# *Curvas estándar de crecimiento infantil mediante métodos estadísticos funcionales. Estudio de caso para la ciudad de Bogotá*

Andrés Nicolás López López
Estadístico


Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Noviembre de 2020

# *Curvas estándar de crecimiento infantil mediante métodos estadísticos funcionales. Estudio de caso para la ciudad de Bogotá*

Andrés Nicolás López López
Estadístico

Trabajo de grado presentado para optar al título de
Magister en Ciencias - Estadística

Director
B. Piedad Urdinola, Ph.D.
Doctor en Demografía

Codirector
Rubén Darío Guevara, Ph.D.
Doctor en Ciencias - Estadística

Línea de investigación
Bioestadística

Grupo de investigación
Observatorio demográfico y epidemiológico del Área Andina


Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Noviembre de 2020

**Título en español**

Curvas de crecimiento infantil mediante métodos estadísticos funcionales. Estudio de caso para la ciudad de Bogotá.

**Title in English**

Child growth curves through functional data analysis. A case study for Bogota city.

**Resumen:** Las curvas de crecimiento infantil son una herramienta utilizada en pediatría para caracterizar variables antropométricas de interés en una población determinada de niños y adolescentes. Para el caso colombiano existe un primer conjunto de curvas las cuales buscan mostrar el patrón de crecimiento para diferentes medidas antropométricas en las principales ciudades del país, controlando la variabilidad observada por la edad y el sexo de los individuos en la muestra. Sin embargo, estándares internacionales consideran una población sana y en condiciones óptimas de desarrollo en la construcción de curvas para evidenciar patrones ideales de crecimiento saludable, por otra parte, representar el patrón de crecimiento por variables adicionales como el desarrollo puberal es de particular interés en el estudio del crecimiento humano. A partir de una muestra no probabilística de niños saludables de la ciudad de Bogotá, el presente trabajo busca estimar curvas de crecimiento para contrastar los resultados con las curvas nacionales e internacionales y presentar una metodología para la elaboración de curvas de crecimiento infantil de talla ajustando por desarrollo puberal. En la comparación de las curvas son aplicados modelos de regresión siguiendo lineamientos internacionales mientras que la metodología de estimación propuesta combina modelamiento longitudinal del crecimiento infantil junto a análisis de datos funcionales escasos y métodos de regresión cuantílica.

**Abstract:** Child growth curves are a tool used in pediatrics for characterizing anthropometric variables of interest on a determined population of children and teenagers. The first set of curves for the Colombian case seek to show the growth pattern for different anthropometric measurements on the main cities in the country, controlling the observed variability by the age and sex of the individuals in the sample. Nevertheless, international standards consider a healthy population in optimal growth conditions for the growth curves estimation to depict ideal growth patterns of healthy growth, moreover, representing the growth pattern by additional features such as pubertal development is of particular interest when studying human growth. Through a non-probabilistic sample of healthy children from Bogota city, the following work aims to estimate growth curves for contrasting the results with the national and international curves and to present a methodology of child growth curve estimation for height adjusting for pubertal development. The curve comparison is performed by regression models following international guidelines, while the proposed methodology combines longitudinal modeling of child growth with sparse functional data analysis and quantile regression methods.

**Palabras clave:** Análisis de datos funcionales, datos escasos, crecimiento infantil.

**Keywords:** Functional data analysis, sparse data, child growth.

## Nota de aceptación

Trabajo de tesis
Aprobado

---

Jurado
Ramón Giraldo

---

Jurado
Martha Bohorquez

---

Director
B. Piedad Urdinola

---

Codirector
Rubén Darío Guevara

Bogotá, D.C., Noviembre de 2020

# Dedicado a

A mis padres y a mis hermanas. Siempre agradeceré el estar incondicionalmente a mi lado.

# Agradecimientos

Agradezco en primera instancia a la profesora B. Piedad Urdinola por su constante apoyo, no solamente en la concepción y elaboración de este trabajo, sino desde siempre en mi formación profesional como Estadístico y durante esta maestría. Agradezco muy especialmente el siempre creer en mis capacidades. Agradezco también la atenta retroalimentación del profesor Rubén Guevara, quién además motivó el enfoque del actual trabajo desde sus excepcionales lecciones de análisis de datos funcionales. Extiendo además este agradecimiento al doctor Mauricio Llano por brindar la información necesaria para la elaboración del trabajo, y en general a nuestro grupo de investigación, por sus valiosos aportes.

Siempre estaré agradecido por la oportunidad de convertirme docente becario dada por la Universidad Nacional junto al soporte financiero brindado por la Dirección Académica de la Universidad, el cual me permitió culminar de manera satisfactoria la carga académica de la maestría. Finalmente me gustaría agradecer al departamento de Estadística de la Universidad, en particular a sus docentes, los cuales han generado en mi una gran pasión por la estadística la cual me ha brindado grandes oportunidades y me ha permitido superar diferentes adversidades. Espero seguir sus pasos y en un futuro lograr transformar vidas a partir del amor al conocimento y la enseñanza de la estadística.

# Índice general

# Índice de tablas

# Índice de figuras

# CAPÍTULO 1

# Introducción

Las curvas de crecimiento infantil son herramientas gráficas usadas en pediatría para caracterizar el crecimiento de niños y adolescentes en función de su edad. Estas curvas presentan la distribución de una variable antropométrica en distintas edades para cada sexo, su construcción generalmente se resume en la estimación de cuantiles tales como la mediana y otros seleccionados de manera simétrica respecto a esta. En la práctica clínica, posibles problemas de salud son presentados por individuos en cuantiles extremos de la variable para la edad y el sexo correspondiente. Las curvas de crecimiento pueden ser de referencia para una población dada, con las cuales se pretende describir de manera representativa las características antropométricas de un conjunto de individuos. Por otra parte, las curvas pueden evidenciar características ideales o también llamadas "estándar" en una población de niños saludables (Kelnar et al., 2007). Las curvas de crecimiento infantil convencionales, tanto de referencia como estándar, diferencian únicamente por edad y sexo la distribución de la variable respuesta. Esto es dado por una mayor influencia de la edad y, en menor medida, del sexo en bastantes variables clínicas (Wright & Royston, 1997). Sin embargo, en el estudio del crecimiento infantil pueden existir diferencias en la distribución de la característica antropométrica de interés al controlar por otras covariables, por tanto, las curvas de crecimiento condicionadas (en contraste con las curvas convencionales no condicionadas) consideran variables adicionales en su construcción.

Este trabajo presenta un estudio de caso como primer esfuerzo en la construcción de curvas estándar de crecimiento de la ciudad de Bogotá: la estimación de las curvas es realizada mediante métodos de regresión bajo lineamientos internacionales mientras que la metodología para la construcción de curvas es elaborada usando modelamiento funcional escaso y regresión cuantílica, motivado por el trabajo de Zhang et al. (2015). La propuesta de Cole et al. (2010) es implementada para la estimación paramétrica de la madurez de los individuos previo al análisis funcional. El trabajo presenta una aplicación novedosa en la elaboración de curvas de crecimiento condicionadas por madurez para la variable talla: los cuantiles de crecimiento son obtenidos mediante regresión cuantílica sobre los puntajes de las funciones propias del operador de covarianza, cuyo kernel es estimado mediante un método especializado para la estimación de funciones de covarianza (Xiao et al., 2018). Un modelo no lineal de efectos mixtos es utilizado para diferenciar la madurez de los individuos en la muestra, diferenciando los procesos aleatorios subyacentes por la variación en fase propia de los registros longitudinales de crecimiento en talla.

Una revisión detallada de algunos de los métodos estadísticos de construcción de curvas de crecimiento controlando por edad y sexo puede encontrarse en Wright & Royston (1997). En particular, una de las técnicas más utilizadas en la estimación paramétrica de curvas de crecimiento es la metodología desarrollada por Cole (1988) y Cole & Green (1992) que a partir de la familia de transformaciones de Box-Cox busca encontrar curvas centílicas suaves en función de la edad. Al considerar no sólo los parámetros ubicación y escala de la distribución normal, sino además un parámetro de potencia, este método permite encontrar una distribución paramétrica conocida para los datos antropométricos originales incluso en presencia de sesgo no despreciable en función de la edad.

Además del control inducido en la partición usual por sexo y de la covariable edad de las curvas no condicionadas, las condicionadas permiten considerar tanto información adicional del paciente como la estructura longitudinal del mismo en la construcción de curvas. Por una parte, los registros longitudinales permiten detectar eficientemente la presencia patrones inusuales de crecimiento al entender las características antropométricas actuales en el contexto de las mediciones previas de un individuo dado, destacando así la importancia práctica de considerar la estructura longitudinal de los datos (Wei et al., 2006). Por ejemplo, Cole (1994) resalta que tanto la estatura como el peso en la infancia y pubertad para sujetos en percentiles extremos tienden a alcanzan percentiles centrales, implicando para estos individuos una velocidad de crecimiento inusual sin que presenten necesariamente un problema de salud. Este fenómeno de compensación/descompensación de crecimiento es desapercibido al contar con observaciones individuales de un paciente determinado. Por otra parte, la importancia del control por covariables adicionales es reflejada por los pacientes que experimentan variantes normales en su maduración durante la pubertad, pero resultan en una investigación clínica al presentar diferencias importantes en talla respecto a curvas no condicionadas (Kelly et al., 2014).

El trabajo de Cole (1994) presenta una primera aproximación en la estimación de curvas de crecimiento con información longitudinal, éste desarrolla una versión general del método de Cole & Green (1992) mediante regresión autoregresiva. La propuesta, además de requerir intervalos fijos de medición de la variable, esta limitada a normalidad marginal para cada ventana temporal, que como destaca Wei et al. (2006), no garantiza la normalidad de los errores en modelos condicionales. Propuestas más recientes realizan modelamiento de regresión cuantílica para datos longitudinales flexibilizando los supuestos anteriores e incorporando variables adicionales en el modelamiento de las curvas (Wei et al., 2006). Sin embargo, estas metodologías para la obtención de las curvas no permiten una clara interpretación del efecto de la velocidad de crecimiento, el cual permite la distinción de patrones de crecimiento saludables (Kelly et al., 2014). Cole et al. (2010) considera un modelo para talla en la pubertad que utiliza la información de la muestra de manera conjunta para la estimación de trayectorias individuales de crecimiento, este acercamiento permite resumir el patrón de individual crecimiento mediante tres parámetros, uno de los cuales controla por la maduración de cada individuo y otro por la velocidad de crecimiento. Este modelo se restringe a la variable talla en la pubertad y a diferencia de las curvas de crecimiento no condicionadas, no permite una clara distinción de un comportamiento inusual en talla para los individuos en la muestra. Se observa que la información longitudinal obtenida es modelada mediante métodos en los que la naturaleza continua del patrón de crecimiento subyacente es obviada o la interpretabilidad de las curvas convencionales no es alcanzada. Este trabajo presenta una alternativa en la elaboración de curvas de crecimiento condicionadas por madurez para la variable talla, permitiendo un análogo longitudinal a la propuesta de Cole & Green (1992) diferenciando por madurez.

En función de su edad y por cada una de las visitas al médico especialista, los pacientes proveen observaciones discretizadas de una medida antropométrica continua determinada, estas observaciones ruidosas permiten la obtención de realizaciones continuas del proceso aleatorio para cada paciente. El acercamiento de dimensión infinita para el problema de estimación de curvas de crecimiento a partir de datos funcionales es un tema reciente en la literatura. James et al. (2000) presenta un modelo de efectos mixtos para la descomposición de la variabilidad observada en componentes principales funcionales, lo cual es realizado bajo fuertes supuestos distribucionales para datos de crecimiento infantil (Zhang et al., 2015). Posteriormente, la metodología elaborada por Yao et al. (2005) busca una descomposición funcional de la variabilidad observada mediante la expansión de Karhunen-Loève y estimaciones no paramétricas de las propiedades de primer y segundo orden del proceso aleatorio. Esta propuesta ha sido usada recientemente en el modelamiento de la velocidad de crecimiento en la adolescencia (Simpkin et al., 2017). Basados en estos desarrollos para el contexto funcional del crecimiento humano, el trabajo de Zhang et al. (2015) busca de manera no paramétrica descomponer las curvas de crecimiento infantil mediante componentes principales funcionales, y a partir de los cuantiles multivariados de los puntajes, estimar la distribución subyacente del crecimiento infantil. Estos métodos pertenecen al campo de los datos funcionales escasos, en el cual el número de observaciones discretizadas de los individuos es escasa e irregular en su dominio, la cual es una característica comúnmente encontrada para los datos de crecimiento infantil.

A continuación se describe el contenido de los diferentes capítulos del documento, siendo la presente introducción el capítulo uno del trabajo: el capítulo dos presenta el marco conceptual, en el cual se desarrollan las características generales de los datos funcionales densos y escasos, resaltando sus diferencias en términos de la estimación de los parámetros funcionales y destacando particularmente la obtención de la función de covarianza y de los componentes principales funcionales, los cuales son de gran importancia en el análisis funcional de datos escasos. En el capítulo tres se realiza una revisión de la literatura correspondiente al trabajo nacional e internacional en la elaboración de curvas de crecimiento. Se destacan además diferentes métodos en el análisis de curvas de crecimiento distinguiendo por la naturaleza transversal o longitudinal de los datos junto al avance metodológico en el contexto funcional escaso, y finalmente, se presenta la estimación de cuantiles multivariados a partir de modelos de regresión cuantílica. En el capítulo cuatro se presentan los resultados del análisis de la información, destacando la comparación entre los estándar de crecimiento internacionales y las curvas locales obtenidas. Se resaltan también las características del análisis longitudinal de la información mediante datos funcionales escasos, regresión cuantilica y modelamiento longitudinal de crecimiento humano para la obtención de curvas de crecimiento basadas en información longitudinal.

# CAPÍTULO 2

# Marco conceptual

Los registros de crecimiento infantil a través de la edad del paciente presentan una naturaleza continua aunque las observaciones sean realizaciones discretas en instantes de tiempo determinados: cada individuo provee observaciones discretizadas de la variable antropométrica continua de interés por cada una de sus visitas al médico especialista. Adicionalmente, para un individuo dado las visitas son rara vez igualmente espaciadas, tampoco es usual que las visitas sean realizadas a la misma edad para diferentes individuos. Un posible acercamiento al modelamiento estadístico del crecimiento infantil está dado a partir del análisis de elementos aleatorios funcionales, en los cuales las observaciones representan funciones, en lugar de escalares o vectores. La sección 2.1 presenta el marco teórico del análisis de datos funcionales densos, en los cuales de manera individual cada paciente cuenta con información suficiente en el dominio de interés para caracterizar su comportamiento funcional. Por otra parte, la sección 2.2 muestra el desarrollo de los métodos funcionales desde el punto de vista escaso, en el cual la información a nivel individual no es suficiente, por lo que las observaciones son utilizadas de manera conjunta para el análisis del fenómeno aleatorio funcional.

