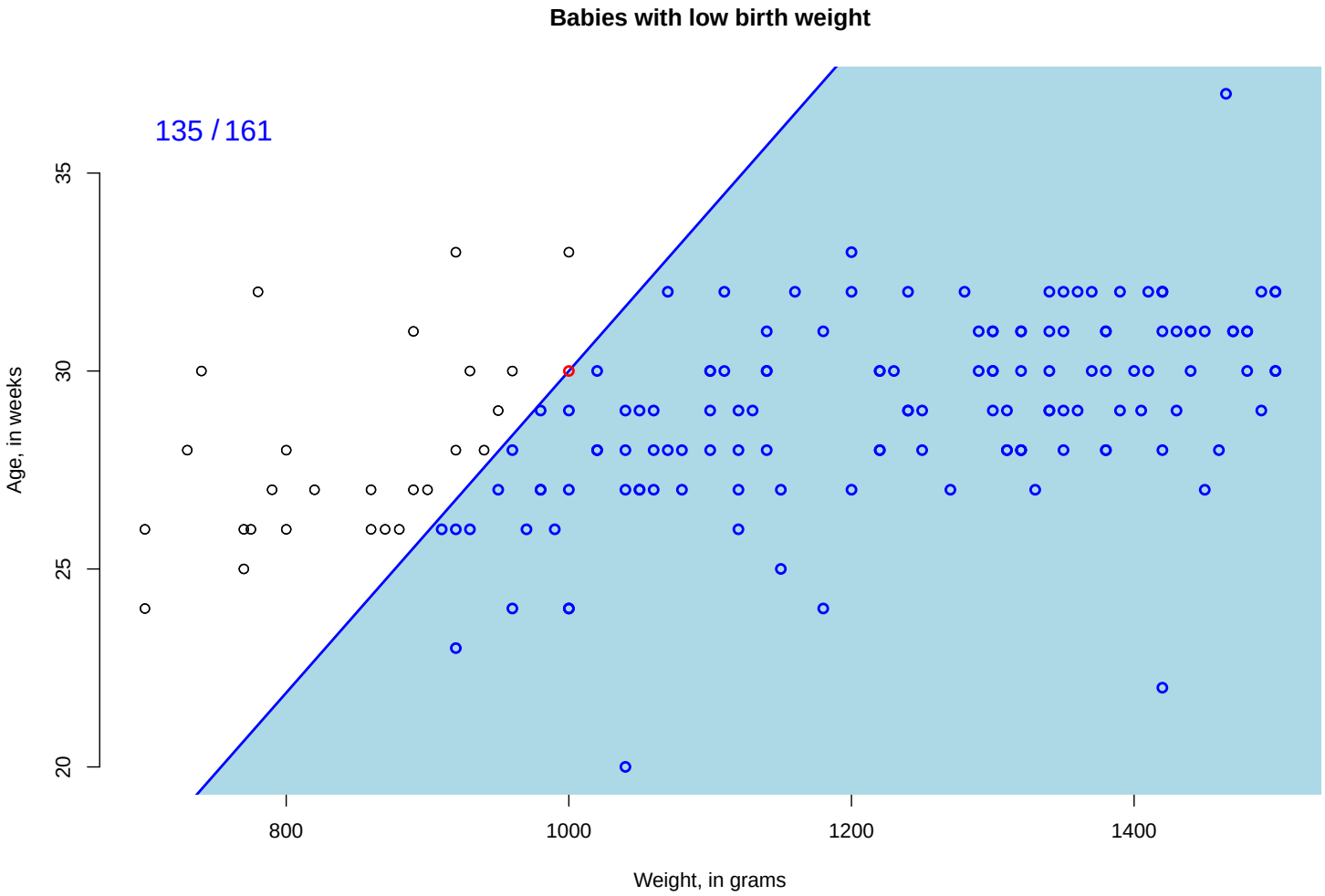
Halfspace data depth

Halfspace data depth

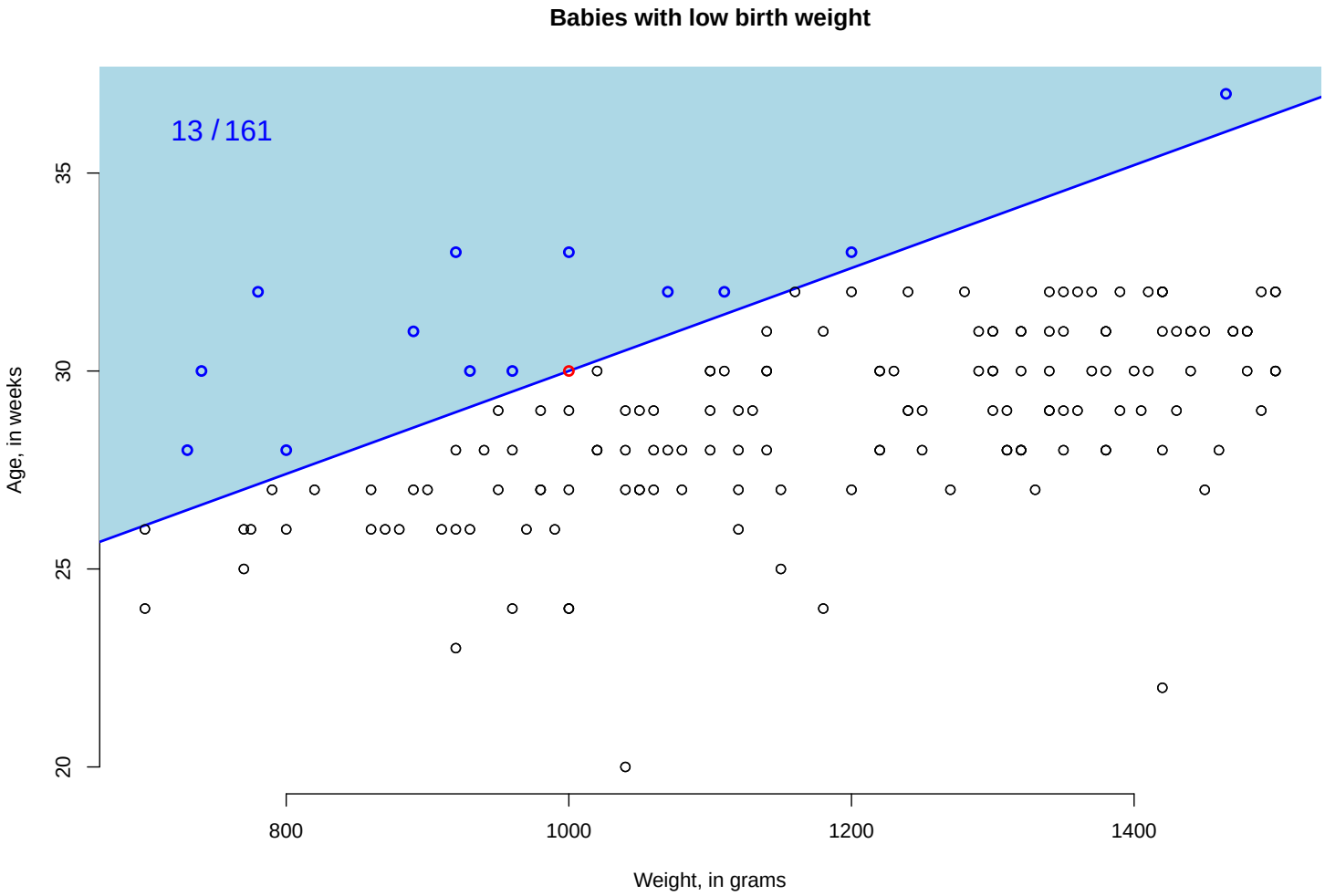
Halfspace data depth



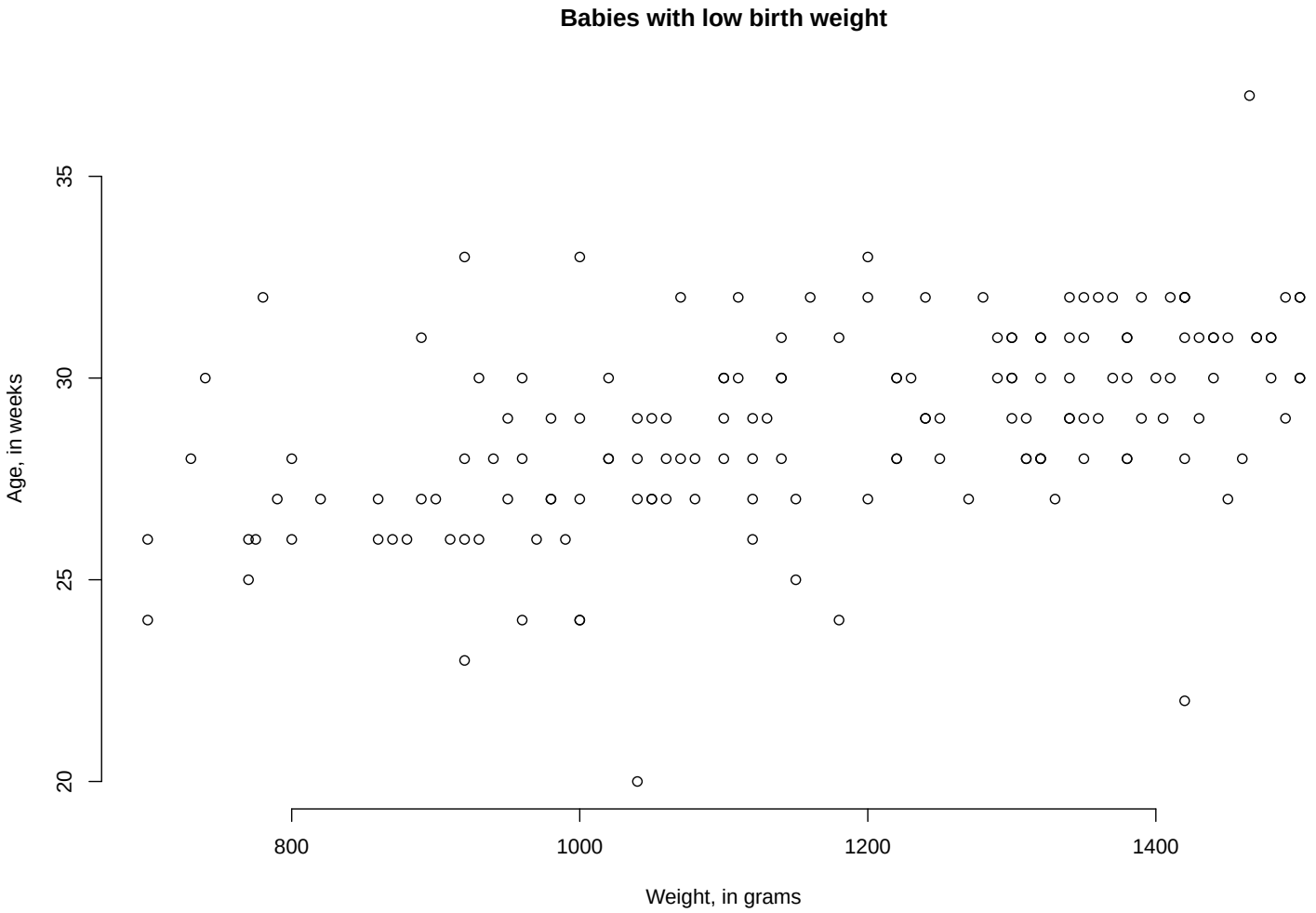