## 2.1. Análisis de datos funcionales densos

Para el análisis de datos funcionales densos generalmente se opta en un primer paso por realizar un suavizamiento de las observaciones discretizadas. El suavizamiento permite una reconstrucción de las observaciones funcionales, removiendo a su vez el ruido de los datos observados. Este primer paso en el análisis permite un posterior estudio de las propiedades de primer y segundo orden de la función aleatoria subyacente a partir de la muestra de observaciones funcionales reconstruidas. Posteriormente, en el estudio de la variabilidad observada de los datos funcionales se presenta el análogo del análisis de componentes principales para el caso funcional. El análisis de componentes principales funcional permite resumir patrones de variabilidad mediante funciones características de los datos, las cuales pueden presentar comportamientos esperados en el estudio del fenómeno, como también pueden revelar características novedosas que reflejen la complejidad presente en los datos (Ramsay & Silverman, 2005). Las siguientes subsecciónes siguen de cerca el trabajo de Kokoszka & Reimherr (2017), Aguilera & Aguilera-Morillo (2013) y Ramsay & Silverman (2005).

### 2.1.1. Expansión en funciones base

En el contexto funcional se tiene una colección finita de observaciones ruidosas, donde para cada individuo, estas se asumen provenientes de una curva de dimensión infinita la cual es evaluada en puntos de un intervalo determinado. Para la $i$-ésima unidad estadística se tiene un conjunto de $n_i$ observaciones discretizadas $\{x_{i1}, ..., x_{ij}, ..., x_{in_i}\}$ de la función $x_i$ en los puntos $t_{i1}, ..., t_{ij}, ..., t_{in_i}$ con $x_{ij} \in \mathbb{R}$, $t_{ij} \in T$ y $T$ un intervalo acotado y además cerrado que representa el dominio sobre los reales donde se definen los datos funcionales. De manera general, se inicia encontrando una expresión funcional para cada curva a partir de los datos discretizados mediante una expansión en funciones base, la cual es una colección de funciones conocidas e independientes entre sí, que permiten aproximar una función arbitraria al tomar una combinación lineal de un número suficiente de dichas funciones (Ramsay & Silverman, 2005). La mayoría de las bases comúnmente utilizadas están conformadas por funciones suaves, y en dado caso, los datos funcionales reconstruidos heredan la suavidad de sus bases correspondientes (Kokoszka & Reimherr, 2017).

Para la observación discretizada ruidosa $x_{ij}$ de la función $x_i$ observada en $t_{ij} \in T$ con error $\epsilon_{ij}$ se tiene su expansión en una base de funciones $\{\phi_j\}_{j=1,..,p}$ de dimensión $p$ como

$$\begin{cases} x_{ij} & = x_i(t_{ij}) + \epsilon_{ij} \\ x_i(t_{ij}) & \approx \sum_{k=1}^{p} c_{ik}\phi_k(t_{ij}) \end{cases} \tag{2.1}$$

Así, para cada individuo, la colección de observaciones discretizadas $\{x_{i1}, ..., x_{ij}, ..., x_{in_i}\}$ es representada mediante un vector de coeficientes $\mathbf{c}_i = (c_{i1}, ..., c_{ip})'$ sobre una base de funciones dada. La formulación de las observaciones como combinación lineal de funciones base recupera la dimensionalidad infinita de los datos funcionales, además permite un marco de referencia común con el cual comparar los datos funcionales cuando los instantes de las observaciones discretizadas $t_{ij}$ difieren entre las curvas (Kokoszka & Reimherr, 2017). La transformación en la expresión 2.1 se resume en la estimación de los coeficientes de la expansión $c_{ij}$ los cuales son generalmente encontrados mediante mínimos cuadrados (ordinarios o ponderados) usando regresión o suavizamiento spline (Aguilera & Aguilera-Morillo, 2013).

#### 2.1.1.1. Suavizamiento de curvas. Regresión spline

La expansión en funciones base permite recuperar la dimensionalidad infinita de la observación funcional para cada individuo a partir del conjunto de datos discretizado; con lo cual, dada una base de funciones $\{\phi_j\}_{j=1,..,p}$ se busca para cada vector de observaciones $\mathbf{x}_i = (x_{i1}, ..., x_{in_i})$ una estimación de los coeficientes $\mathbf{c}_i$ de la expansión en funciones base $x_i(t) = \mathbf{c}_i'\boldsymbol{\phi}(t)$ con $\boldsymbol{\phi}(t) = (\phi_1(t), ..., \phi_p(t))'$. Como resalta Aguilera & Aguilera-Morillo (2013), los coeficientes estimados $\hat{\mathbf{c}}_i$ son usualmente calculados mediante la minimización de la siguiente expresión

$$SSE(\mathbf{c}_i) = (\mathbf{x}_i - \mathbf{c}_i'\boldsymbol{\phi}(t))'\mathbf{W}(\mathbf{x}_i - \mathbf{c}_i'\boldsymbol{\phi}(t)), \tag{2.2}$$

donde $\mathbf{W}$ induce una posible estructura de autocorrelación o de no estacionalidad en los errores $\epsilon_{ij}$ del modelo. Al ser $\mathbf{W} = \mathbf{I}_p$ se asume independencia y varianza constante de los errores.

Diferentes bases de funciones pueden ser utilizadas para un conjunto de datos dado e idealmente las características de la base seleccionada deben coincidir con aquellas de las funciones a ser estimadas, por tanto, las particularidades de los datos determinan la base de funciones a seleccionar (Aguilera & Aguilera-Morillo, 2013; Ramsay & Silverman, 2005). Entre las diferentes bases de funciones usadas para el suavizamiento de curvas, la base *B-spline* es comúnmente aplicada en datos funcionales no periódicos al estar conformada por funciones spline (Horváth & Kokoszka, 2012). Un spline es una función definida en subintervalos usando polinomios de grado $p$ en la cual tanto los polinomios como sus $p-1$ derivadas presentan restricciones de continuidad en sus intersecciones, limitando así la flexibilidad respecto al ajuste de polinomios independientes en cada subintervalo y garantizando suavidad de la función spline resultante (Hastie et al., 2009). En el análisis funcional de datos de crecimiento humano es normalmente utilizada la base de funciones *B-spline* para la reconstrucción de las observaciones funcionales (Ramsay & Silverman, 2005).

Una vez seleccionada la base *B-spline*, se define su flexibilidad en la reconstrucción de las curvas. En el problema de mínimos cuadrados anteriormente mencionado esto se resume en la selección de la dimensión de la base y el grado de los polinomios del spline. Respecto al grado de los polinomios, en general se utilizan splines cúbicos (en los cuales cada polinomio es de grado 3) para obtener una función suave a simple vista, sin embargo, el grado puede incrementar dependiendo del interés en la suavidad de las derivadas funcionales (Hastie et al., 2009). La dimensión de la base depende de la delimitación de los subintervalos del spline. Como enuncian Hastie et al. (2009), se busca definir los subintervalos de manera correspondiente a la suavidad esperada de la función resultante, ubicando intervalos más cortos donde se espera una mayor curvatura de las funciones resultantes.

#### 2.1.1.2. Suavizamiento de curvas. Suavizamiento spline

La elección de la dimensión de la base es de gran importancia en la reconstrucción de las curvas funcionales, por lo que alternativas que no requieren determinar de antemano los subintervalos del spline han sido planteadas y comparadas en el contexto de datos funcionales (Aguilera & Aguilera-Morillo, 2013). Estas alternativas están basadas en la penalización del problema de mínimos cuadrados por la rugosidad de las curvas resultantes, y en lugar de buscar una reconstrucción de las curvas mediante la localización óptima de subintervalos de un spline, el problema se direcciona a determinar de manera eficiente el grado de regularización requerido en la rugosidad de la estimación de las observaciones funcionales. La rugosidad de la función es medida a través de su $p$-ésima derivada, la cual es elevada al cuadrado para obtener una medida resumen de rugosidad de la curva a través de la integral de la función resultante (Hastie et al., 2009). Bajo el suavizamiento spline, se busca entonces minimizar

$$PENSSE_p(\mathbf{c}_i, \lambda) = SSE(\mathbf{c}_i) + \lambda \int [D^p x_i(t)]^2 dt, \tag{2.3}$$

en donde $SSE(\mathbf{c}_i)$ está dado por la ecuación 2.2 y $\lambda$ es un parámetro de ajuste de la regularización o suavizamiento, el cual es usualmente estimado mediante validación cruzada. Al minimizar la ecuación 2.3, $SSE(\mathbf{c}_i)$ busca un ajuste funcional cercano a los datos originales mientras que el término $\lambda \int [D^p x(t)]^2 dt$ fuerza suavidad en la función resultante. Normalmente se considera $p = 2$ y se toma como base de funciones la base *B-spline* de grado 3 con subintervalos definidos por los valores únicos de la variable, en este caso, la función resultante es denominada spline cúbico suavizado (Hastie et al., 2009).

### 2.1.2. Reconstrucción de curvas de crecimiento

La reconstrucción de curvas de crecimiento se ha basado en la estimación de funciones paramétricas conocidas (Anderson et al., 2019). Ramsay & Silverman (2005) presentan ejemplos en los que un modelo es seleccionado para una medida antropométrica dada en un intervalo de edad determinado, y mediante los parámetros estimados del modelo se caracteriza el patrón de crecimiento infantil, reduciendo la dimensionalidad del problema y simplificando posibles comparaciones entre curvas. Los parámetros estimados para cada curva pueden ser utilizados como medidas resumen del patrón de crecimiento, en lugar de utilizar los datos originales, para evaluar asociaciones con posibles variables de interés (Cole et al., 2010). La expresión funcional de los datos permite, además de un acercamiento a la naturaleza continua del crecimiento infantil, una caracterización de la razón de cambio de las curvas. En el problema actual, la información adicional recolectada mediante la primera derivada funcional se traduce en la velocidad de crecimiento infantil.

La implementación computacional de métodos modernos como la estimación mediante funciones spline ha permitido modelar de manera menos restrictiva las características del patrón de crecimiento humano (Cameron, 2002). Sin embargo, la flexibilidad del planteamiento en funciones base puede no capturar características propias del fenómeno de interés. Un ejemplo claro es la reconstrucción de curvas para la talla infantil en la cual la combinación lineal de funciones puede llevar a un decrecimiento en las curvas reconstruidas, principalmente para las edades mayores donde se presentan cambios sutiles o nulos en talla en las visitas médicas consecutivas, implicando velocidades negativas para las curvas ajustadas en dichas edades (Ramsay & Silverman, 2005). Alternativas recientes en la reconstrucción de curvas combinan modelos spline con una parametrización individual del crecimiento para el análisis de la variable talla en la pubertad (Cole et al., 2010).

El ajuste de curvas con restricciones en su estructura desarrollado en el capítulo 6 de Ramsay & Silverman (2005) muestra de una forma alterna el problema de estimación para volver a un escenario sin restricción funcional. Al contar con una función positiva $x$, esta se puede definir como el exponencial de una función sin restricción alguna $W$ que a su vez puede mostrarse como una expansión sobre una base de funciones apropiada

$$x(t) = e^{W(t)} \tag{2.4}$$

$$W(t) \approx \sum_{k=1}^{p} c_k \phi_k(t). \tag{2.5}$$

Se buscan en este caso los coeficientes de la expansión en funciones base de manera similar al escenario sin restricciones al minimizar la suma de cuadrados, pero incorporando la penalización por rugosidad de la función $W$ en lugar de la función $x$, al ser $x$ una función positiva y $W$ una función sin restricción alguna. El procedimiento requiere la utilización de métodos numéricos, involucrando así un mayor trabajo computacional. Para el escenario de crecimiento infantil en el que se ajusta una curva monótona de crecimiento $x$ cuya primera derivada $Dx$ es estrictamente creciente, Ramsay & Silverman (2005) expresan $Dx$ a través de una transformación exponencial de la función $W$ sin restricciones, de manera semejante al resultado de la ecuación 2.5:

$$Dx(t) = e^{W(t)}. \tag{2.6}$$

Al integrar en ambos lados de la igualdad en la ecuación 2.6 resulta la función de crecimiento original

$$x(t) = \beta_0 + \int_{t_0}^{t} e^{W(u)} du, \tag{2.7}$$

donde $\beta_0$ es estimado a partir de los datos y $t_0$ es el límite inferior sobre el cual se suaviza la función $x$. Nuevamente, se requieren métodos numéricos en la estimación de los coeficientes de la base de funciones.

Una vez son ajustadas las curvas crecientes, las mediciones seriales de la talla infantil son transformadas en funciones monótonas de crecimiento. Estas observaciones de dimensión infinita revelan información de interés respecto al fenómeno aleatorio, como por ejemplo el desarrollo puberal temprano, promedio o tardío para cada paciente, el cual es evidenciado mediante las curvas de velocidad de crecimiento (Simpkin et al., 2017). Maduradores tempranos, aquellas personas con un desarrollo puberal temprano, presentan una máxima velocidad de crecimiento a menor edad que los maduradores promedio. Por otra parte, los maduradores tardíos presentan una máxima velocidad de crecimiento a mayor edad que los maduradores promedio. Generalmente, la edad de máxima velocidad de crecimiento (o APHV por sus siglas en inglés) se calcula de manera retrospectiva, lo cual presenta una desventaja para determinar patrones de crecimiento saludables al controlar por la edad de entrada a la pubertad del paciente (Kelly et al., 2014).

### 2.1.3. Elementos aleatorios funcionales

Las definiciones en la siguiente subsección están basadas en la obra de Horváth & Kokoszka (2012) y Kokoszka & Reimherr (2017). Una vez las observaciones discretas son convertidas en elementos funcionales, estos datos de dimensión infinita son considerados como elementos de un espacio vectorial dotado de un producto interno cuya respectiva norma induce un espacio completo. Este espacio vectorial denominado $L^2$ está conformado por las funciones cuadrado-integrables

$$L^2 = \left\{ f : \int_T f^2(t)dt < \infty \right\}.$$

El producto interno definido para dos elementos $f$, $g \in L^2$ está dado por

$$\langle f, g \rangle = \int_T f(t)g(t)dt,$$

el cual a su vez induce la norma $\|f\|^2 = \langle f, f \rangle$. Este espacio vectorial es conveniente al permitir la noción de ortogonalidad entre las curvas, haciendo familiar su estructura a la del espacio euclídeo. Se resalta que los elementos de $L^2$ no son estrictamente funciones, sino clases de equivalencia conformadas por funciones, en las cuales dos funciones pertenecen a la misma clase si difieren en un conjunto de medida cero. En el contexto probabilístico, los objetos aleatorios $X : \Omega \to \Lambda$ para el caso funcional se definen como funciones medibles sobre un espacio de probabilidad $(\Omega, F, P)$. El conjunto $\Lambda$ se toma como un espacio vectorial de Hilbert separable, generalmente igual a $L^2$, para extender definiciones en estadística tales como el valor esperado y la covarianza (Horváth & Kokoszka, 2012).

Las propiedades de primer orden de la función aleatoria $X$ se resumen en la función media $EX = \mu$ en $L^2$, la cual determina la centralidad de la distribución de la función aleatoria en el intervalo $T$ y es igual a

$$E\left[\langle X, y\rangle\right] = \langle \mu, y\rangle \ \forall y \in L^2.$$

Para caracterizar las propiedades de segundo orden del fenómeno aleatorio se destaca que las realizaciones de los objetos aleatorios en el espacio descrito son consideradas datos funcionales, es decir que $X(\omega)$ para $\omega \in \Omega$ pertenece a $\Lambda = L^2$, con lo cual se tiene que $\|X\|^2 : \Omega \to \mathbb{R}$ denota una variable aleatoria. Se dice que una función aleatoria es cuadrado integrable si su segundo momento es finito, es decir, si $E\|X\|^2 < \infty$ con la norma inducida por el producto interno en $L^2$. Suponiendo $\mu = 0$, el operador de covarianza $\Gamma$ para una función cuadrado integrable $X$ está dado por

$$\Gamma(y)(t) = \int c(t,s)y(s)ds \text{ , con } y \in L^2 \text{ y } \Gamma(y) \in L^2,$$

donde $c$ es igual a la función de covarianza de $X$, igual a

$$c(t,s) = E\left[X(t)X(s)\right].$$

Se nota que $\Gamma$ es un operador integral con kernel igual a la función de covarianza. Por la definición del operador de covarianza, se resalta que las características de segundo orden para la función aleatoria $X$ pueden ser identificadas a partir de la función de covarianza correspondiente (Kokoszka & Reimherr, 2017).

En el caso muestral, los parámetros poblacionales $\mu$ y $c$ son estimados a partir de una colección de $n$ observaciones funcionales $x_1, ..., x_n$ de manera análoga a las medidas clásicas. Por tanto, la función media muestral en el instante $t$ se obtiene como el promedio en $t$ de las $n$ curvas observadas

$$\hat{\mu}(t) = \frac{\sum_{i=1}^{n} x_i(t)}{n}. \tag{2.8}$$

Una vez estimada la función media, la superficie de covarianza muestral evaluada en los puntos $t$ y $s$ es igual a

$$\hat{c}(t,s) = \frac{\sum_{i=1}^{n}(x_i(t) - \hat{\mu}(t))(x_i(s) - \hat{\mu}(s))}{n}. \tag{2.9}$$

En el contexto inferencial de curvas de crecimiento infantil, $X$ denota la función de crecimiento aleatoria de interés y $x_1, ..., x_n$ corresponde a la realización de una muestra aleatoria $X_1, ..., X_n$ de tamaño $n$ como enuncia Kokoszka & Reimherr (2017). Las funciones estimadas $\hat{\mu}$ y $\hat{c}$ son consideradas como los estimadores de los parámetros poblacionales correspondientes para el objeto aleatorio $X$.

### 2.1.4. Análisis en componentes principales funcionales

Las definiciones en la siguiente subsección están basadas en la obra de Ramsay & Silverman (2005) y Kokoszka & Reimherr (2017). La expansión en base de funciones en la subsección 2.1.1 asume conocidos los elementos en la base para obtener observaciones funcionales, sin embargo, es posible encontrar un sistema de funciones ortonormales para la reconstrucción de las funciones a partir de los datos funcionales (Kokoszka & Reimherr, 2017). Este sistema de funciones obtenido a partir de la información disponible en la muestra permite caracterizar las principales fuentes de variación en los datos respecto a la función media. Dado un sistema ortonormal arbitrario de funciones $u_1, ..., u_q$ en $L^2$ de dimensión $q > 1$, es posible descomponer la función aleatoria $X \in L^2$ en su proyección sobre el espacio generado por los elementos del sistema de funciones como

$$Proy_{\{u_1,...,u_q\}}(X) = \sum_{k=1}^{q} \langle X, u_k \rangle u_k. \tag{2.10}$$

Para el operador de covarianza $\Gamma$ correspondiente a la función aleatoria $X$ se definen las funciones propias $v_j$ y valores propios $\lambda_j$ como la solución a la ecuación $\Gamma(v_j) = \lambda_k v_j$, con $\lambda_j$ perteneciente a los números reales y $v_j$ función unitaria que satisface $\langle v_l, v_j \rangle = 0$ para $l \neq j$. Al ser $\Gamma$ un operador de Hilbert-Schmidt, simétrico y definido positivo se tiene

$$\sup\{\langle \Gamma(x), x \rangle : \|x\| = 1, \langle x, v_l \rangle = 0, 1 \leq l \leq j-1\} = \lambda_j,$$

el cual es alcanzado en $x = v_j$. Asumiendo $\lambda_1 > ... > \lambda_q$, el anterior resultado permite obtener la descomposición óptima de $X$ en un sistema ortonormal de $q$ funciones a partir de la minimización de la diferencia esperada entre la función aleatoria $X$ y su proyección presentada en la ecuación 2.10. Se obtiene entonces

$$X \approx \sum_{k=1}^{q} \langle X, v_k \rangle v_k. \tag{2.11}$$

La ecuación 2.11 enuncia que la mejor aproximación en términos de error esperado para la función aleatoria $X$ a partir de su proyección en un sistema ortonormal de $q$ funciones corresponde a la proyección de $X$ en las primeras $q$ funciones propias de $\Gamma$. De manera análoga se obtiene al centrar la función aleatoria $X$ con media $\mu$ la mejor aproximación de la función aleatoria centrada como

$$X - \mu \approx \sum_{k=1}^{q} \langle X - \mu, v_k \rangle v_k. \tag{2.12}$$

La ecuación 2.12 corresponde a los primeros $q$ componentes de la expansión de Karhunen-Loève para la función aleatoria $X$, esta descomposición enuncia que cualquier función cuadrado integrable $X$ con función media $\mu$ y operador de covarianza $\Gamma$ puede expresarse en $t$ como

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \langle X - \mu, v_k \rangle v_k(t). \tag{2.13}$$

Los objetos funcionales $v_k$ son denominados componentes principales funcionales, los cuales son elementos determinísticos que resumen los principales patrones de variabilidad respecto a la función media del proceso funcional. La aleatoriedad en la expansión de Karhunen-Loève es reflejada en los coeficientes de Fourier de la función aleatoria $X$ sobre los componentes principales funcionales $\xi_k = \langle X - \mu, v_k \rangle$ para $k \geq 1$ donde $E(\xi_k) = 0$ y $V(\xi_k) = \lambda_k$. Estas variables aleatorias son denominadas puntajes, y sus varianzas sobre las funciones propias correspondientes proveen una descomposición decreciente de la variabilidad de $X$ en la varianza de su proyección sobre los diferentes componentes principales funcionales.

La función media poblacional $\mu$ y el kernel de covarianza poblacional $c$ correspondiente a $\Gamma$ son estimados a partir de los datos funcionales para posteriormente encontrar los componentes principales funcionales muestrales $\{\hat{v}_k\}_{k=1,...,q}$ como las funciones propias del operador de covariaza muestral. Diferentes métodos han sido planteados para reducir el problema de estimación funcional de los componentes principales a un escenario matricial equivalente para la función de covarianza muestral (Ramsay & Silverman, 2005).

La dimensión del sistema de funciones $q$ en el análisis de componentes principales para el caso funcional se determina normalmente mediante el comportamiento decreciente de los valores propios estimados. Se recomienda seleccionar $q$ como el valor correspondiente a $\lambda_q$ en el que se estabilice el patrón decreciente de los valores propios estimados. Además, la selección de $q$ también es usualmente soportada por el porcentaje acumulado de varianza explicada por las funciones propias estimadas, en el cual la suma acumulada de los valores propios estimados respecto a la suma total es calculada y generalmente se selecciona $q$ para el cual el porcentaje acumulado sea superior al 85 %. Métodos de selección adicionales basados en criterios de información y validación cruzada han sido implementados, sin embargo, lo importante es complementar la selección de $q$ con una revisión de la forma de las funciones propias estimadas: comportamientos irregulares y aleatorios pueden indicar un componente principal poco informativo (Kokoszka & Reimherr, 2017).

Una vez se determina el valor de $q$, una aproximación óptima de cada curva a partir de los primeros $q$ componentes principales muestrales es obtenida mediante la expansión de Karhunen-Loève. Para el $i$-ésimo individuo en el punto $t$ se tiene

$$x_i(t) - \hat{\mu}(t) \approx \sum_{k=1}^{q} \hat{\xi}_{ik} \hat{v}_k(t) \tag{2.14}$$

$$x_i(t) \approx \hat{\mu}(t) + \hat{\mathbf{d}}_i' \hat{\mathbf{v}}(t), \tag{2.15}$$

con $\hat{\mathbf{d}}_i = (\hat{\xi}_{i1}, ..., \hat{\xi}_{iq})'$ y $\hat{\mathbf{v}}(t) = (\hat{v}_1(t), ..., \hat{v}_q(t))'$. Se nota la semejanza con la expansión en funciones base presentada en la ecuación 2.1, sin embargo, este acercamiento permite la obtención de las funciones $\{\hat{v}_k\}_{k=1,...,q}$ que reflejan los principales modos de variación en los datos respecto a la función media a partir de las funciones observadas, además, se espera que $p >> q$ con $p$ igual dimensión de la base de funciones en la expansión en funciones base (Kokoszka & Reimherr, 2017). En la reconstrucción de las curvas se destaca que seleccionar valor muy pequeño de $q$ puede suavizar características importantes del proceso aleatorio, mientras que un valor muy grande puede añadir ruido en las curvas reconstruidas.

## 2.2. Análisis de datos funcionales escasos

En el análisis de datos funcionales densos se cuenta con una cantidad considerable de observaciones por sujeto las cuales se encuentran distribuidas a lo largo del dominio de interés, esto permite caracterizar el comportamiento funcional de manera individual para cada sujeto, lo cual es necesario en la estimación de los parámetros funcionales presentados en la sección anterior. Sin embargo, en el estudio del crecimiento infantil las observaciones de cada individuo pueden estar registradas en intervalos de edad diferentes y en una cantidad variable de instantes en el tiempo llegando incluso a unas pocas observaciones para cada paciente. Esta característica inherente de los datos longitudinales contrasta con el escenario anterior y en el contexto funcional es definida como datos funcionales escasos, en el sentido de proveer información individual incompleta del proceso aleatorio funcional para el dominio de interés.

Un acercamiento estadístico para incorporar en el análisis la estructura de dependencia de las observaciones de un mismo sujeto es realizado a partir del modelamiento de efectos mixtos, que mediante la incorporación de efectos fijos y aleatorios, busca caracterizar las trayectorias individuales de crecimiento junto a las fuentes de variación presentes en datos. Esta perspectiva longitudinal de modelamiento se verá posteriormente reflejada en el análisis de datos funcionales escasos. Siguiendo la notación de Rabe-Hesketh & Skrondal (2008), la especificación del modelo de efectos mixtos para $x_{ij}$, la $j$-ésima observación del $i$-ésimo individuo, está dada por

$$x_{ij} = \mu + \zeta_i + \epsilon_{ij}, \tag{2.16}$$

con lo cual se obtiene un efecto fijo $\mu$ que caracteriza la estructura media del modelo, un efecto aleatorio $\zeta_i$ con media cero y varianza $\psi$ el cual es independiente entre sujetos e induce la estructura de covarianza en las medidas repetidas y finalmente un componente de error aleatorio $\epsilon_{ij}$ con media cero y varianza $\theta$ el cual es independiente por cada individuo y ocasión en la cual es medido. Como destaca Rice & Wu (2001) en el contexto de datos funcionales escasos con un enfoque de modelamiento de efectos mixtos, se asume que la observación discretizada $x_{ij}$ del individuo $i$ en el instante $t_{ij}$ proviene del siguiente modelo

$$x_{ij} = \mu(t_{ij}) + \zeta_i(t_{ij}) + \epsilon_{ij}, \tag{2.17}$$

para $i = 1, ..., n$ y $j = 1, ..., m_i$ con $\mu(t_{ij})$ igual a la función media evaluada en $t_{ij}$, $\zeta_i(t_{ij})$ igual a la función de error específica para el sujeto $i$ evaluada en $t_{ij}$ y $\epsilon_{ij}$ igual a un error de medición aleatorio con varianza $\sigma^2$ y esperanza cero. Se obtiene así que bajo este modelo el conjunto longitudinal de mediciones de un individuo dado es representado como la suma de una función media poblacional, una función aleatoria individual, y un ruido aleatorio (Rice & Wu, 2001). De manera semejante al modelo longitudinal, el componente de error específico $\zeta_i$ induce la correlación para observaciones de un mismo individuo mientras que el error de medición $\epsilon_{ij}$ es independiente e idénticamente distribuido entre ocasiones e individuos. Los dos componentes de error son independientes entre si y $\zeta_i$ es independiente entre sujetos diferentes. Se asume adicionalmente que el intervalo acotado y cerrado $T$ sobre el cual se definen los datos funcionales escasos es observado de manera relativamente densa para los instantes $t_{ij}$ (Kokoszka & Reimherr, 2017).