Halfspace data depth



Halfspace data depth

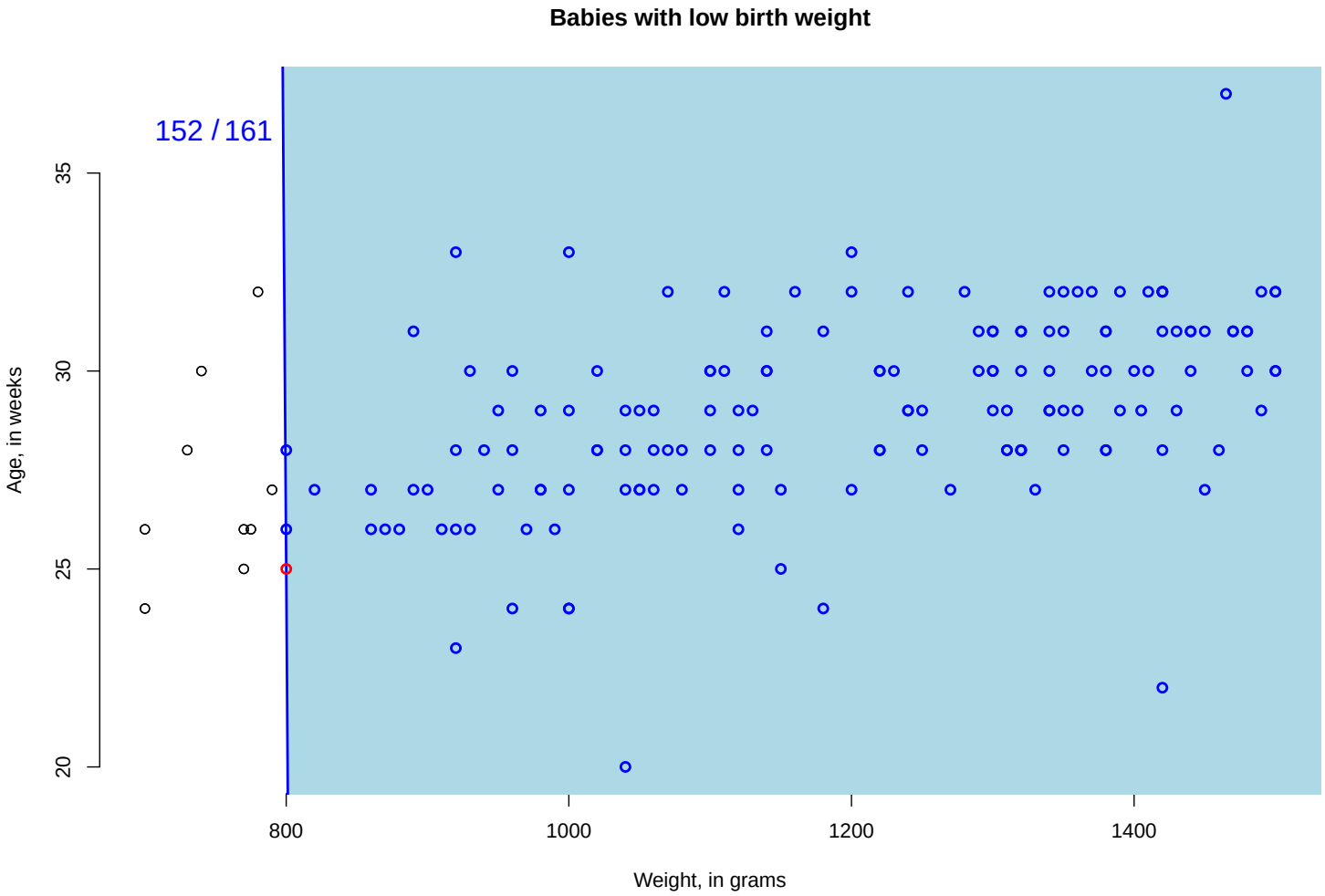


Halfspace data depth

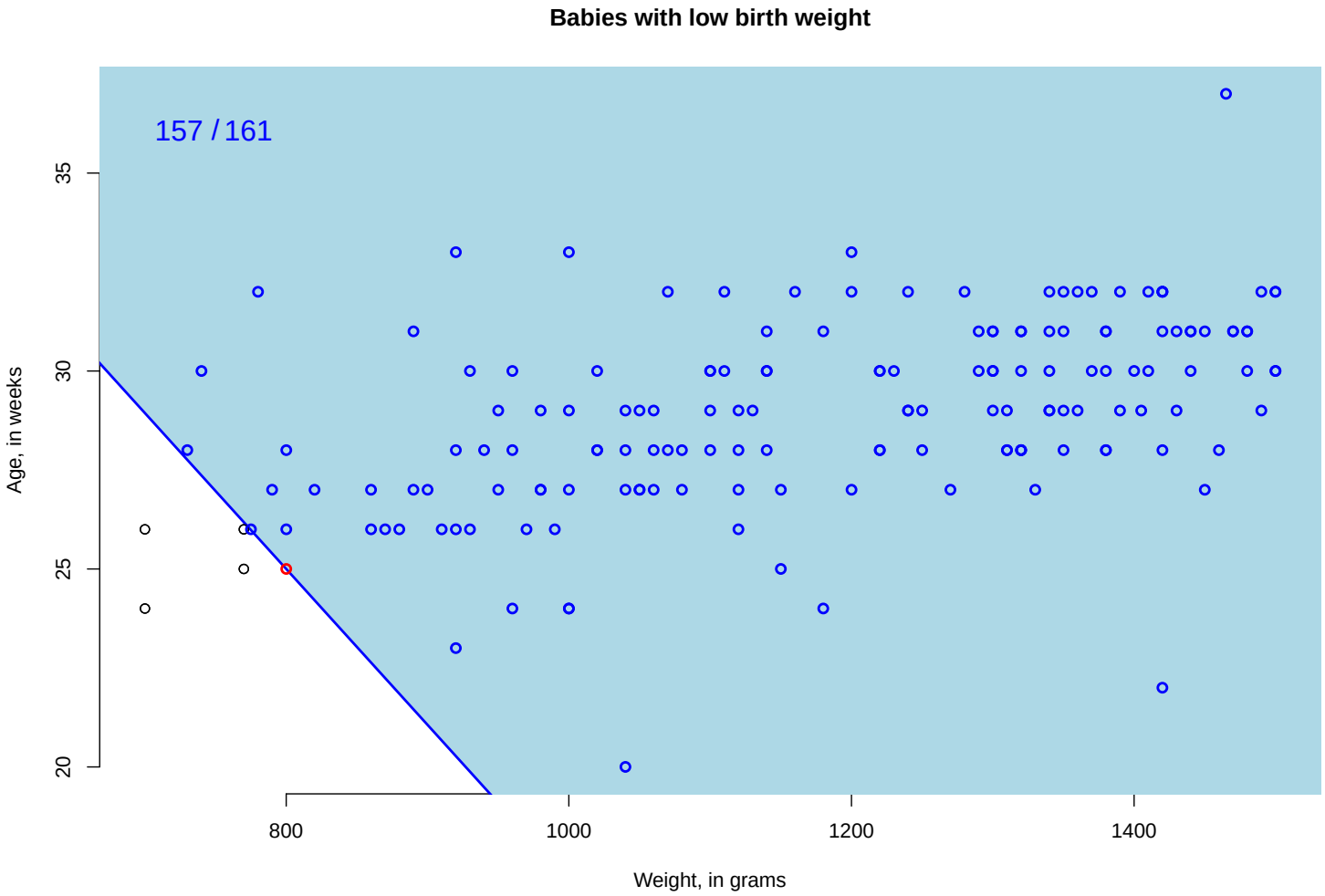


Halfspace data depth