Para finalizar, se resalta que bajo el modelo en la ecuación 2.17 la suma entre la función media $\mu$ y la desviación específica $\zeta_i$ correspondiente al $i$-ésimo individuo es igual a la función aleatoria $X_i$, por lo tanto, en el instante $t_{ij}$ se tiene que

$$x_{ij} = X_i(t_{ij}) + \epsilon_{ij}. \tag{2.18}$$

### 2.2.1. Estimación de parámetros

Como destaca Kokoszka & Reimherr (2017), para los datos funcionales escasos no se aplica un suavizamiento a trayectorias individuales, por el contrario, se busca combinar de manera apropiada la información de todos los individuos para luego obtener trayectorias individuales reconstruidas. La estimación se basa en el conocimiento de las propiedades de primer y segundo orden bajo el modelo propuesto para las observaciones, donde la estimación de los parámetros funcionales debe ser adaptada para el contexto escaso de los datos funcionales. Bajo el modelo propuesto en la ecuación 2.17, el valor esperado marginal es igual a

$$E(x_{ij}|t_{ij}) = E(X_i(t_{ij})) = \mu(t_{ij}). \tag{2.19}$$

Por otra parte, la covarianza marginal está dada por

$$\mathrm{Cov}(x_{ij}, x_{il}|t_{ij}, t_{il}) = E\left((x_{ij} - E(x_{ij}|t_{ij}))(x_{il} - E(x_{il}|t_{il}))\right), \tag{2.20}$$

por lo cual se tiene

$$\begin{aligned}
\mathrm{Cov}(x_{ij}, x_{il}|t_{ij}, t_{il}) &= E((\zeta_i(t_{ij}) + \epsilon_{ij})(\zeta_i(t_{il}) + \epsilon_{il})|t_{ij}, t_{il}) \\
&= E(\zeta_i(t_{ij})\zeta_i(t_{il})|t_{ij}, t_{il}) + E(\epsilon_{ij}\epsilon_{il}|t_{ij}, t_{il}) \\
&= \mathrm{Cov}(\zeta_i(t_{ij}), \zeta_i(t_{il})|t_{ij}, t_{il}) + \mathrm{Cov}(\epsilon_{ij}, \epsilon_{il}|t_{ij}, t_{il}) \\
&= \mathrm{Cov}(X_i(t_{ij}), X_i(t_{il})|t_{ij}, t_{il}) + \mathrm{Cov}(\epsilon_{ij}, \epsilon_{il}|t_{ij}, t_{il}).
\end{aligned}$$

Se concluye así que la función de covarianza marginal evaluada en $j, l$ es igual a la suma de dos cantidades: la función de covarianza de la función aleatoria $X$ evaluada en $(t_{ij}, t_{il})$ y la varianza del error de medición, igual a $\sigma^2$ cuando $j$ es igual a $l$ y 0 en otro caso. Es decir

$$\mathrm{Cov}(x_{ij}, x_{il}|t_{ij}, t_{il}) = c(j, l) + \sigma^2\delta_{jl}. \tag{2.21}$$

En la estimación de la función media a partir de datos funcionales escasos es claro que el promedio punto a punto calculado para datos funcionales densos puede resultar en estimaciones poco confiables. Alternativas basadas en polinomios locales, espacios de Hilbert con kernel reproducible o expansión en funciones base son adoptadas (Kokoszka & Reimherr, 2017). Por ejemplo, Yao et al. (2005) utiliza regresión local univariada, Xiao et al. (2016) y Zhang et al. (2015) aplican suavizamiento spline. La característica común entre los métodos de estimación de la función media en $t$ es la combinación de información disponible para toda la muestra.

La estimación de la superficie de covarianza tampoco permite una analogía directa con los datos funcionales densos, por lo que diferentes alternativas han sido propuestas para la solución de este problema, entre estas está el trabajo de Yao et al. (2005) que implementa métodos de suavizamiento bivariados basados en los productos de desvíos respecto a la media estimada, por otra parte Xiao et al. (2018) discretiza el problema de estimación mediante la expansión en funciones base. A diferencia del escenario funcional denso, los métodos consideran en su estimación la varianza del error de medición, ya que este error es significativo y no puede ser suavizado en la presencia de datos funcionales escasos.

De manera similar al caso funcional denso, el análisis en componentes principales es de gran relevancia para el análisis funcional escaso y nuevamente se basa en la superficie de covarianza estimada. El procedimiento para encontrar las funciones propias y los valores propios generalmente se basa en discretizar la función de covarianza estimada (Xiao et al., 2016), sin embargo, es importante resaltar que la estimación de los puntajes debe hacerse de manera especialmente cuidadosa para el escenario funcional escaso (Kokoszka & Reimherr, 2017): la definición de los puntajes en el análisis en componentes principales funcional evidenciada en la ecuación 2.12 asume curvas observadas de manera densa, haciendo que el procedimiento tradicional para la evaluación del producto interno no sea aplicable en el caso escaso. Diferentes propuestas han sido implementadas para solucionar este problema de estimación: Yao et al. (2005) calcula los puntajes mediante esperanza condicionada, mientras que Zhang et al. (2015) propone modelos de regresión en la estimación de los puntajes. Ambos trabajos usan los puntajes estimados junto a los componentes principales funcionales muestrales en la reconstrucción de las trayectorias funcionales. Con un enfoque diferente, Xiao et al. (2018) sugiere una estimación de las curvas sin necesidad de realizar análisis en componentes principales funcionales.

La gran importancia del análisis en componentes principales para el escenario funcional escaso esta basada en la estandarización de las observaciones longitudinales a partir de los puntajes de las observaciones sobre las funciones propias estimadas. En el estudio del crecimiento infantil, el análisis de componentes principales funcional escaso confiere una escala común para los registros antropométricos de sujetos medidos en diferentes edades. Esta escala común de las observaciones escasas sobre los componentes principales funcionales permite una comparación de las curvas reconstruidas en todo el dominio de interés.

# CAPÍTULO 3

# Revisión de literatura

En este capítulo se realiza una corta revisión del trabajo nacional e internacional en el desarrollo de curvas de crecimiento resaltando tanto características descriptivas como prescriptivas en la elaboración de las curvas. Se destacan las curvas colombianas actuales junto a sus limitaciones. Se incorpora finalmente el trabajo adelantado en el contexto de datos funcionales escasos y las metodologías de estimación de curvas. En la sección 3.1 son mencionados diferentes ejemplos de construcción de curvas en países desarrollados y en vía de desarrollo junto a sus ventajas, limitaciones y constantes mejoras. La sección 3.2 muestra los métodos actuales en la estimación de curvas de crecimiento diferenciando por la naturaleza transversal o longitudinal de los datos. Finalmente en la sección 3.3 se presentan dos metodologías en el análisis de datos funcionales escasos para la reconstrucción de trayectorias de crecimiento.

## 3.1. Aplicación de las curvas de crecimiento

La utilidad práctica de las curvas de crecimiento en endocrinología pediátrica radica en la evaluación temprana de posibles riesgos en salud en el desarrollo infantil, esto ha permitido evidenciar la gran importancia de la recolección y el análisis de la información antropométrica. Generalmente, las curvas son elaboradas de manera inicial para dominios particulares de la población, en otros casos se construyen resúmenes estadísticos de la medida antropométrica de interés a partir de muestras no probabilísticas de la población determinada. Posteriormente, estas curvas de crecimiento son refinadas para superar las limitaciones metodológicas y posibles defectos técnicos presentados en su elaboración. La evolución en la elaboración de curvas de crecimiento infantil evidenciada en los esfuerzos para subsanar limitaciones técnicas y metodológicas en su construcción muestra el campo de mejora en la elaboración de curvas de crecimiento nacionales, esto con el objetivo de identificar de manera eficiente individuos con problemas de salud al incrementar la sensibilidad (capacidad de identificar correctamente pacientes enfermos) y la especificidad (capacidad de descartar correctamente pacientes sanos) de las curvas (Kelnar et al., 2007).

### 3.1.1. Curvas de referencia y curvas estándar

Curvas de referencia de crecimiento han sido desarrolladas en diferentes países, las cuales están caracterizadas por la búsqueda de representatividad muestral para niños y adolescentes de la población en un amplio rango de edad (Kelnar et al., 2007). Kuczmarski et al. (2000) presentaron curvas de referencia para Estados Unidos obtenidas a partir de información transversal proveniente de cinco encuestas de salud diferentes con diseños muestrales probabilísticos complejos, buscando con esto una representatividad nacional de las curvas para las diferentes edades, lo cual es de particular importancia en la práctica clínica para la estimación de percentiles extremos. Freeman et al. (1995) elaboraron curvas de referencia para el Reino Unido procurando alcanzar una representatividad nacional utilizando diferentes fuentes de información, de manera similar al caso norteamericano.

Las curvas de referencia para Estados Unidos y Reino Unido son elaboradas como mejoras ante versiones anteriores en las que se presentaban, entre otros, inconvenientes de representatividad nacional en la muestra: el trabajo de Hamill et al. (1977) presenta las curvas anteriores para Estados Unidos, las cuales evidencian bastantes inconvenientes metodológicos para niños menores de 36 meses dado que su información fue tomada de un estudio longitudinal que no buscaba una representatividad nacional, además, en este estudio los intervalos de edad usados para la medición no eran adecuados para identificar de manera apropiada el rápido cambio en el patrón de crecimiento en este periodo de edad (Kuczmarski, 2002). Para las curvas de crecimiento del Reino Unido presentadas por Tanner et al. (1966), además de evidenciar problemas de representatividad nacional, estas fueron basadas en información de curvas precedentes, por lo que Freeman et al. (1995) destaca la necesidad de una actualización en las curvas dada la tendencia secular del crecimiento.

Curvas de crecimiento internacionales permiten simplificar comparaciones entre las diferentes regiones y países asumiendo un patrón de crecimiento homogéneo en sus niños y adolescentes. Curvas estándar globales han sido propuestas por la Organización Mundial de la Salud (OMS) cuya aplicabilidad universal está fundamentada en la incorporación de una gran diversidad étnica, genética y cultural de los individuos de la muestra; junto a una clara evidencia de un patrón homogéneo de crecimiento para niños saludables y bien nutridos en edad preescolar para diferentes países (De Onis et al., 2004). Las curvas estándar para la primera infancia de la OMS contienen participantes de ciudades en seis países, los niños son seleccionados de una población socioeconómicamente privilegiada de personas saludables bajo condiciones favorables en el desarrollo de su potencial de crecimiento. Además, las madres de los participantes del estudio se involucran en prácticas saludables y relacionadas con el crecimiento del niño tales como amamantamiento y no fumar. Los países involucrados en el estudio combinan una muestra longitudinal con una transversal de la misma población, donde el componente longitudinal comprende a los niños desde su nacimiento hasta los 24 meses de edad. A partir de los 24 meses hasta los 71 fue realizado un estudio transversal. La combinación de ambos diseños en el análisis obedece a un comportamiento de crecimiento menos lineal para los niños menores respecto a los mayores, el cual requiere información en edades determinadas para una correcta caracterización del patrón de crecimiento, también está dado por los costos adicionales asociados, en tiempo y dinero, de un estudio longitudinal respecto a un estudio transversal (De Onis et al., 2004). En resumen, el estudio busca un patrón de crecimiento universal que sea replicable en los diferentes países bajo condiciones ideales de desarrollo y crecimiento.

En la elaboración del estándar internacional de crecimiento, la información recolectada para Latinoamérica proviene únicamente de la ciudad de Pelotas en Brasil. La descripción del protocolo e implementación del estudio en este lugar es detallada por Araújo et al. (2004). Se destaca que el componente longitudinal es obtenido a partir de una muestra obtenida de tres hospitales que representan cerca del 90 % de los nacimientos de esta ciudad, mientras que en el componente transversal, buscando una semejanza con los niños del componente longitudinal, fue configurado un diseño muestral basado en el vecindario de los participantes del componente longitudinal. De manera semejante a las curvas de crecimiento para Estados Unidos y Reino Unido, las curvas de crecimiento de la OMS se presentan como una mejora respecto a una versión anterior, puesto que consideran metodologías estadísticas modernas en su construcción y además recolectan una muestra multinacional buscando aplicabilidad internacional. Las curvas de crecimiento fueron probadas antes de su despliegue en cuatro países diferentes a los utilizados en su construcción (Italia, Argentina, Maldivas y Pakistán) con el objetivo de comparar las curvas con la evaluación clínica (Onyango et al., 2007). La concordancia general de los resultados obtenidos sugirió la validez de las curvas de crecimiento para estos países.

Las curvas desarrolladas por Kuczmarski et al. (2000) y Freeman et al. (1995) han sido complementadas con la información de las curvas estándar de crecimiento de la OMS y posteriormente implementadas en los países correspondientes. Para Estados Unidos se recomienda utilizar las curvas de la OMS en niños menores de 24 meses, de lo contrario se sugiere continuar con las curvas originales (Grummer-Strawn et al., 2010). Para el Reino Unido también fue efectuada una amalgama de las curvas de la OMS con las anteriores curvas (Wright et al., 2010). Las mejoras efectuadas a las curvas originales con la información del estándar internacional corresponden a la incorporación de niños alimentados predominante o exclusivamente con leche materna por al menos 4 meses en las curvas de la OMS, la cual es una característica de gran importancia del estándar de crecimiento como modelo normativo para el crecimiento y el desarrollo infantil (WHO, 2006a).

Como puede observarse, en el caso de un estándar de crecimiento infantil se busca una muestra homogénea de una población objetivo, cuyas características sean favorables en el crecimiento de sus niños y adolescentes. Por ejemplo, las curvas de crecimiento estándar de la OMS consideran una población en ideales condiciones socioeconómicas, con una baja morbilidad y dispuesta a seguir diferentes recomendaciones alimenticias. Con lo cual, a pesar de la amplia diversidad genética, étnica y cultural en los países de estudio para la construcción de las curvas, éstas son unificadas para la elaboración de un estándar internacional dada la gran similitud observada en el patrón de crecimiento de los participantes en cada país. Este supuesto en la construcción de estándares de crecimiento ha sido revisado, y en general se encuentra que no siempre es apropiado y en particular no es justificado para el perímetro cefálico (Natale & Rajagopalan, 2014).

Por otra parte, las curvas de referencia infantil buscan constantemente la caracterización representativa del comportamiento de crecimiento a partir de información principalmente transversal, lo cual es particularmente informativo para diferentes subpoblaciones de estudio en países en vía de desarrollo en los que se evidencian diferencias importantes en el peso y la estatura desde el nacimiento a los 7 años de edad al segregar por regiones, urbano y rural, y por niveles de riqueza (Habicht et al., 1974). En síntesis, mientras en un estándar de crecimiento se priorizan factores socioeconómicos y nutricionales respecto a la diversidad genética, étnica y cultural; en una referencia de crecimiento se busca representatividad en un instante dado para diferentes dominios de una población sin relación alguna con el estado de salud.