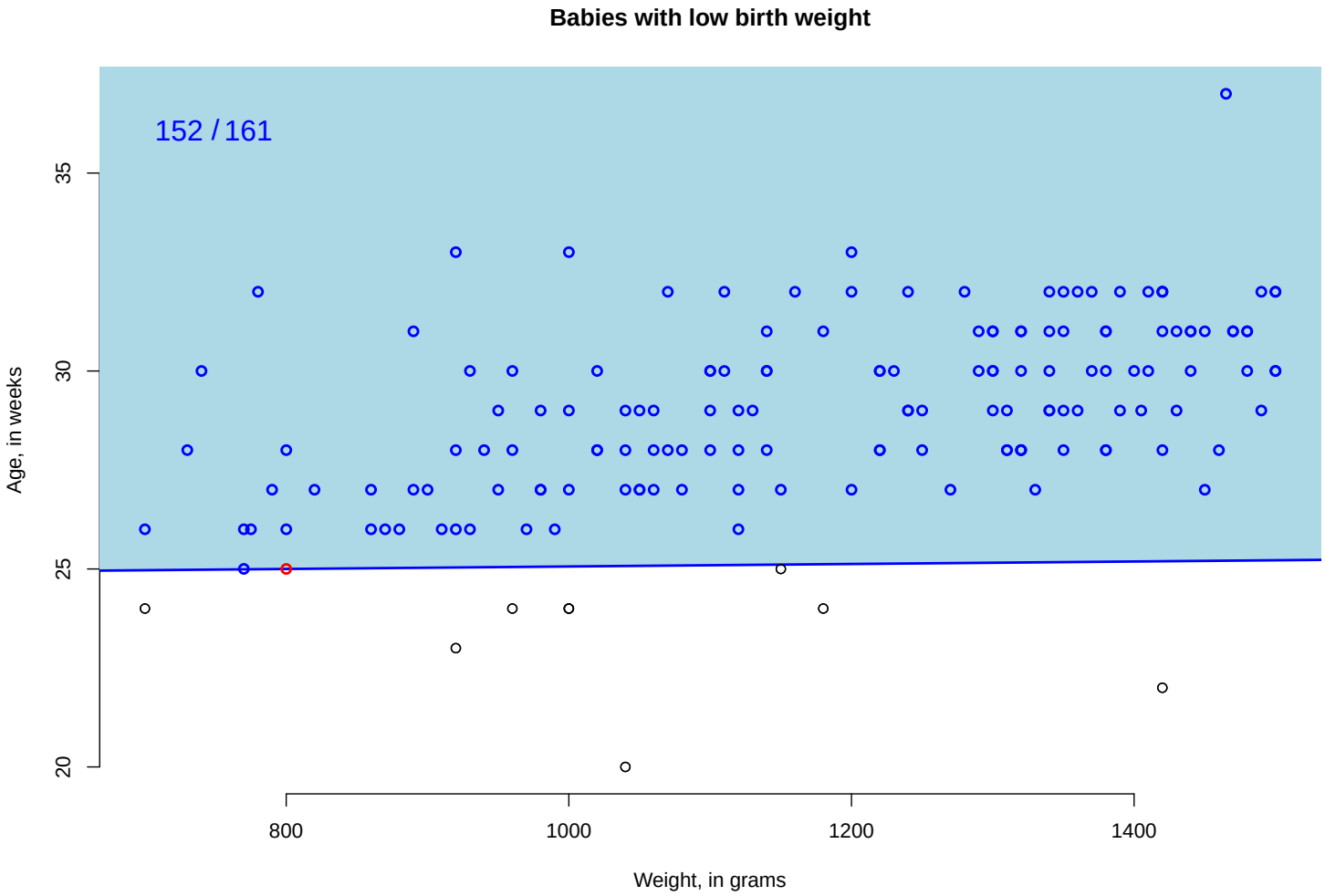
Halfspace data depth



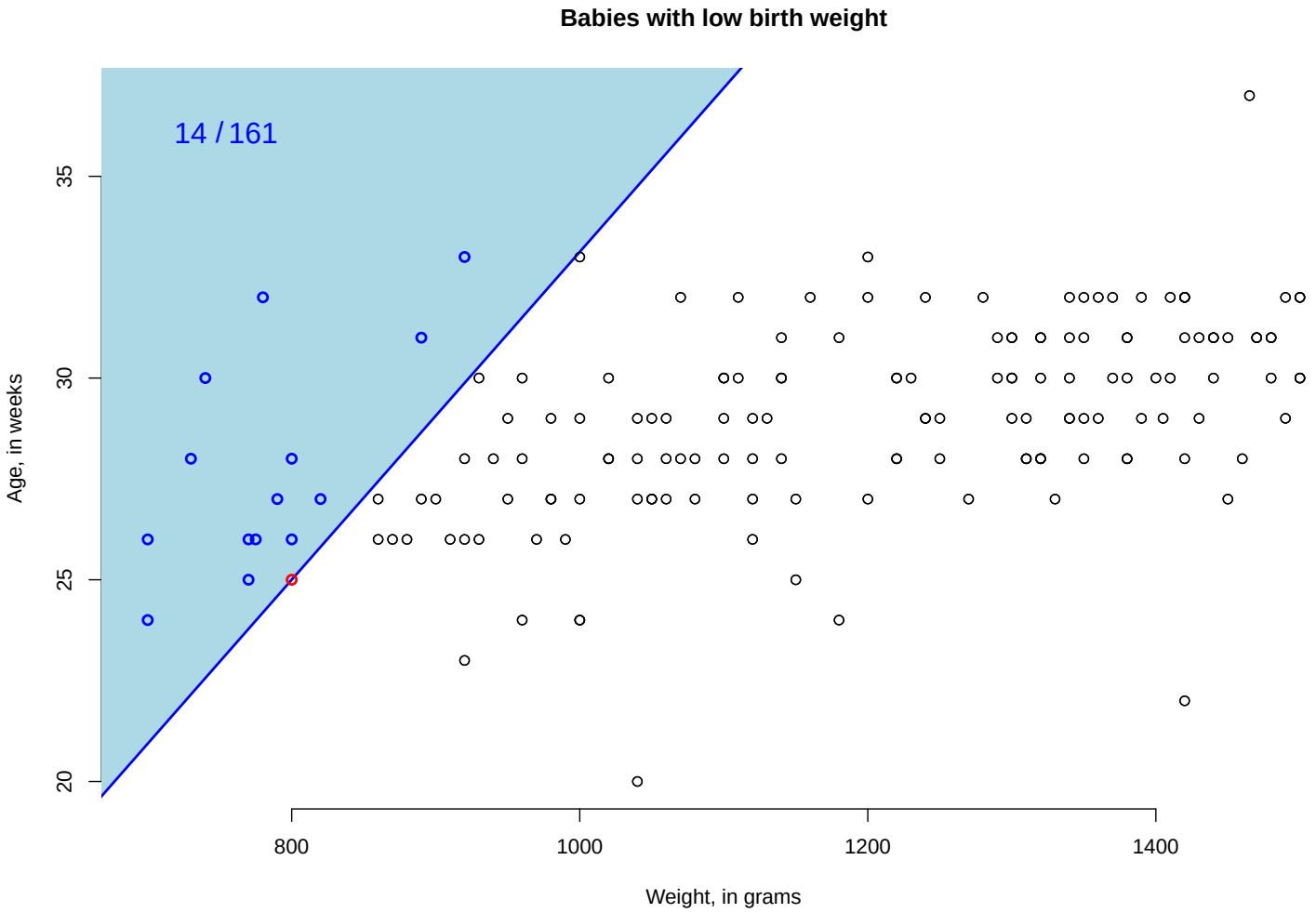
Halfspace data depth



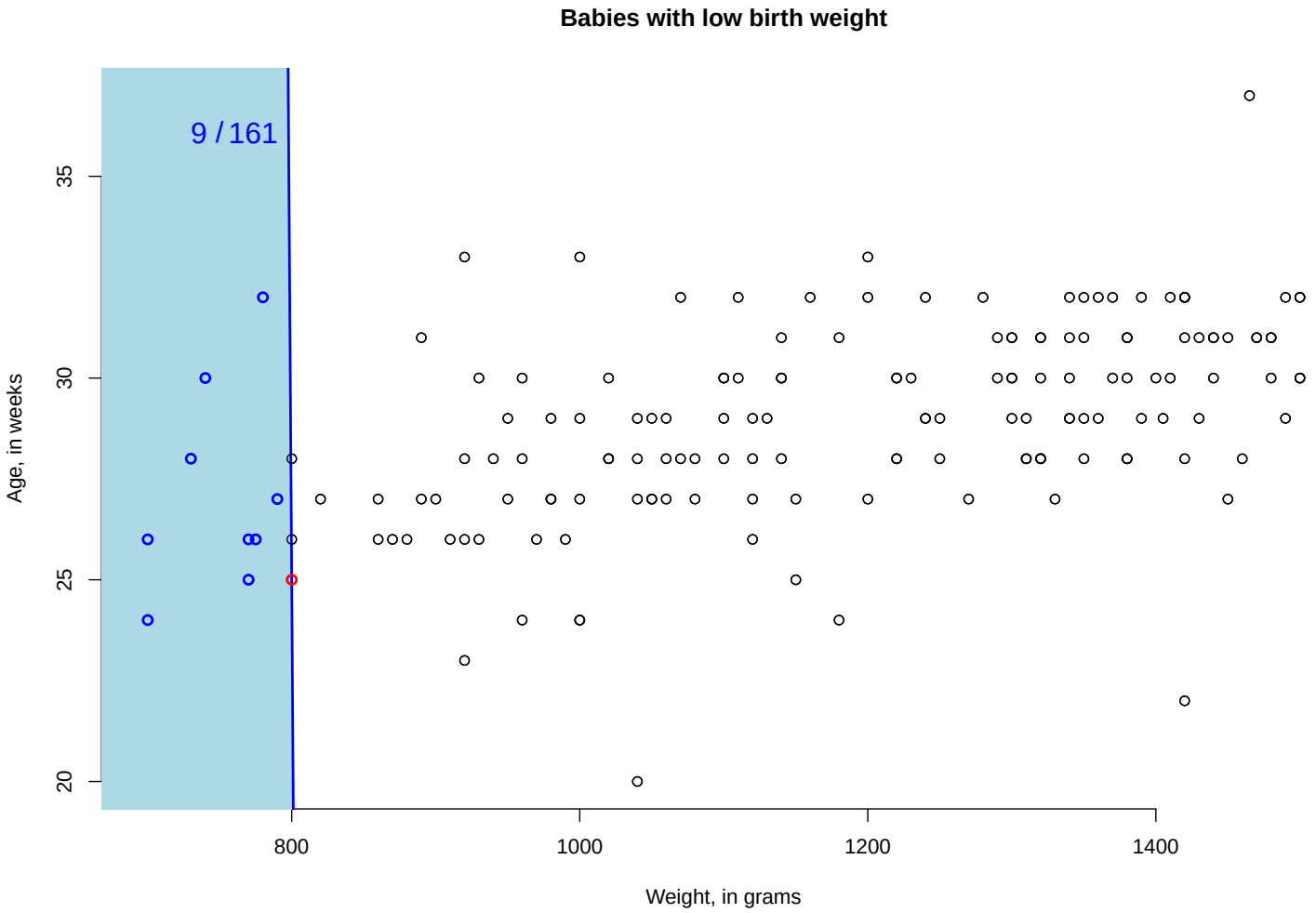
Halfspace data depth



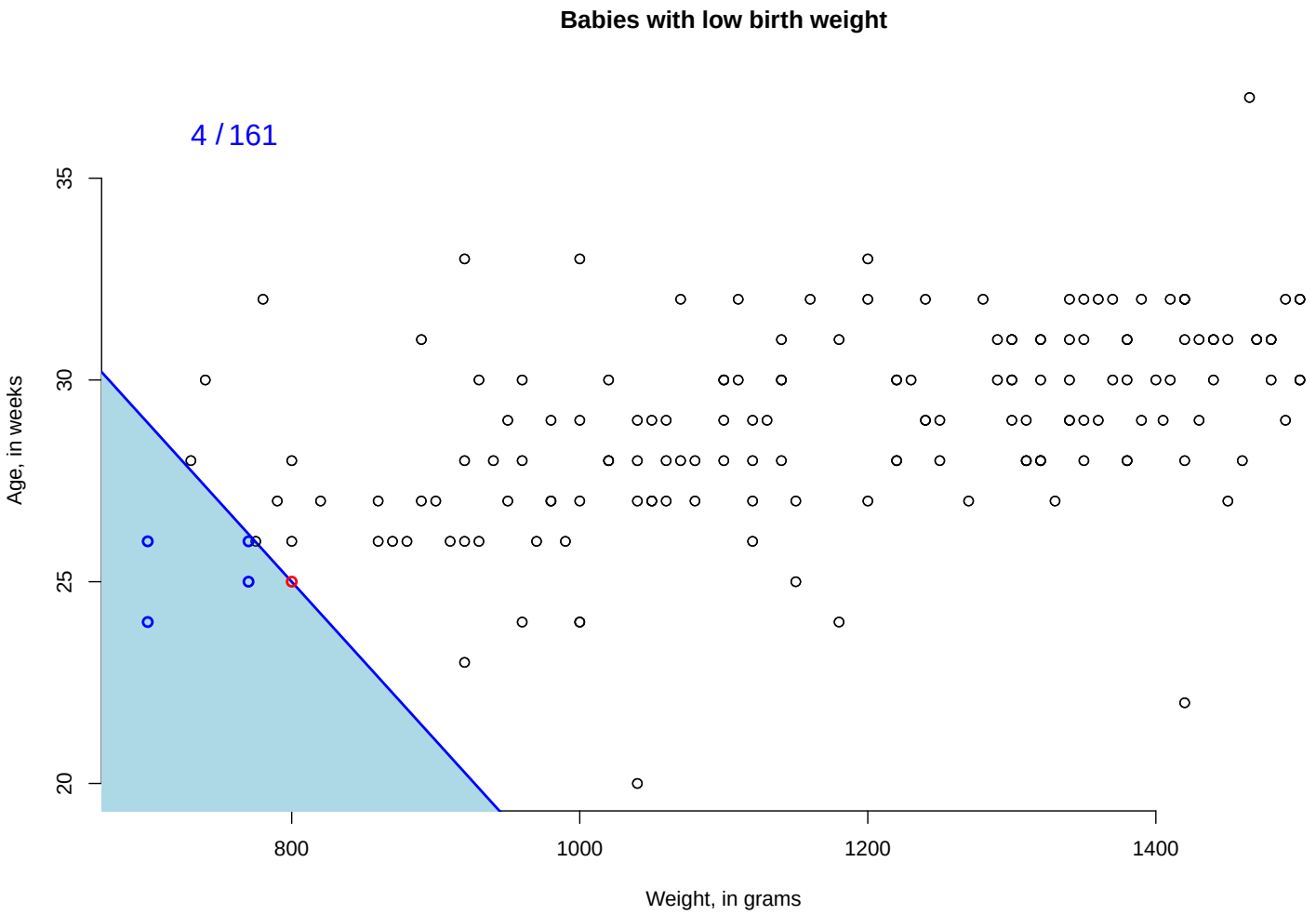
Halfspace data depth



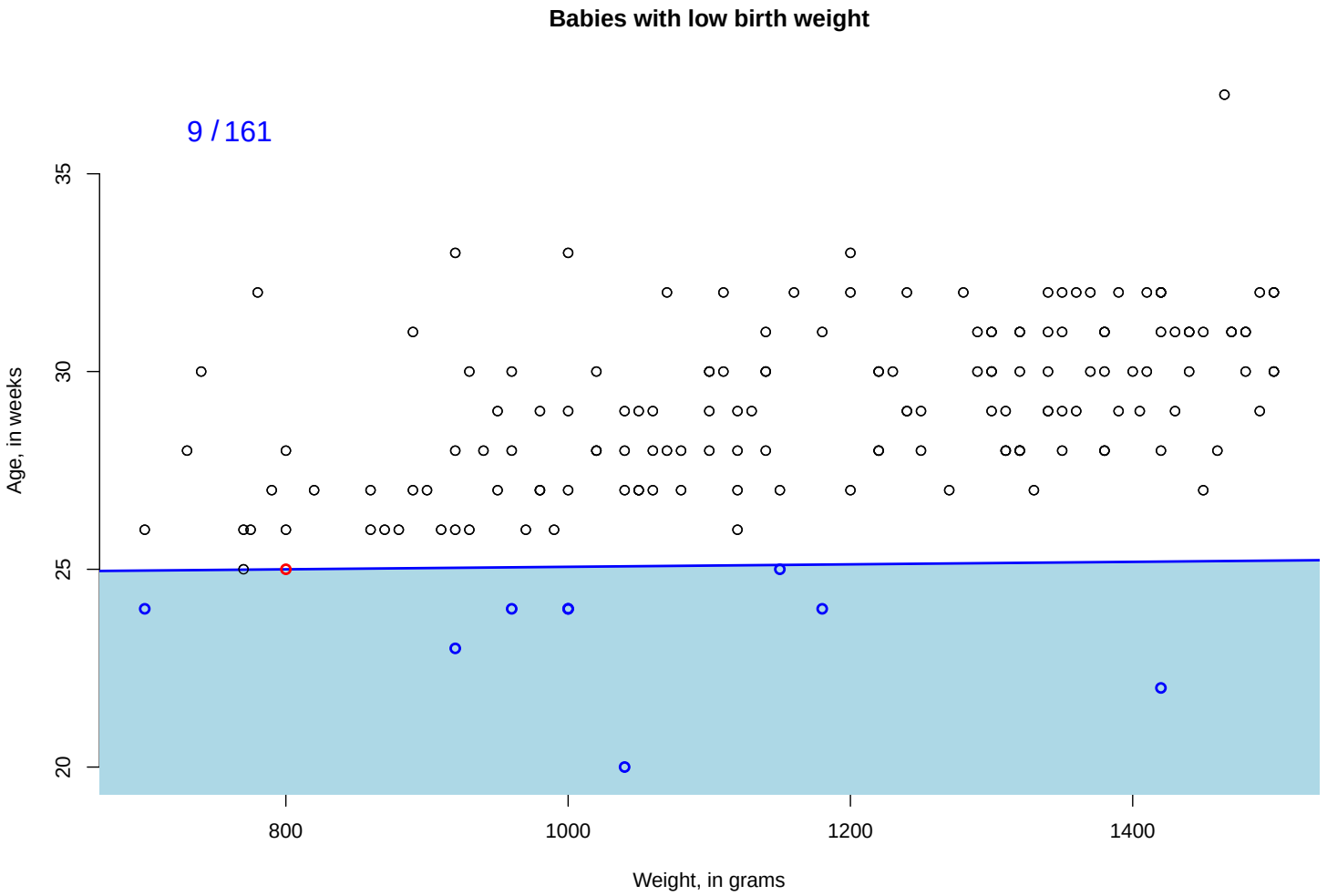
Halfspace data depth



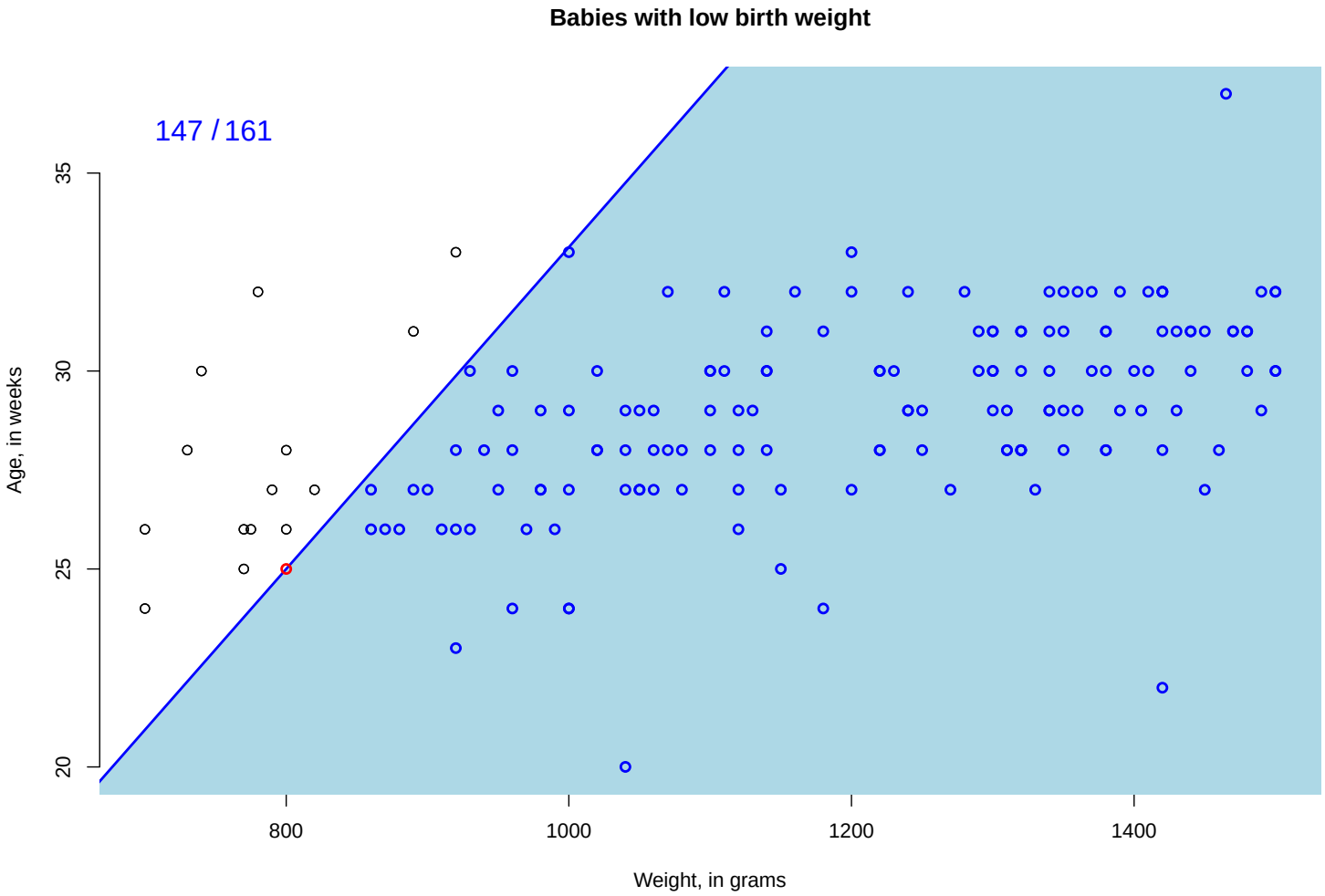
Halfspace data depth