### 3.1.2. Curvas latinoamericanas de crecimiento infantil

Las curvas estándar de crecimiento infantil internacionales de la OMS representan una alternativa ante la falta de curvas nacionales como herramienta diagnóstica de crecimiento infantil. Para Latinoamérica estas curvas han sido adoptadas por diferentes países, sin embargo, los estándares internacionales en general han sido encontrados poco apropiados para determinar problemas de crecimiento inusual dadas las características inherentes de cada población (Silva et al., 2010; Orden & Apezteguía, 2016). Particularmente se destaca que aunque Brasil fue el único país latinoamericano considerado en la construcción de las curvas de crecimiento de la OMS, tanto la talla como el peso en este país son en general mayores para ambos sexos en la mayoría de las edades respecto a las curvas de crecimiento estándar (Silva et al., 2010). A pesar de las diferencias encontradas, tan sólo algunos pocos países en Latinoamérica han elaborado curvas propias de crecimiento locales: Argentina (Lejarraga et al., 2009), Colombia (Durán et al., 2016) y Venezuela (de Blanco et al., 2013) han construido sus propias curvas, además, curvas Ecuatorianas de referencia fueron recientemente publicadas (Tarupi et al., 2020). Respecto a los estándares de crecimiento infantil en la región, son escasos los estudios relacionados en grupos socioeconómicamente privilegiados para países de América del Sur (Natale & Rajagopalan, 2014).

Para el caso argentino fue elaborado un conjunto inicial de curvas de referencia por Lejarraga & Orfila (1987). Dado que estas primeras curvas de crecimiento no lograban representar el crecimiento de niños alimentados con leche materna y además presentaban una gran dificultad de aplicación en la práctica clínica por la metodología manual usada de suavizamiento de cuantiles, las curvas fueron actualizadas con la información de la OMS de manera semejante a las curvas de otros países (Grummer-Strawn et al., 2010; Wright et al., 2010): para los primeros 2 años de vida la información longitudinal de la OMS fue utilizada y, en lugar del suavizamiento experto de los percentiles realizado anteriormente, una metodología estadística de suavizamiento fue aplicada (Lejarraga et al., 2009). Se destaca que desde los dos años hasta la madurez se utilizó la misma información que aquella de las curvas originales: cuatro diferentes fuentes de información, tres de las cuales provienen de diferentes ciudades. Estudios posteriores sugieren diferencias importantes en el patrón de crecimiento de adolescentes en zonas urbanas de Argentina y las curvas finales de crecimiento (Onyango et al., 2007). En Ecuador, de manera semejante al caso argentino, se consideran las curvas internacionales de la OMS como norma para los niños menores de 5 años. En estas curvas se enfatiza el efecto significativo de las características geográficas, socioeconómicas y genéticas en el crecimiento infantil a partir de los 5 años. El estudio de Tarupi et al. (2020) considera muestras transversales en diferentes periodos para las distintas regiones que conforman el país, logrando así el primer conjunto de curvas de crecimiento de referencia para Ecuador.

Para el caso colombiano fue recientemente presentada la construcción de curvas de referencia de crecimiento para cuatro variables antropométricas: talla, peso, índice de masa corporal y perímetro craneal (Durán et al., 2016). El estudio prospectivo fue el primero de su clase en el país y fue realizado a partir de una muestra mayoritariamente transversal ya que se hizo un seguimiento longitudinal únicamente durante el primer año de vida de 540 niños en la muestra, para los demás participantes las mediciones de las variables fueron únicas. Los niños y adolescentes seleccionados en la muestra representan estratos medios y altos de las principales ciudades del país (Bogotá, Medellín, Cali y Barranquilla), además, se asume que los participantes viven en condiciones favorables para el desarrollo de su crecimiento.

Es claro que las características particulares de cada población motivan la elaboración de curvas de crecimiento locales y la alternativa de estándares internacionales permite una solución ante su ausencia, sin embargo, adaptar curvas internacionales a escenarios locales puede llevar a diagnósticos erróneos de los niños y adolescentes, por lo cual se hace necesaria una recolección de información nacional para un entendimiento del crecimiento local saludable. Diferentes países han optado por realizar una mezcla de muestras poblacionales en diferentes instantes temporales e incluso se realizan mezclas de diseños longitudinales con transversales con el objetivo práctico de ampliar el tamaño de la muestra estadística y de caracterizar de manera apropiada de los cuantiles distribucionales en las diferentes edades. Este diseño longitudinal mixto permite combinar las ventajas de ambos diseños a costo de una mayor dificultad en el análisis de la información (Kelnar et al., 2007).

## 3.2. Métodos actuales de estimación de curvas

En la obtención de curvas de crecimiento no condicionadas, tanto estándar como de referencia, se buscan en general curvas percentílicas suaves de la variable antropométrica de interés en función de la edad diferenciando por el sexo de los individuos en la muestra. Los primeros procedimientos en la construcción de las curvas asumen una distribución teórica para la variable de interés, la cual es ajustada en diferentes intervalos de edad con el objetivo de obtener cuantiles esperados de la variable aleatoria. Cada cuantil esperado es suavizado a través de los intervalos de edad mediante un procedimiento que varía desde un ajuste manual de los percentiles hasta métodos estadísticos de suavizamiento. Este acercamiento ha permitido la evolución metodológica de herramientas en la estimación de curvas de crecimiento basadas en datos transversales, alcanzando técnicas estadísticas avanzadas tales como modelos generalizados aditivos. El modelamiento basado en datos transversales bajo un enfoque paramétrico es comúnmente usado en la caracterización del patrón de crecimiento dada su versatilidad en la estimación de cuantiles y su amplio reconocimiento en la comunidad académica. Éste es aplicado para datos transversales y de única ocasión, en los cuales se cuenta con observaciones individuales para cada paciente.

En la construcción de curvas de crecimiento condicionadas, las propuestas existentes no permiten modelar eficientemente las características propias de los datos longitudinales de crecimiento infantil (Wei et al., 2006), y aunque los métodos actuales para el escenario longitudinal modelan el fenómeno de interés, no permiten una elaboración análoga de las curvas transversales[1]. Al reconocer cada paciente con sus observaciones en cada visita médica como unidad de análisis, el acercamiento longitudinal en la estimación de las curvas presenta oportunidades de mejora, que desde el análisis de datos funcionales escasos se traduce en la evaluación de la atipicidad de las observaciones funcionales reconstruidas.

La siguiente sección presenta los métodos de construcción de curvas de crecimiento basados en datos transversales de mayor relevancia junto al proceso de estimación de curvas de la OMS. Posteriormente se describe un modelo longitudinal general para medidas repetidas en la reconstrucción de las trayectorias de crecimiento, de particular importancia en el modelamiento funcional escaso, junto a un modelo específico para la variable talla en la pubertad, el cual será aplicado en la clasificación homogénea de individuos por maduración de manera previa al modelamiento funcional.

[1] Curvas estimadas con información longitudinal omiten la estructura de los datos para obtener curvas de crecimiento semejantes a las curvas meramente transversales: las curvas de la OMS y las curvas de referencia Colombianas, por ejemplo, amalgaman en un mismo modelo información longitudinal y transversal.

### 3.2.1. Métodos basados en datos transversales

El suavizamiento de cuantiles observados es implementado como un paso inicial en el modelamiento de curvas de crecimiento para algunas curvas existentes (Kuczmarski et al., 2000), sin embargo, dado el incremento del error estándar para los cuantiles observados más extremos, en general se prefiere utilizar cuantiles esperados bajo una distribución determinada en la estimación de las curvas. Esto además porque dichos cuantiles son de particular interés en la interpretación de las curvas de crecimiento, dado que las observaciones extremas en la curva pueden reflejar alguna condición inusual en la salud del individuo correspondiente. A continuación se describen los métodos paramétricos de regresión comúnmente utilizados en la estimación de las curvas de crecimiento junto al procedimiento de estimación de las curvas de crecimiento de la OMS para la variable talla.

#### 3.2.1.1. Modelos paramétricos de regresión

Una de las técnicas más relevantes en la caracterización del patrón de crecimiento a través de la edad usando cuantiles esperados es la metodología LMS, cuyas siglas corresponden a las curvas de asimetría, $L$, localización, $M$, y escala, $S$, requeridas en la obtención de normalidad marginal de la variable antropométrica en función de la edad mediante una transformación adecuada de Box-Cox. Los parámetros estimados permiten determinar los cuantiles esperados de la variable de interés en términos de los cuantiles de la distribución normal estándar.

Siguiendo el desarrollo de Cole (1988) se tiene que para el $i$-ésimo elemento de una muestra observada $y_1$, ..., $y_n$ se asume la siguiente transformación de potencia

$$y_i^{(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda & \text{si } \lambda \neq 0 \\ \log y_i & \text{si } \lambda = 0 \end{cases} \tag{3.1}$$

Bajo el supuesto de normalidad de las observaciones transformadas, el parámetro $\lambda$ es estimado de manera que maximice la verosimilitud de la muestra. Cole (1988) sugiere encontrar $\lambda$ como el valor que minimiza la varianza $\sigma^2$ de las observaciones escaladas $g_i^{(\lambda)}$ dadas por

$$g_i^{(\lambda)} = \frac{y_i^{(\lambda)}}{\dot{y}^{\lambda}},$$

para $i = 1, ..., n$ y con $\dot{y}$ igual a la media geométrica de las observaciones originales. Este resultado es equivalente al estimador de máxima verosimilitud de $\lambda$ bajo el modelo propuesto, su ventaja radica al notar que la transformación propuesta es adimensional, con lo cual su varianza es análoga al coeficiente de variación de las observaciones originales. A partir de la muestra observada se encuentra una estimación de la potencia ($\lambda$), la mediana ($\mu$) y la desviación estándar ($\sigma$) de manera independiente para $p$ diferentes grupos de edad, obteniendo así triplas $(\lambda_i, \mu_i, \sigma_i)$ con $i = 1, ..., p$ que representan $p$ distintas estimaciones discretizadas para las curvas $L$, $M$ y $S$ respectivamente. Las curvas finales son obtenidas mediante un suavizamiento de las $p$ estimaciones para cada parámetro de manera independiente. Finalmente, cualquier cuantil distribucional en el instante $t$ puede ser estimado a partir de $L(t)$, $M(t)$ y $S(t)$ bajo normalidad de las observaciones transformadas.

Mediante máxima verosimilitud penalizada, un desarrollo posterior del método permite la estimación de las funciones $L$, $M$ y $S$ sin requerir un suavizamiento de las estimaciones puntuales y bajo los mismos supuestos distribucionales (Cole & Green, 1992). Al igual que Cole (1988), se asume que la distribución de la variable antropométrica de interés varía con una covariable, como por ejemplo la edad, y que los los parámetros que caracterizan en $t$ la distribución de la variable respuesta están dados por valores de funciones de localización, dispersión y simetría en el punto determinado. Sin embargo, al incorporar las funciones en la verosimilitud de los datos y regularizar la rugosidad de los parámetros funcionales, la solución del problema lleva a la estimación de los parámetros mediante splines cúbicos suavizados. Como enuncia Hastie et al. (2009) el problema se reduce a la estimación eficiente de la regularización, es decir, de los grados de libertad efectivos de cada spline (notados como $gl(\cdot)$ en el presente trabajo). A diferencia del trabajo de Cole (1988), esta metodología no requiere la partición subjetiva en grupos de edad dada la incorporación de la penalización en el problema de estimación. Desde su formulación, este procedimiento se ha convertido en un estándar para la elaboración de curvas de crecimiento.

La propuesta LMS de Cole & Green (1992) para la construcción de curvas de crecimiento ha sido generalizada para controlar no solo la asimetría sino además la curtosis en los datos mediante la adición de un parámetro adicional $\tau$ al modelo original (Rigby & Stasinopoulos, 2004; Rigby & Stasinopoulos, 2006). Sea $Y$ una variable aleatoria real positiva definida a través de la transformación de la variable aleatoria $Z$ como

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & \text{si } \lambda \neq 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & \text{si } \lambda = 0 \end{cases} \tag{3.2}$$

Rigby & Stasinopoulos (2004) asumen $Z$ con distribución de potencia exponencial estándar con parámetro de curtosis $\tau$ y Rigby & Stasinopoulos (2006) asumen $Z$ con distribución $t$ de parámetro $\tau$. Dada la distribución de la variable transformada $Z$, la distribución $\mathcal{D}$ de $Y$ en función de los parámetros $\mu > 0$ (ubicación), $\sigma > 0$ (escala), $\nu$ (asimetría) y $\tau$ (curtosis) es conocida. El método ha sido enmarcado como un modelo de regresión generalizado aditivo para localización, escala y forma, también llamado GAMLSS por sus siglas en inglés (Rigby & Stasinopoulos, 2005), mediante el cual los diferentes parámetros permiten modelar de manera flexible la variable respuesta en términos de variables explicativas, adicionalmente, incorpora funciones de enlace para los parámetros buscando una apropiada definición de los mismos. Dadas las covariables en el modelo de regresión, se asume que las observaciones de la variable $Y$ son independientes en la estimación de los parámetros, restringiendo su uso al caso transversal.

La mayoría de las curvas de crecimiento actuales consideran como única covariable la variable edad y su metodología de estimación está basada en el método GAMLSS o particularmente en el método LMS. Ejemplos de aplicación basados en el principio de construcción de curvas mediante LMS son las curvas de referencia para Estados Unidos y Reino Unido (Kuczmarski et al., 2000; Freeman et al., 1995), las curvas de referencia colombianas (Durán et al., 2016), las ecuatorianas (Tarupi et al., 2020), las argentinas (Lejarraga et al., 2009), entre otros. La metodología GAMLSS usando la distribución de potencia exponencial fue seleccionada como método de construcción para los estándares de crecimiento internacionales de la OMS (WHO, 2006c). Su implementación metodológica es descrita en el material suplementario de la construcción de las curvas (WHO, 2006b) y es resumida para la variable talla a continuación.

#### 3.2.1.2. Procedimiento de estimación de la OMS

Se asume que la variable antropométrica de interés tiene una distribución de Box-Cox potencia exponencial de parámetros $\mu$, $\sigma$, $\nu$ y $\tau$ (Rigby & Stasinopoulos, 2004), esto es denotado como BCPE($\mu$,$\sigma$,$\nu$,$\tau$). Bajo el enfoque GAMLSS un primer modelo de regresión arbitrario tanto para $\mu$ como para $\sigma$ es asumido manteniendo $\nu = 1$ y $\tau = 2$ (lo cual se simplifica a una distribución normal de parámetros $\mu$ y $\sigma$), usando la variable antropométrica de interés como variable respuesta, la variable edad como única variable regresora y splines cúbicos como método de suavizamiento para las curvas $\mu(t)$ y $\sigma(t)$ con $t \in T$. Estas curvas son resumidas a partir de los $gl$ de los splines correspondientes. Los $gl$ de este modelo inicial para $\mu$ y $\sigma$ se mantienen constantes y se busca la transformación de potencia $x^{\lambda}$ de la variable edad con la cual se alcanza un menor deviance global, el objetivo de la transformación es obtener una mejor bondad de ajuste en el modelo final para edades de rápido crecimiento. Los valores $\lambda$ considerados son 0.05 a 1.00 con incrementos de 0.05.

Una vez encontrado $\lambda$, se define una grilla de valores para $gl(\mu)$ y $gl(\sigma)$ sobre la cual se selecciona el mejor modelo en términos de $AIC$ y $GAIC(k)$, siendo este último una generalización del $AIC$ que penaliza por la complejidad del modelo y está dado por $GAIC(k) = -2L - kp$ donde $L$ corresponde a la verosimilitud de los datos y $p$ es igual a los $gl$ del modelo. Dadas las características de la penalización, $GAIC(3)$ permite la obtención de curvas más suaves respecto al $AIC$, mientras que el $AIC$ mejora el ajuste local de las curvas obtenidas respecto al $GAIC(3)$ (WHO, 2006b). Para seleccionar la mejor combinación de $gl(\mu)$ y $gl(\sigma)$ la OMS recomienda usar ambos criterios en paralelo, y en caso de discordancia, sugiere utilizar $AIC$ para seleccionar $gl(\mu)$ y $GAIC(3)$ para seleccionar $gl(\sigma)$. Únicamente $GAIC(3)$ es utilizado para la selección de $gl(\nu)$ y de $gl(\tau)$ en caso de ser necesario.

Una vez seleccionados $gl(\mu)$ y $gl(\sigma)$, se evalúa la adecuación del modelo propuesto usando los residuales cuantílicos aleatorizados (Dunn & Smyth, 1996). Mediante gráficos de oruga (Buuren & Fredriks, 2001) y estadísticos $Z$ (Royston & Wright, 2000) se determina la necesidad de modelar la asimetría, y posteriormente la curtosis, de los datos. En general no hay necesidad práctica de caracterizar la curtosis en los datos antropométricos por lo que se modelan únicamente los $gl$ para los parámetros de tendencia, dispersión y simetría (WHO, 2006b). Particularmente para la variable talla en función de la edad, en general se asume normalidad en los datos, por lo cual usualmente no se modela la simetría en estas curvas (Cole, 1988; WHO, 2006b). En caso de considerar la presencia de asimetría en los datos, un siguiente paso busca $gl(\nu)$ manteniendo constantes $gl(\mu)$ y $gl(\sigma)$, una vez encontrado $gl(\nu)$ son nuevamente estimados $gl(\mu)$ y $gl(\sigma)$. El valor de $\lambda$ es actualizado en un último paso del algoritmo manteniendo los $gl$ encontrados en los pasos anteriores.

El algoritmo previamente descrito es implementado en la construcción de las curvas de la OMS para diferentes variables antropométricas. Este busca la cantidad óptima de suavizamiento a través de los $gl()$ de los splines correspondientes en lugar de utilizar validación cruzada para estimar cada parámetro de regularización. Este principio para la selección de modelos es brevemente mencionado en el capítulo 5, sección 5 de la obra de Hastie et al. (2009), en el cual se destaca su utilidad al utilizar diferentes métodos de suavizamiento en un mismo modelo generalizado aditivo. Para el presente estudio, el método de estimación propuesto es conveniente en la selección parsimoniosa de los parámetros a modelar bajo GAMLSS.

### 3.2.2. Métodos basados en datos longitudinales

Como resaltan Wright & Royston (1997), el entendimiento del patrón de crecimiento se logra mediante un seguimiento del desarrollo antropométrico, aun así, las curvas de crecimiento infantil convencionales son estimadas a partir de estudios de corte transversal que proveen información relevante para mediciones simples de un individuo dado (Kelnar et al., 2007). En la práctica clínica, sin embargo, se realiza un seguimiento de las múltiples observaciones de un individuo a partir de curvas basadas en registros transversales, en las que se espera el paciente se encuentre de manera aproximada en el mismo cuantil por la mayor parte de su crecimiento[2]. Cole (1994) destaca que, al contar con mediciones múltiples, es errónea la interpretación de los datos antropométricos longitudinales usando curvas transversales, es relevante así un modelamiento longitudinal en la construcción de curvas de crecimiento al contar con observaciones repetidas de un individuo dado. A continuación se describe el modelamiento longitudinal general en la reconstrucción de las trayectorias individuales, de importancia en la fundamentación del modelamiento funcional escaso, posteriormente se destaca un modelo particular para la caracterización de la variable talla en la adolescencia basado en la superposición mediante traslación y rotación de la función media y denominado SITAR por sus siglas en inglés (*superimposition by translation and rotation*). Este modelo es utilizado de manera novedosa para estimar la madurez de los individuos previo al modelamiento funcional escaso, lo cual a su vez permite diferenciar los fenómenos aleatorios funcionales subyacentes.

#### 3.2.2.1. Modelamiento longitudinal general

En el trabajo de Rice & Wu (2001) se presenta el modelo de efectos mixtos para datos longitudinales presentado en la ecuación 2.17, el cual es configurado a un escenario funcional escaso con covariables haciendo uso de funciones spline para modelar las curvas individuales mediante efectos aleatorios. Bajo este modelo, la función fija $\mu$ y la función aleatoria $\zeta_i$ para $i = 1, ..., n$ son modeladas en $t_{ij}$ como

$$\mu(t_{ij}) = \sum_{k=1}^{p} \beta_k \bar{B}_k(t_{ij}) \quad \text{y} \quad \zeta_i(t_{ij}) = \sum_{k=1}^{q} \gamma_{ik} B_k(t_{ij}),$$

con $\{\bar{B}_j\}_{j=1,...,p}$ y $\{B_j\}_{j=1,...,q}$ bases de funciones para las funciones spline en $T$. Los valores $\beta_1, ..., \beta_p$ son efectos fijos mientras que los valores $\gamma_1, ..., \gamma_q$ son aleatorios. Como $E(x_{ij}|t_{ij}) = \mu(t_{ij})$ se tiene que la estructura de covarianza marginal bajo el modelo de efectos aleatorios está dada por

$$\text{Cov}(x_{ij}, x_{il}|t_{ij}, t_{il}) = \sum_{k=1}^{q} \sum_{s=1}^{q} \Gamma_{ks} B_k(t_{ij}) B_s(t_{il}) + \sigma^2 \delta_{jl}, \tag{3.3}$$

donde $\Gamma_{ks}$ es igual a $\text{Cov}(\gamma_{ik}, \gamma_{is})$ e induce la estructura de covarianza marginal para los coeficientes aleatorios.

[2]Exceptuando su adolescencia, en la cual, aun siendo un paciente saludable, puede presentar cambios importantes en los cuantiles de crecimiento dependiendo de su reloj biológico. Al final de su adolescencia es esperado que el paciente vuelva a su cuantil original (Cole, 1994).

Una vez definidas las dimensiones de las bases de funciones, el modelo de efectos mixtos no paramétrico propuesto puede verse como un modelo lineal clásico de efectos mixtos. De manera matricial, para el $i$-ésimo individuo se puede expresar el modelo como

$$\mathbf{x}_i = \bar{\mathbf{B}}_i\boldsymbol{\beta} + \mathbf{B}_i\boldsymbol{\gamma}_i + \mathbf{e}_i,$$

con $\mathbf{x}_i$ el vector observaciones para el $i$-ésimo individuo de dimensión $m_i$, $\bar{\mathbf{B}}_i$ un arreglo matricial de dimensión $m_i \times p$ y $\mathbf{B}_i$ un arreglo matricial de dimensión $m_i \times q$. Para el modelo lineal, el vector $\boldsymbol{\beta}$ de $p$ parámetros caracteriza el promedio marginal poblacional y el vector $\boldsymbol{\gamma}_i$ de $q$ parámetros determina la desviación aleatoria de las mediciones para el individuo $i$ respecto a la media general con $i = 1, ..., n$.

Finalmente $\mathbf{e}_i$ denota el vector de errores para el $i$-ésimo sujeto bajo el modelo multinivel. La estimación de los parámetros de modelo $\beta$, $\sigma$ y $\Gamma$ es realizada mediante el algoritmo EM para finalmente obtener el mejor predictor lineal insesgado para cada vector $\boldsymbol{\gamma}_i$ como

$$\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\Gamma}}\mathbf{B}_i^t \left(\mathbf{B}_i\hat{\boldsymbol{\Gamma}}\mathbf{B}_i^t + \hat{\sigma}^2\mathbf{I}_{m_i}\right)^{-1} (\mathbf{x}_i - \bar{\mathbf{B}}_i\hat{\boldsymbol{\beta}}). \tag{3.4}$$

Y a su vez, las trayectorias individuales son estimadas como

$$\hat{X}_i(t) = \sum_{k=1}^{p} \hat{\beta}_k \bar{B}_k(t) + \sum_{k=1}^{q} \hat{\gamma}_{ik} B_k(t). \tag{3.5}$$

Como resalta Rice & Wu (2001), $\hat{X}_i$ combina información de la muestra y del sujeto particular. Esta característica es recurrente en el análisis de información funcional escasa.

#### 3.2.2.2. Superposición mediante traslación y rotación (SITAR)

El trabajo de Cole et al. (2010) presenta una alternativa para el modelamiento de medidas repetidas de talla en la pubertad, en la cual se resumen las curvas de crecimiento de manera individual para cada sujeto a partir de un modelo de efectos fijos y aleatorios incorporando el uso de splines. El modelo SITAR asume que la $j$-ésima observación del $i$-ésimo sujeto en el instante $t_{ij}$ es modelada como

$$x_{ij} = \alpha_i + h\left(\frac{t_{ij} - \beta_i}{\exp(-\gamma_i)}\right) + \epsilon_{ij}, \tag{3.6}$$

en el modelo propuesto $h$ corresponde a un spline cúbico natural[3] que representa una trayectoria media de crecimiento de talla en función de la edad. Los parámetros $\alpha_i$, $\beta_i$ y $\gamma_i$ corresponden a efectos aleatorios que modifican la función media para ajustarla a la trayectoria individual correspondiente. Este modelo no lineal considera el mecanismo generador de los datos para el $i$-ésimo sujeto como un modelo constante para la talla en la adolescencia representado por la función media $h$, con traslaciones y rotaciones específicas dadas por los efectos aleatorios $\alpha_i$, $\beta_i$ y $\gamma_i$ en el ajuste de las curvas de crecimiento individuales.

[3]El cual corresponde a un spline cúbico con restricciones de linealidad en los subintervalos extremos de la variable independiente (Hastie et al., 2009).

Siguiendo la notación de Pinheiro & Bates (2006) siendo $\boldsymbol{\phi}_i = (\alpha_i, \beta_i, \gamma_i)'$ y $\nu_{ij} = t_{ij}$, el modelo no lineal de efectos mixtos correspondiente al descrito en la ecuación 3.6 está dado por

$$x_{ij} = f(\boldsymbol{\phi_i}, \nu_{ij}) + \epsilon_{ij}, \tag{3.7}$$

$$\boldsymbol{\phi_i} = \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix} = \mathbf{I}_3 \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + \mathbf{I}_3 \times \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} = \mathbf{a} + \mathbf{b}_i, \tag{3.8}$$

para $i = 1, ..., n$ y $j = 1, ..., m_i$, con $\mathbf{I}_3$ matriz identidad de dimensión 3, $\mathbf{b}_i \sim N(0, \boldsymbol{\Psi})$ y $\epsilon_{ij} \sim N(0, \sigma^2)$. Se tiene además que

$$f(\boldsymbol{\phi_i}, \nu_{ij}) = \alpha_i + h\left(\frac{\nu_{ij} - \beta_i}{\exp(-\gamma_i)}\right). \tag{3.9}$$

Como enuncia Pinheiro & Bates (2006), los efectos fijos $\mathbf{a}$ representan el valor medio para los parámetros $\boldsymbol{\phi_i}$ en la población mientras que los efectos aleatorios $\mathbf{b}_i$ representan las desviaciones de los $\boldsymbol{\phi_i}$ respecto a sus valores medios para el $i$-ésimo paciente. Los efectos aleatorios se asumen independientes entre pacientes y los errores $\epsilon_{ij}$ se asumen independientes entre pacientes y ocasiones, también se consideran independientes de los efectos aleatorios. El método de estimación del modelo 3.7 utilizado por el autor es máxima verosimilitud, donde la verosimilitud es aproximada para el modelo no lineal mediante el algoritmo de Lindstrom y Bates (Lindstrom & Bates, 1990), el cual es descrito en la obra de Pinheiro & Bates (2006), capítulo 7 sección 2.

En el modelo planteado por Cole et al. (2010) tres efectos aleatorios determinan las trayectorias individuales y permiten una clara interpretación de las curvas en el contexto del problema: $\alpha_i$ induce una traslación vertical del $i$-ésimo sujeto respecto a la función $h$ y representa el tamaño del individuo respecto a la curva promedio, con valores positivos (negativos) para individuos más grandes (pequeños) que el individuo promedio. $\beta_i$ induce una traslación horizontal del $i$-ésimo sujeto respecto a la función $h$ y representa la madurez del individuo respecto a la curva promedio, con valores negativos para maduradores tempranos y positivos para maduradores tardíos. Finalmente, el parámetro $\gamma_i$ estandariza la duración de la etapa de crecimiento, el cual es parametrizado como $\exp(-\gamma_i)$. Valores positivos de $\gamma_i$ estiran el eje de abscisas, implicando una corta etapa de crecimiento y una mayor velocidad respecto a la curva promedio, valores negativos, por el contrario, contraen el eje, implicando una larga etapa de crecimiento y una menor velocidad respecto a la función $h$. La unidad de medida de $\alpha_i$ es las misma de $x_{ij}$ (centímetros), la de $\beta_i$ es la misma de $t_{ij}$ (años) y $\gamma_i$ es adimensional.