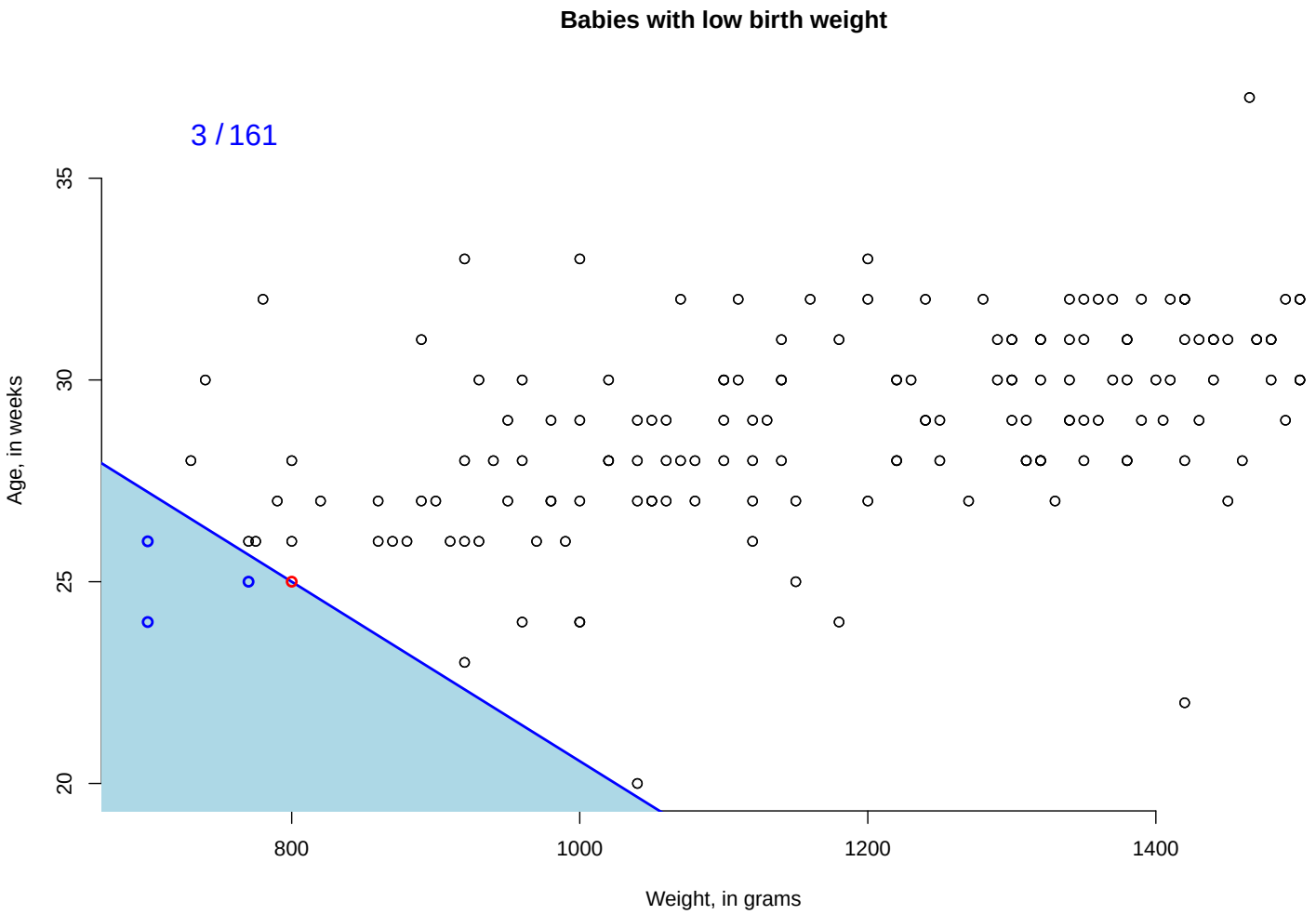
Halfspace data depth



Halfspace data depth



Halfspace data depth



Unparametrized Curves

Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.