Respecto a los efectos fijos en la caracterización del crecimiento promedio de talla en función de la edad, se tienen en total 3 efectos fijos correspondientes al vector $\mathbf{a}$ junto a los efectos fijos asociados al spline cúbico natural, cuya cantidad depende del número de subintervalos en donde es definido. Cole et al. (2010) utilizan cuantiles en la distribución de la variable edad para determinar la ubicación de los nodos en los cuales está definido el spline, y aunque la selección de los $gl$ para definir de la cantidad de subintervalos no es discutida en el trabajo, Cole et al. (2010) usan 4 y 6 $gl$ en el modelamiento de la talla entre los 8 y los 18 años de edad[4].

[4]Por ejemplo, para 4 $gl$ el trabajo de Cole et al. (2010) considera los cuartiles $Q_1$, $Q_2$ y $Q_3$ de la variable edad para determinar los nodos internos del spline natural.

## 3.3. Datos funcionales escasos

En el análisis de datos funcionales escasos se busca estimar para cada individuo el comportamiento funcional en el dominio de interés, esta reconstrucción es alcanzada al utilizar de manera conjunta la información contenida en la muestra para la estimación de los parámetros funcionales. La estimación de la función de covarianza es de importancia en la recuperación de las trayectorias individuales al caracterizar el comportamiento conjunto de las observaciones, además, esta función permite determinar las principales fuentes de variación presentes en los datos. Diferentes propuestas han sido elaboradas para entender el comportamiento funcional en un conjunto de observaciones discretizadas escasas, en particular se destaca el trabajo de Yao et al. (2005) en la predicción de puntajes muestrales y Xiao et al. (2018) en la estimación de funciones de covarianza. Bajo estos dos acercamientos, se busca la estandarización de los individuos a partir de puntajes en el análisis de componentes principales funcional escaso para explorar el concepto de curvas de crecimiento infantil longitudinales.

Yao et al. (2005) presenta un procedimiento no paramétrico en la estimación de la función de covarianza, que a su vez permite determinar los principales modos de variación respecto a la función media mediante componentes principales. El trabajo elabora un algoritmo en la estimación de los puntajes muestrales condicionados a las observaciones individuales y a la función de covarianza estimada, los puntajes son utilizados en una posterior reconstrucción de las observaciones funcionales. El procedimiento de análisis en componentes principales con esperanza condicionada es denominado PACE por sus siglas en inglés (*Principal components analysis through conditional expectation*). Xiao et al. (2018) propone un método específico con enfoque matricial para la estimación eficiente de funciones de covarianza, este presenta una extensión al caso funcional escaso para la metodología de estimación de la función de covarianza implementada en Xiao et al. (2016) bajo un escenario denso. La propuesta de estimación para datos funcionales densos es denominada *Fast Covariance Function Estimation*, o FACE por sus siglas en inglés. En el presente trabajo se denota como FACE-S la alternativa funcional escasa. A continuación se presenta una breve descripción de los métodos mencionados.

### 3.3.1. Análisis en componentes principales con esperanza condicionada (PACE)

El procedimiento no paramétrico de Rice & Wu (2001) es direccionado al análisis de componentes principales en el trabajo de Yao et al. (2005). Esta técnica provee un marco para el modelamiento de datos escasos en el que el dominio es observado de manera relativamente densa al considerar conjuntamente las observaciones en la muestra. Para $x_{ij}$, la $j$-ésima observación del $i$-ésimo individuo en el instante $t_{ij}$, se asume el modelo descrito en la ecuación 2.18. Además, por la expansión de Karhunen-Loève presentada en la ecuación 2.13, el componente funcional en el modelo es planteado como una combinación lineal de funciones propias. El modelo considerado es entonces igual a

$$x_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \langle X_i - \mu, v_k \rangle v_k(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik} v_k(t_{ij}) + \epsilon_{ij}, \;\; t_{ij} \in T. \quad (3.10)$$

#### 3.3.1.1. Estimación de la función media

La estimación de la función media $\mu$ para el modelo propuesto es realizada mediante regresión local lineal univariada, en la cual se utilizan de manera conjunta los datos escasos de todos los individuos en la muestra. Para el instante $t$ se estima $\mu$ como $\hat{\beta}_0$, el cual se encuentra al minimizar la expresión

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} K_\mu \left( \frac{t_{ij} - t}{h_\mu} \right) (x_{ij} - (\beta_0 + \beta_1(t - t_{ij})))^2,$$

donde $K_\mu$ corresponde a una función kernel univariada utilizada para el suavizamiento de la función media y $h_\mu$ representa el ancho de banda para la estimación de $\mu$. El procedimiento permite que la función media estimada no requiera que la muestra sea observada de manera regular o igualmente espaciada en el dominio de estudio, en contraste a la ecuación 2.8 para el escenario funcional denso. Por otra parte, de manera similar al caso funcional denso, en el escenario escaso la función media estimada es requerida para estimar la función de covarianza.

#### 3.3.1.2. Estimación de la función de covarianza

En la estimación de la función de covarianza se destaca que el modelo propuesto en la ecuación 2.18 comprende un error de medición no despreciable reflejado de manera aditiva en la covarianza de las observaciones discretizadas. Como se observa en la ecuación 2.21, los dos componentes que contribuyen a la función de covarianza de las observaciones son $c$, la función de covarianza de $X$, y $\sigma^2$, la varianza del error de medición. El proceso de estimación de los componentes es realizado una vez estimada la función media, con la cual se calcula el producto de desvíos para cada individuo respecto a esta. Para $i = 1, ..., n$ y $1 \leq j \leq l \leq m_i$ se define $G_{ijl} = G_i(t_{ij}, t_{il})$ como $(x_{ij} - \hat{\mu}(t_{ij}))(x_{il} - \hat{\mu}(t_{il}))$. Estos desvíos son usados como insumo en la estimación de la función de covarianza, la cual es obtenida mediante una versión bivariada del suavizamiento de la función media junto a un algoritmo para la estimación de la varianza en el error de medición.

Para la estimación de la función de covarianza de $X$ se realiza inicialmente una regresión local lineal bivariada, la cual es una generalización a dos variables del procedimiento descrito para la estimación de la función media. En el suavizamiento se excluyen aquellos desvíos $G_{ijl}$ en los que $j = l$ los cuales corresponden a las varianzas de las observaciones discretizadas, que al incorporar un error de medición no despreciable en el caso escaso, presentan un aumento de $\sigma^2$ en su varianza. Adicionalmente, como destacan los autores, dado que la covarianza de $X$ es máxima a lo largo de la diagonal, es necesario hacer un ajuste a la diagonal de la superficie de covarianza recién estimada (Yao et al., 2003). El ajuste se realiza mediante una regresión local bivariada, lineal en la dirección paralela a la diagonal y cuadrática en la dirección perpendicular a esta, excluyendo nuevamente los desvíos en la diagonal. Esto se logra al realizar el suavizamiento sobre los ejes rotados 45 grados. La función diagonal de la superficie resultante en $t$ se nota como $\tilde{c}(t)$. Para la varianza del error de medición, $\sigma^2$, un suavizamiento mediante regresión local lineal univariada es ejecutado, esta vez sobre los desvíos $G_{ijj}$. Esto permite encontrar una estimación de la función de varianza, que en $t$ es igual a $V(t) = c(t, t) + \sigma^2$. Finalmente se obtiene una estimación de la varianza del error a partir de la diferencia entre las funciones $\hat{V}$ y $\tilde{c}$.

El proceso de estimación descrito es aplicable a datos longitudinales irregularmente espaciados en los que el número de observaciones por individuo es pequeño. Se destaca la diferencia con el enfoque tradicional funcional denso, en el cual la estimación de la función de covarianza depende de observar de manera regular y en todo el dominio de interés las curvas funcionales en la muestra, como se destaca en la ecuación 2.9. En su trabajo, Yao et al. (2005) usan una herramienta gráfica denominada *design plot* la cual consiste en un plano cartesiano junto a las parejas $(t_{ij}, t_{ik})$ cuyo objetivo es determinar si la información muestral cubre de manera densa el plano definido por el dominio $T$. Bajo PACE, el procedimiento de estimación de la función de covarianza es apropiado cuando las parejas de puntos son lo suficientemente densas en el plano. Un ejemplo de aplicación para secuenciamiento celular presentado por Madrigal et al. (2018) muestra el uso de esta herramienta, en el cual los autores deciden truncar el dominio para evitar un efecto erróneo de frontera. Una vez estimada la función de covarianza del proceso aleatorio, esta es utilizada para obtener versiones discretas de las ecuaciones propias con el objetivo de estimar la función $v_k$ y el escalar $\lambda_k$ para los primeros $q$ componentes de la expansión de Karhunen-Loève. Diferentes métodos computacionales para la estimación de los componentes principales funcionales mediante una aproximación matricial equivalente pueden encontrarse en el capítulo 8.4 de la obra de Ramsay & Silverman (2005).

Al ser obtenida la función media estimada junto a las funciones y valores propios del operador de covarianza estimado, se procede a predecir para cada individuo los primeros $q$ puntajes sobre las funciones propias estimadas. Este procedimiento es considerablemente diferente en el escenario escaso (Kokoszka & Reimherr, 2017) y no es siempre requerido para la reconstrucción de las curvas funcionales (ver por ejemplo Xiao et al. (2018)).

#### 3.3.1.3. Predicción de curvas y puntajes muestrales

Para encontrar los puntajes muestrales $\hat{\xi}_{ik}$ la propuesta de Yao et al. (2005) está basada en supuestos de normalidad conjunta para $\xi_{ik}$ y $\epsilon_{ij}$ bajo el modelo propuesto en la ecuación 3.10. Para el $i$-ésimo sujeto observado en los instantes $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})'$ se definen los siguientes vectores $\mathbf{x}_i = (x_{i1}, ..., x_{im_i})'$ como el vector de observaciones, $\mathbf{X}_i = (X_i(t_{i1}), ..., X_i(t_{im_i}))'$ como el vector de realizaciones de la función aleatoria $X_i$ en $\mathbf{t}_i$, $\boldsymbol{\mu}_i = (\mu(t_{i1}), ..., \mu(t_{im_i}))'$ como la función media evaluada en $\mathbf{t}_i$ y $\mathbf{v}_{ik} = (v_k(t_{i1}), ..., v_k(t_{im_i}))'$ como la $k$-ésima función propia evaluada en $\mathbf{t}_i$. El objetivo es determinar la distribución condicional de los puntajes dadas las observaciones $\mathbf{x}_i$ con lo cual encontrar un predictor $\hat{\xi}_{ik}$ para los puntajes individuales. Se resalta que $E(\xi_{ik}, \mathbf{x}_i) = (0, \boldsymbol{\mu}_i)$ y además, al notar que $\text{Cov}(\mathbf{x}_i, \mathbf{x}_i) = \boldsymbol{\Sigma}_{\mathbf{x}_i} = \text{Cov}(\mathbf{X}_i, \mathbf{X}_i) + \sigma^2 \mathbf{I}_{m_i}$, se tiene que el elemento $(j, l)$ de la matriz $\boldsymbol{\Sigma}_{\mathbf{x}_i}$ está dado por la ecuación 2.21. Adicionalmente, dado que $V(\xi_{ik}) = \lambda_k$ y $\text{Cov}(\xi_{ik}, x_{ij}) = \lambda_k \Gamma(v_k)(t_{ij}) = \lambda_k v_k(t_{ij})$ se obtiene que la covarianza entre el puntaje $\xi_{ik}$ y el vector de observaciones $\boldsymbol{x}_i$ es igual a la siguiente matriz definida por bloques

$$\text{Cov}(\xi_{ik}, \mathbf{x}_i) = \left(\begin{array}{c|ccc} \lambda_k & \lambda_k v_k(t_{i_1}) & \dots & \lambda_k v_k(t_{im_i}) \\ \hline \lambda_k v_k(t_{i_1}) & c(t_{i1}, t_{i1}) + \sigma^2 & \dots & c(t_{i1}, t_{im_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_k v_k(t_{im_i}) & c(t_{im_i}, t_{i1}) & \dots & c(t_{im_i}, t_{im_i}) + \sigma^2 \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{\mathbf{x}_i} \end{array}\right)$$

Bajo los supuestos de normalidad mencionados previamente, se puede caracterizar la distribución condicional $\xi_{ik}|\mathbf{x}_i$ como una distribución normal con media igual a

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i) = \lambda_k \mathbf{v}_{ik}'\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i).$$

El valor esperado de esta distribución corresponde a la mejor predicción del $k$-ésimo puntaje para el $i$-ésimo individuo (Yao et al., 2005; Kokoszka & Reimherr, 2017). Una vez encontradas las estimaciones de los parámetros correspondientes, se obtienen los puntajes muestrales como

$$\hat{\xi}_{ik} = E(\xi_{ik}|\mathbf{x}_i) = \hat{\lambda}_k \hat{\mathbf{v}}_{ik}'\hat{\boldsymbol{\Sigma}}_{\mathbf{x}_i}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_i) \tag{3.11}$$

Considerando los primeros $q$ componentes principales funcionales, de manera similar a la propuesta de Rice & Wu (2001) se obtiene una estimación de las trayectorias individuales

$$\hat{X}_i^q(t) = \hat{\mu}(t) + \sum_{k=1}^{q} \hat{\xi}_{ik}\hat{v}_k(i) \tag{3.12}$$

Nuevamente se destaca que la reconstrucción de las trayectorias combina información individual y de la muestra en conjunto.

### 3.3.2. Estimación eficiente de la función de covarianza para datos escasos (FACE-S)

Bajo el mismo modelo de PACE con error no despreciable presentado en la ecuación 2.18 y mediante una descomposición en base de funciones para la función de covarianza semejante a la usada en el modelo longitudinal descrito en la subsección 3.2.2, la metodología propuesta por Xiao et al. (2018) busca estimar de manera eficiente y especializada la función de covarianza, dada su gran importancia en el análisis de datos funcionales escaso.