A ***path* (or parametrized curve)** is a continuous map $\gamma : [a, b] \to \mathbb{R}^d$, where $a < b$. $S_\gamma = \gamma([a, b])$, the image of γ, is the ***locus*** of γ.

Two paths γ_1 and γ_2 are said equivalent, noted $\gamma_1 \mathcal{R} \gamma_2$, if $S_{\gamma_1} = S_{\gamma_2}$ and if they visit the points of S_{γ_1} in a same order.

An ***unparameterized curve*** $\mathcal{C} := \mathcal{C}_\gamma$ is the equivalence class of γ up to the equivalence relation $\mathcal{R}$. All members γ of this class have the same locus $S_\mathcal{C} = \gamma([a, b])$ and visit it in the same order.

The Space of Unparametrized Curves

Let $\mathcal{C}([0,1],\mathbb{R}^d)$ be the space of continuous functions from $[0,1]$ to $\mathbb{R}^d$. The ***space of unparametrized curves*** is

$$\Gamma = \{\mathcal{C}_\gamma : \gamma \in \mathcal{C}([0,1],\mathbb{R}^d)\}.$$

Proposition

Endowing Γ with the metric

$$d_\Gamma(\mathcal{C}_1,\mathcal{C}_2) = \inf\{\|\gamma_1-\gamma_2\|_\infty, \gamma_1 \in \mathcal{C}_1,\ \gamma_2 \in \mathcal{C}_2\}, \quad \mathcal{C}_1,\mathcal{C}_2 \in \Gamma,$$

where $\|\gamma\|_\infty = \sup_{t\in[0,1]}|\gamma(t)|_2$, the metric space (Γ, d_Γ) inherits the property of separability and completeness from $\mathcal{C}([0,1],\mathbb{R}^d)$.

Every probability measure defined on Γ is regular and tight, and there exists a non-atomic measure on (Γ, d_Γ).

The Length of a Curve

Let $\mathcal{C}$ be an unparametrized curve.
Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrization of $\mathcal{C}$, $\gamma \in \mathcal{C}$.
Let $a = \tau_0 < \tau_1 < \cdots < \tau_N = b$ be a subdivision of $[a, b]$.

The ***length of*** $\mathcal{C}$:

$$L(\mathcal{C}) = \sup_{\tau} \left\{ \sum_{i=1}^{N} |\gamma(\tau_i) - \gamma(\tau_{i-1})|_2 \ : \tau \text{ is a partition of } [a, b] \right\}.$$

An unparametrized curve $\mathcal{C}$ is called *rectifiable* if $L(\mathcal{C})$ is finite.

Proposition [Väisälä, 2006] : the normal parametrization

Let $\gamma : [a, b] \to \mathbb{R}^d$ be a rectifiable path, $\ell = L(\mathcal{C}_\gamma)$,

$$\phi : [a, b] \to [0, \ell], \quad \phi(t) = L\left(\gamma^\star_{|[a,t]}\right).$$

There exists a unique path $\gamma\star : [0, \ell] \to \mathbb{R}^d : \gamma = \gamma^\star \circ \phi$ and $L\left(\gamma^\star_{|[0,t]}\right) = t$.

The line integral along a curve

Let $\mathcal{C}$ be a rectifiable unparameterized curve, $\ell = L(\mathcal{C})$.
Let $\gamma : [0, \ell] \to \mathbb{R}^d$ be the normal parametrization of $\mathcal{C}$, $\gamma \in \mathcal{C}$.

For a non-negative Borel function $f : \mathbb{R}^d \to \mathbb{R}^d$, the ***line integral*** of f over $\mathcal{C}$ is :

$$\int_{\mathcal{C}} f(s)ds := \int_0^{\ell} f\left(\gamma(t)\right) dt.$$

The probability measure associated to $\mathcal{C}$ is

$$\forall A \in \mathcal{B}\left(\mathbb{R}^d\right), \quad \mu_{\mathcal{C}}(A) = \frac{1}{L(\mathcal{C})} \int_{\mathcal{C}} \mathbb{1}_A(s)ds.$$

Roughly speaking, $\mu_{\mathcal{C}}(A)$ can be interpreted as the "portion" of the length of curve $\mathcal{C}$ inside A to the total length of $\mathcal{C}$.

The Statistical Model

$$\mathcal{P} = \{P, \text{ a probability on } (\Gamma, d_\Gamma) \ : \ P\left(\{\mathcal{C} \in \Gamma \ : \ 0 < L(\mathcal{C}) < \infty\}\right) = 1\}.$$

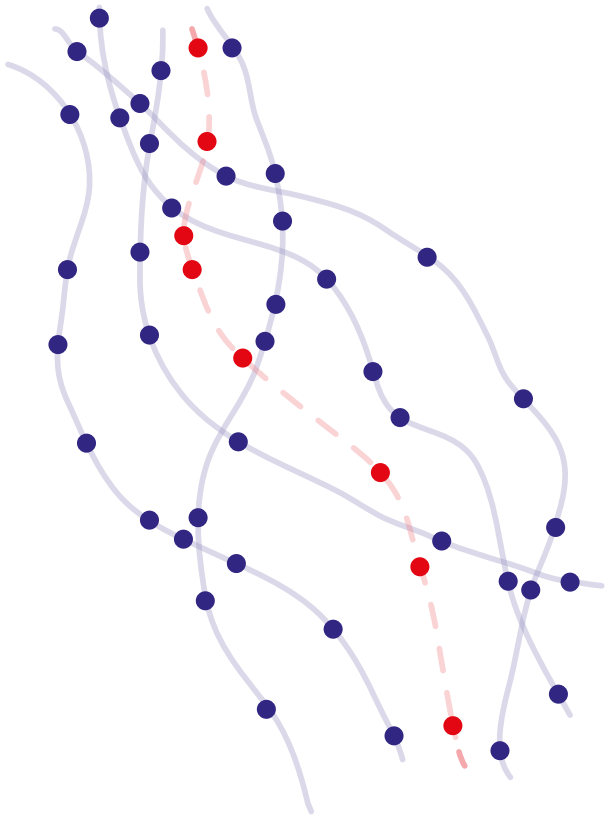
Let $\mathcal{X}$ be a random element of Γ with distribution $P \in \mathcal{P}$.

$$\forall A \in \mathcal{B}\left(\mathbb{R}^d\right), \quad Q_P(A) = \int_\Gamma \mu_{\mathcal{C}}(A) dP(\mathcal{C}).$$

We define a random vector X of $\mathbb{R}^d$:

$$X \sim Q_P \quad \text{and} \quad \mathcal{L}\left(X | \mathcal{X} = \mathcal{C}\right) = \mu_{\mathcal{C}}$$

$$\begin{cases} \mathcal{X}_1, \ldots, \mathcal{X}_n \text{ are i.i.d. from } P \in \mathcal{P}, \\ \text{and, for all } i = 1 \ldots n, \\ \qquad X_{i,1}, \ldots, X_{i,m_i} \text{ are i.i.d. and } \mathcal{L}(X_{i,j} | \mathcal{X}_i) = \mu_{\mathcal{X}_i}. \end{cases}$$



The sample (blue points)

$$\begin{cases} \mathcal{X}_1, \ldots, \mathcal{X}_n \text{ are i.i.d. from } P \in \mathcal{P}, \\ \text{and, for all } i = 1 \ldots n, \\ \qquad X_{i,1}, \ldots, X_{i,m_i} \text{ are i.i.d.} \\ \qquad \mathcal{L}(X_{i,j} | \mathcal{X}_i) = \mu_{\mathcal{X}_i}. \end{cases}$$

The curve $\mathcal{C}$ (red points)

$$\begin{cases} Y_1, \ldots, X_{b+c} \text{ are i.i.d.} \\ \mathcal{L}(Y_i) = \mu_{\mathcal{C}}. \end{cases}$$

Depth for a curve $\mathcal{C}$ with respect to P

Let $\mathcal{C}$ be a rectifiable curve in Γ.

The ***Tukey curve depth*** of $\mathcal{C}$ w.r.t. P is defined as,

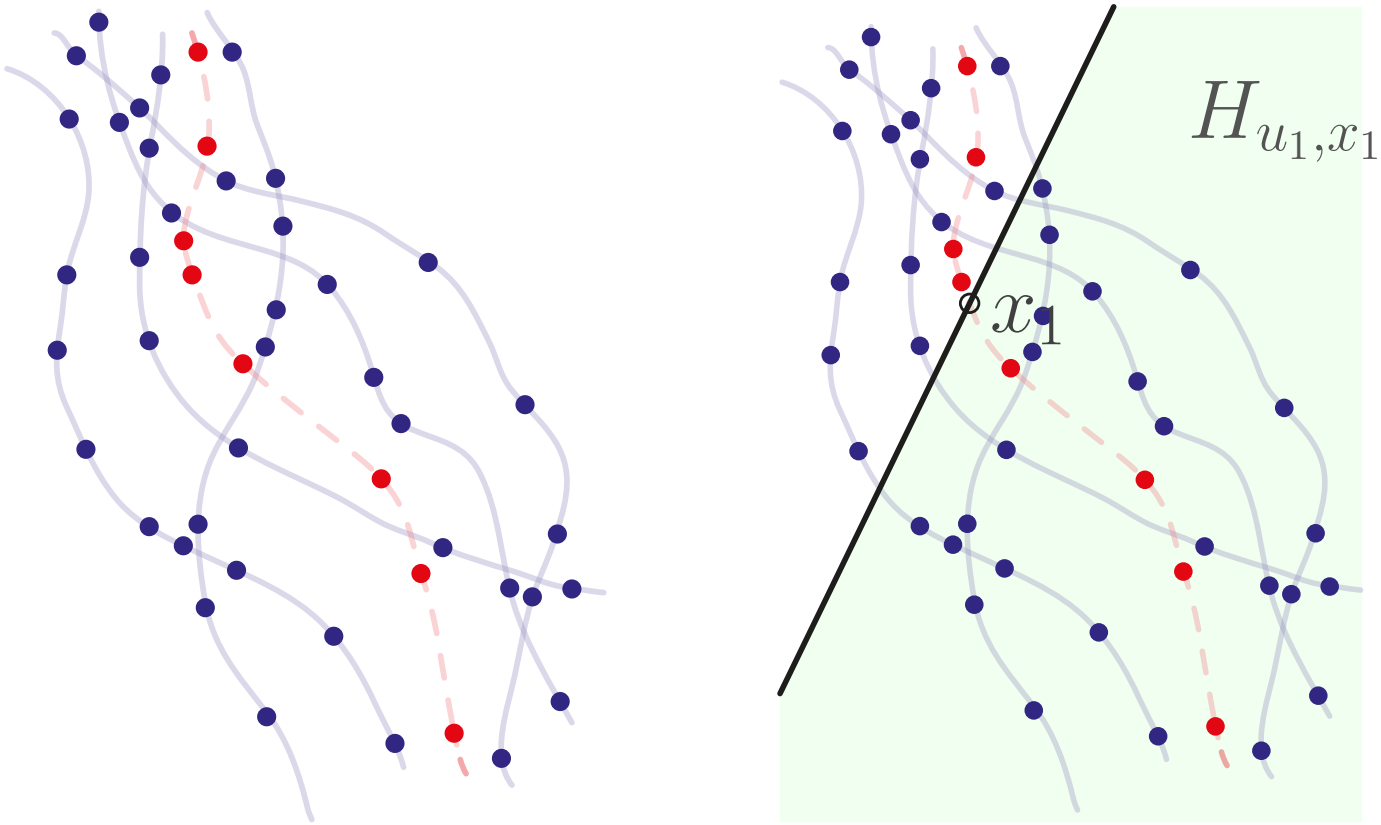
$$D(\mathcal{C}, P) := \int_{\mathcal{C}} D(s|Q_P, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(s),$$

where

$$D(x|Q_P, \mu_{\mathcal{C}}) := \inf_{u \in \mathcal{S}_d} \left\{ \frac{Q_P(H_{u,x})}{\mu_{\mathcal{C}}(H_{u,x})} \right\},$$

with the convention $0/0 = 0$.

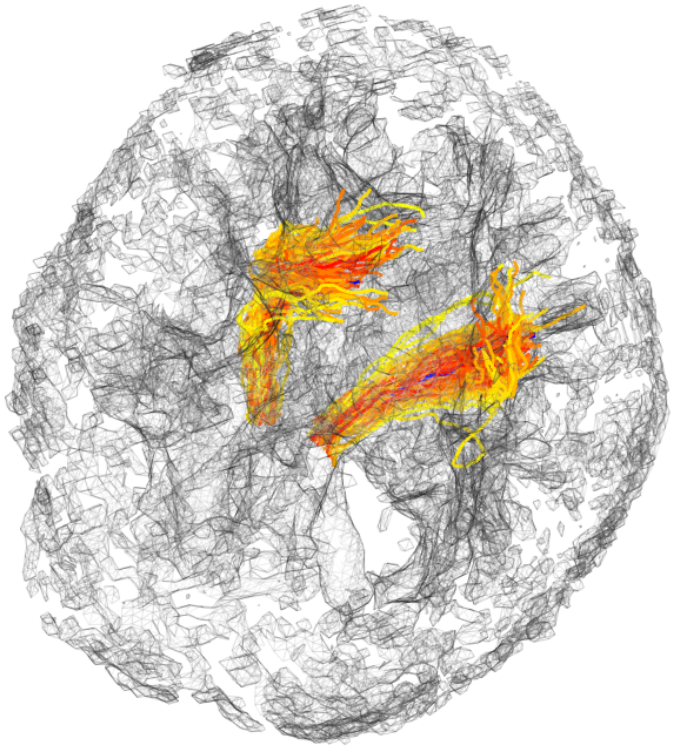
We obtained various theoretical properties of D (e.g., it takes values in $[0, 1]$, similarity invariance, vanishing at infinity). We also defined an empirical version of $D(\mathcal{C}, P)$ and proved its (weak) consistency.

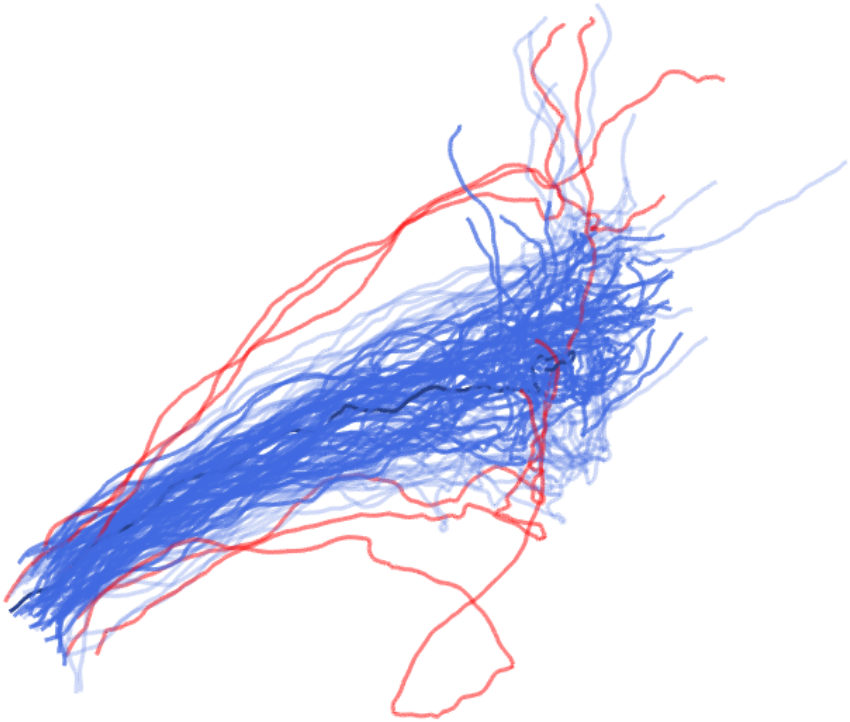


For u and x fixed, recall that $\mu_{\mathcal{C}}(H_{u,x})$ measures which fraction of length of the curve $\mathcal{C}$ delves into the half-space $H_{u,x}$, whereas $Q_P(H_{u,x})$ measures which expected fraction of length of a curve $\mathcal{X} \sim P$ delves into $H_{u,x}$. Consequently, the ratio $Q_P(H_{u,x})/\mu_{\mathcal{C}}(H_{u,x})$ is small when we expect curves generated according to P to enter less deeply into $H_{u,x}$ than the curve $\mathcal{C}$.

The results using our R package `CurveDepth`

`http://biostatisticien.eu/DataDepthFig4Left/`





Curve Registration for 64 brain bundles

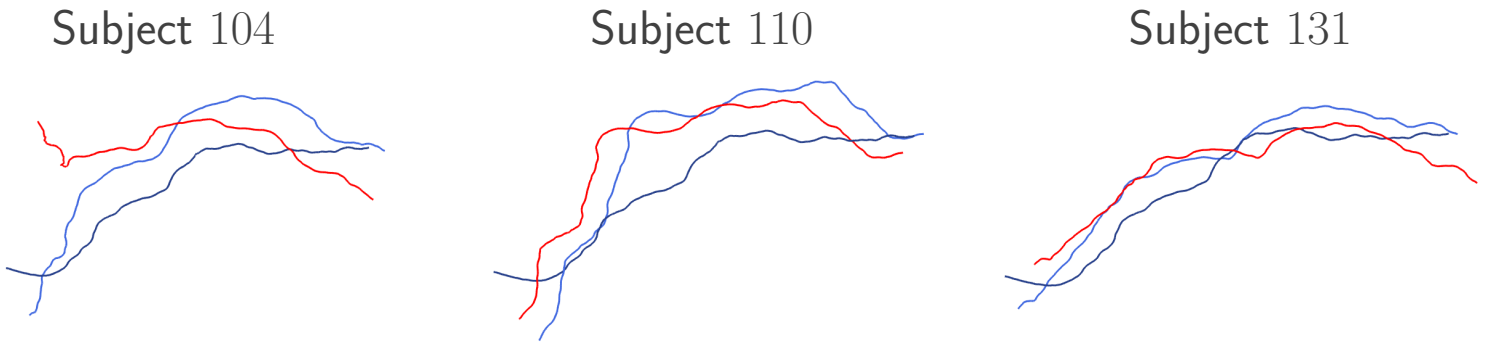
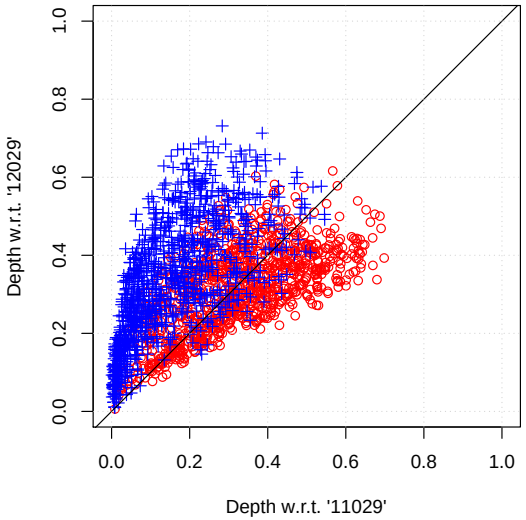
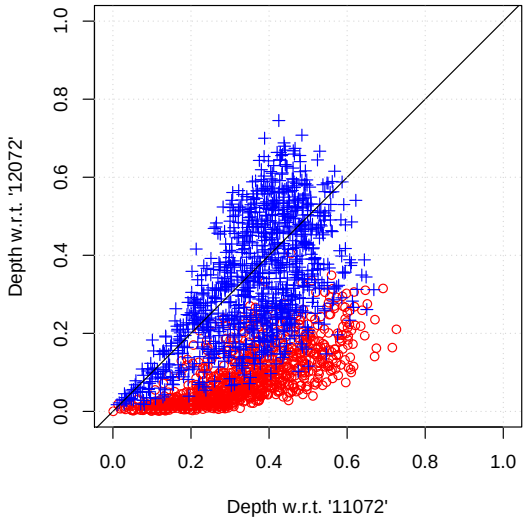


FIGURE – Illustration of the registration process. The red and the dark blue curves are respectively the deepest curves before registration of the respective subject and subject 235, the subject whose deepest curve is the *deepest of all*. We bring the red curve as close as possible (in terms of the distance) to the dark blue curve. The transformed curve (after registration) is the light blue curve. Distances from each curve to the deepest one (dark blue) before (red) and after (light blue) registration are 10.271 and 3.245 (for subject 104), 4.539 and 3.395 (for subject 110), 3.329 and 2.084 (for subject 131), respectively.

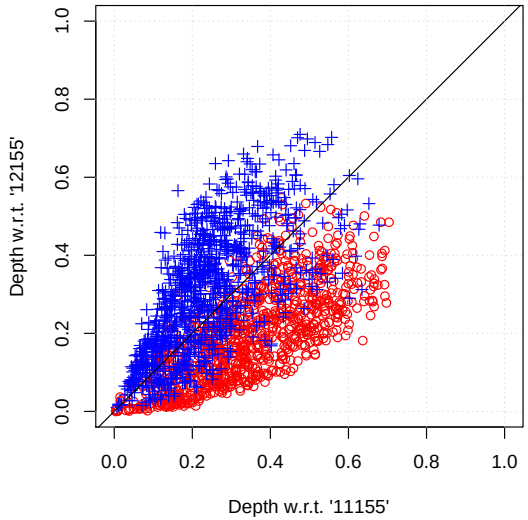
11029 *vs.* 12029 (DZ)



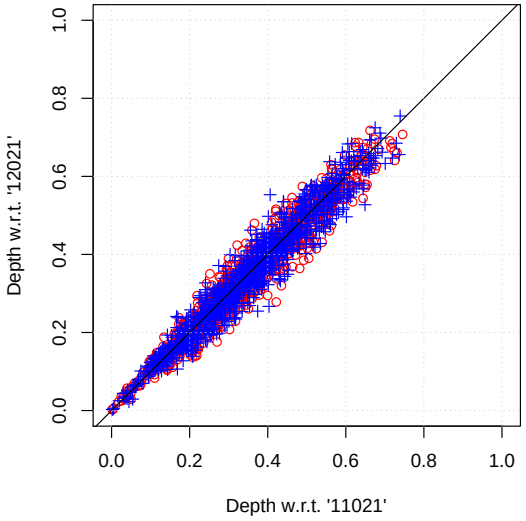
11072 *vs.* 12072 (DZ)



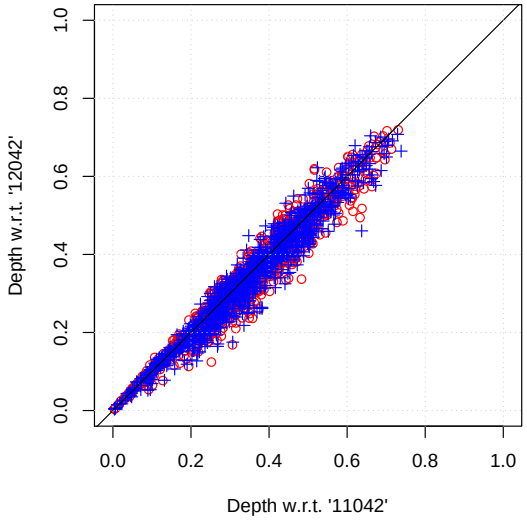
11155 *vs.* 12155 (DZ)



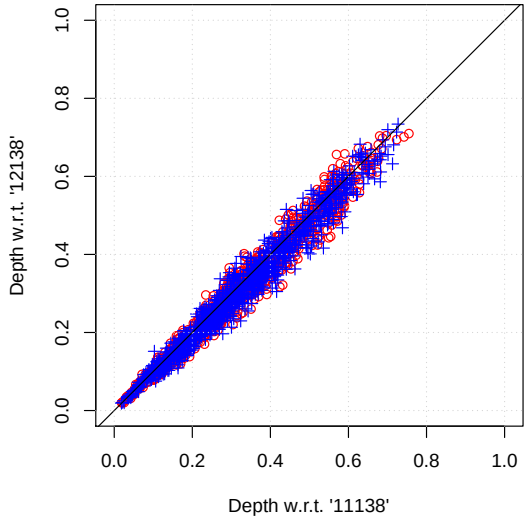
11021 *vs.* 12021 (MZ)



11042 *vs.* 12042 (MZ)



11138 *vs.* 12138 (MZ)



Thank you for your attention ! I

Kanchibhotla, S. C., Mather, A. A., Wen, W., Schofield, P. R., and Kwok, J. (2013).

Genetics of ageing-related changes in brain white matter integrity - a review.

Ageing Research Reviews, 12 :391–401.

Sachdev, P., Lammel, A., Trollor, J., Lee, T., Wright, M., Ames, D., Wen, W., Martin, N., Brodaty, H., Schofield, P., and the OATS research team (2009).

A comprehensive neuropsychiatric study of elderly twins : The older australian twins study.

Twin Research and Human Genetics, 12(6) :573–582.

Teipel, S. J., Grothe, M. J., Filippi, M., Fellgiebel, A., Dyrba, M., Frisoni, G., Meindl, T., Bokde, A., Hampel, H., Klöppel, S., Hauenstein, K., and the EDSD study group (2014).

Fractional anisotropy changes in alzheimer's disease depend on the underlying fiber tract architecture : A multiparametric dti study using joint independent component analysis.

Journal of Alzheimer's Disease, 41 :69–83.

Tukey, J. W. (1975).

Mathematics and the Picturing of Data.

In James, R. D., editor, *International Congress of Mathematicians 1974*, volume 2, pages 523–532.

Thank you for your attention ! II


Väisälä, J. (2006).

Lectures on n-Dimensional Quasiconformal Mappings.

Lecture Notes in Mathematics. Springer, Berlin Heidelberg.