Buscando solucionar diferentes problemas metodológicos en procesos de estimación de funciones de covarianza propuestos previamente, FACE-S resulta en un modelamiento simultáneo de la función de covarianza y de la varianza del error de medición. Xiao et al. (2018) resaltan en particular dos inconvenientes en el algoritmo de estimación elaborado por Yao et al. (2005): el primero está relacionado con el procedimiento estadístico para la estimación de los parámetros funcionales, el cual involucra técnicas de suavizamiento de multiple propósito que no están enfocadas en la estimación específica de funciones de covarianza. El segundo problema está relacionado con el algoritmo de búsqueda del ancho de banda para el suavizamiento en la estimación de las funciones, que al ser realizado mediante validación cruzada dejando un individuo fuera, es computacionalmente costoso en PACE. FACE-S propone un acercamiento eficiente en la estimación de la función de covarianza al simplificar el cálculo del error de validación cruzada dejando un individuo afuera. Respecto a la estimación de la función media bajo FACE-S, detallada en el material suplementario del artículo, esta es efectuada mediante suavizamiento spline como el destacado en la subsección 2.1.1.2 y utilizando validación cruzada para determinar el parámetro de suavizamiento.

Al igual que PACE, FACE-S utiliza los desvíos respecto a la función media estimada $G_{ijl}$ como insumo en la estimación de la función de covarianza, pero en lugar de realizar un suavizamiento bivariado, FACE-S parametriza la función de covarianza y vectoriza sus componentes, cambiando el problema de estimación a mínimos cuadrados. Dada la representación en funciones base observada en la ecuación 3.3, la función de covarianza de $X$ es modelada como

$$H(t_{ij}, t_{il}) = \sum_{k=1}^{q} \sum_{s=1}^{q} \Gamma_{ks} B_k(t_{ij}) B_s(t_{il}) \tag{3.13}$$

Bajo el modelo propuesto en la ecuación 3.13 se define la matriz $\mathbf{\Gamma} = (\Gamma_{ks})_{1 \leq k \leq q, 1 \leq l \leq q}$ como una matriz de coeficientes simétrica, garantizando con esto la simetría de la función de covarianza estimada $\hat{c}$ (Xiao et al., 2018). Adicionalmente, para el $i$-ésimo sujeto con $i = 1, ..., n$ se define $n_i = m_i(m_i + 1)/2$ y dada su observación $j$-ésima con $1 \leq j \leq m_i$ se tienen los siguientes vectores de dimensión $m_i - j + 1$: $\mathbf{G}_{ij} = (G_{ijj}, ..., G_{ijm_i})'$, $\mathbf{H}_{ij} = (H(t_{ij}, t_{ij}), ..., H(t_{ij}, t_{im_i}))'$ y $\boldsymbol{\delta}_{ij} = (1, \mathbf{0}'_{m_i - j})'$ con $\mathbf{0}'_p$ vector de ceros de dimensión $p$. Los $m_i$ componentes son a su vez configurados en forma de vectores de dimensión $n_i$ como $\mathbf{G}_i = (\mathbf{G}'_{i1}, ..., \mathbf{G}'_{im_i})'$, $\mathbf{H}_i = (\mathbf{H}'_{i1}, ..., \mathbf{H}'_{im_i})'$ y $\boldsymbol{\delta}_i = (\boldsymbol{\delta}'_{i1}, ..., \boldsymbol{\delta}'_{im_i})'$ con $i = 1, ..., n$.

Para la estimación bajo mínimos cuadrados ponderados, la estructura de dependencia presente en las observaciones longitudinales es capturada mediante una matriz de pesos $\mathbf{W}_i$ de dimensión $n_i \times n_i$, cuya inversa está definida en función de la matriz de covarianza de $\mathbf{G}_i$. Con los objetos definidos, la suma de cuadrados ponderados es igual a

$$SCP(\mathbf{\Gamma}, \sigma) = \sum_{i=1}^{n} \left(\mathbf{H}_i + \boldsymbol{\delta}_i \sigma^2 - \mathbf{G}_i\right)' \mathbf{W}_i \left(\mathbf{H}_i + \boldsymbol{\delta}_i \sigma^2 - \mathbf{G}_i\right) \tag{3.14}$$

Para la estimación de los parámetros $\mathbf{\Gamma}$ y $\sigma$ que determinan la función de covarianza en la ecuación 3.14, el problema de minimización es regularizado al penalizar la rugosidad de la función de covarianza para evitar un sobreajuste en la estimación. A diferencia de PACE, el procedimiento propuesto logra obtener una estimación de la función de covarianza sin requerir la elección de un ancho de banda o de un kernel, sin embargo, si requiere la selección de un parámetro de suavizamiento en la penalización. Para esto, Xiao et al. (2018) propone un algoritmo eficiente para aproximar el error de validación cruzada dejando un individuo afuera. El algoritmo descrito para la estimación eficiente de la función de covarianza junto al procedimiento de predicción de curvas presentado a continuación han sido utilizados en métodos de regresión funcional para el estudio del crecimiento infantil usando como covariables variables funcionales escasas (Leroux et al., 2018).

#### 3.3.2.1. Predicción de curvas

De manera similar al procedimiento de predicción de curvas de PACE, la predicción en FACE-S está soportada en supuestos de normalidad para realizar la reconstrucción de las trayectorias individuales. Sin embargo, en lugar de aplicar los supuestos en la predicción de puntajes sobre funciones propias muestrales, los utiliza para encontrar estimaciones de las realizaciones discretas $(X_i(l_{i1}), ..., X_i(l_{ik}))'$ en $l_{i1}, ..., l_{ik}$ para la función $X_i$ usando el vector de observaciones $\mathbf{x}_i$ y los parámetros funcionales estimados. El procedimiento busca la distribución condicional $(X_i(l_{i1}), ..., X_i(l_{ik}))|\mathbf{x}_i$ bajo supuestos de normalidad conjunta.

Una vez es estimada la función media y los componentes de la función de covarianza, el vector de predicciones $(\hat{X}_i(l_{i1}), ..., \hat{X}_i(l_{ik}))$ se toma como el valor esperado de la distribución de las nuevas observaciones condicionado al vector de observaciones correspondiente. En $t$ se evalúa la base de funciones $\mathbf{B}(t) = (B_1(t), ..., B_q(t))'$, con lo cual se definen las matrices $\mathbf{H}_i^0 = [\mathbf{B}(t_{i1}), ..., \mathbf{B}(t_{im_i})]'$ y $\mathbf{H}_i^1 = [\mathbf{B}(l_{i1}), ..., \mathbf{B}(l_{ik})]'$. Se definen además los vectores $\hat{\boldsymbol{\mu}}_{i0}$ y $\hat{\boldsymbol{\mu}}_{i1}$ como la función media estimada evaluada en $(t_{i1}, ..., t_{im_i})$ y $(l_{i1}, ..., l_{ik})$ respectivamente. Las predicciones de $X_i$ en los instantes $l_{i1}, ..., l_{ik}$ bajo FACE-S están dadas por

$$(\hat{X}_i(l_{i1}), ..., \hat{X}_i(l_{ik}))' = \hat{\boldsymbol{\mu}}_{i1} + \left(\mathbf{H}_i^1 \hat{\boldsymbol{\Gamma}} (\mathbf{H}_i^0)'\right) \hat{\mathbf{V}}_i^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{i0}), \tag{3.15}$$

con $\hat{\mathbf{V}}_i = \left(\mathbf{H}_i^0 \hat{\boldsymbol{\Gamma}} (\mathbf{H}_i^0)' + \hat{\sigma}^2 \mathbf{I}_{m_i}\right)$. Xiao et al. (2018) destaca que en contraste a PACE, FACE-S no requiere la estimación de funciones propias ni tampoco de puntajes individuales en la predicción de $X_i$, sin embargo, la expansión de Karhunen-Loève puede ser utilizada sobre la función de covarianza estimada de manera análoga a Yao et al. (2005). Esto permite obtener bajo la ecuación 3.10 una reconstrucción funcional de las curvas usando los primeros $q$ componentes principales mediante la función de covarianza estimada con el algoritmo FACE-S.

## 3.4. Cuantiles multivariados

La metodología de componentes principales funcionales para el análisis de datos escasos de crecimiento infantil fue propuesta por Zhang et al. (2015). Mediante un acercamiento no paramétrico, este trabajo propone un algoritmo para transformar datos observados de manera escasa en elementos funcionales mediante componentes principales incorporando un error de medición no despreciable como en Yao et al. (2005) y Xiao et al. (2018). De particular interés en el trabajo es la elaboración de curvas de crecimiento infantil para información de tipo longitudinal, que aplicando un modelamiento funcional escaso y regresión cuantílica, busca diferentes percentiles de interés en la distribución conjunta de los primeros $k$ puntajes muestrales de los componentes principales funcionales estimados correspondientes. La reconstrucción funcional de los datos escasos mediante componentes principales junto al ordenamiento multivariado de los puntajes correspondientes permite una caracterización del patrón de crecimiento de manera análoga al escenario transversal.

Zhang et al. (2015) destacan que la distribución conjunta de los puntajes sobre las primeras $k$ funciones propias en el análisis de componentes principales funcionales rara vez sigue una distribución paramétrica conocida, por lo que sugieren determinar los cuantiles de los puntajes estimados de manera no paramétrica basado en la propuesta de Wei (2008). El acercamiento de Wei (2008) busca la construcción de percentiles multivariados para vectores $k$ dimensionales mediante la unión de intervalos unidimensionales, los cuales corresponden a los valores centrales de la distribución del proceso aleatorio estandarizado, proyectado sobre las líneas diametrales de la esfera unitaria. La unión de los intervalos resultantes es un conjunto central con una medida de probabilidad determinada. Usando, por ejemplo, los puntajes muestrales sobre los primeros dos componentes principales en el análisis de datos funcionales de crecimiento infantil, se obtiene una colección de contornos concéntricos bidimensionales con los cuales se construyen cuantiles de crecimiento de la medida antropométrica de interés. Los contornos encontrados mediante este método son interpretados de manera semejante a los percentiles obtenidos en los modelos transversales bajo una concepción longitudinal.

Siguiendo la notación de Wei (2008) y la descripción del algoritmo propuesto, se define $\mathbb{S}^{k-1}$ como la esfera unitaria en $\mathbb{R}^k$ y la dirección espacial como la línea conectando un par de puntos opuestos $(\mathbf{u}, -\mathbf{u})$ en $\mathbb{S}^{k-1}$. Para cualquier par de puntos $\mathbf{x}$ y $\mathbf{y}$ en $\mathbb{R}^k$ se escribe $\mathbf{x} \overset{(u)}{=} \mathbf{y}$ si tanto $\mathbf{x}$ como $\mathbf{y}$ se encuentran en la misma dirección espacial definida por $(\mathbf{u}, -\mathbf{u})$. Finalmente, se dice que $\mathbf{x}$ es más atípico que $\mathbf{y}$ hacia $\mathbf{u}$ si $\langle \mathbf{x}, \mathbf{u} \rangle \geq \langle \mathbf{y}, \mathbf{u} \rangle$ y se nota $\mathbf{x} \geq_{(u)} \mathbf{y}$. Para un vector aleatorio $\mathbf{Y}$ de dimensión $k$ se define un vector central $\boldsymbol{\mu}$ y una matriz diagonal $\mathbf{S} = \text{diag}(s_i)_{i=1,...,k}$ con $s_i$ parámetro de escala. Se tiene $\mathbf{Z} = \mathbf{S}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$ como el vector estandarizado. Para cualquier $\tau \in (0, 1)$ se define el $100\tau\,\%$ intervalo para $\mathbf{Z}$ a lo largo de la dirección espacial determinada por $(\mathbf{u}, -\mathbf{u})$ como el intervalo cerrado definido por los cuantiles condicionales $l_\tau(\mathbf{u})$ y $u_\tau(\mathbf{u})$ con $l_\tau(\mathbf{u}) \leq u_\tau(\mathbf{u})$ que contiene todos los puntos $\mathbf{x}$ con $\mathbf{x} \overset{(u)}{=} \mathbf{u}$ tales que

$$P(\mathbf{Z} \leq_{(u)} \mathbf{x} | \mathbf{Z} \overset{(u)}{=} \mathbf{u}) \leq \frac{1+\tau}{2} \tag{3.16}$$

Y además

$$P(\mathbf{Z} \leq_{(u)} \mathbf{x} | \mathbf{Z} \overset{(u)}{=} \mathbf{u}) \geq \frac{1-\tau}{2} \tag{3.17}$$

La unión de los conjuntos en las direcciones espaciales de $\mathbb{S}^{k-1}$ está dada por

$$\mathcal{R}_\tau = \left\{ \mathbf{S}\mathbf{x} + \boldsymbol{\mu} : \mathbf{x} \in \bigcup_{\mathbf{u} \in \mathbb{S}^{k-1}} [l_\tau(\mathbf{u}), u_\tau(\mathbf{u})] \right\} \tag{3.18}$$

Con lo que se define el $100\tau$-ésimo cuantil multivariado para $\mathbf{Y}$ como los límites del conjunto $\mathcal{R}_\tau$ y se nota $\mathcal{C}_\tau$. Como resalta Wei (2008), si $k = 1$ se tiene que $\mathcal{R}_\tau$ es el intervalo $[F^{-1}(\frac{1-\tau}{2}), F^{-1}(\frac{1+\tau}{2})]$ y $\mathcal{C}_\tau$ es igual a los puntos límites en dicho intervalo.

En la estimación del $100\tau\,\%$ cuantil multivariado $\mathcal{C}_\tau$ para $\mathbf{Y}$, Wei (2008) propone transformar los puntos muestrales en un sistema de coordenadas polares. En lugar de caracterizar las observaciones mediante coordenadas cartesianas, estas son representados mediante coordenadas polares, con las cuales un punto en el plano es identificado mediante un radio respecto al origen y uno o varios ángulos, dependiendo la dimensionalidad del problema. Asumiendo $k = 2$ se obtiene una muestra de $n$ vectores $\mathbf{y}_i = \{(y_{i1}, y_{i2})'\}_{i=1,..,n}$ la cual sigue la distribución conjunta $\mathbf{F}$ de $\mathbf{Y}$. En un primer paso del algoritmo la observación $\mathbf{y}_i$ es estandarizada mediante la transformación $\mathbf{z}_i = \hat{\mathbf{S}}(y_i - \hat{\mu})$, seguidamente $\mathbf{z}_i = (z_{i1}, z_{i2})$ es expresado en coordenadas polares equivalentes $(r_i^{(1)}, \theta_i^{(1)})$ y $(r_i^{(2)}, \theta_i^{(2)})$ con

$$\begin{aligned} \theta_i^{(1)} &= \arctan\left(\frac{z_{i2}}{z_{i1}}\right) \cap [0, \pi) \qquad r_i^{(1)} = z_{i1}\cos(\theta_i^{(1)}) + z_{i2}\text{sen}(\theta_i^{(1)}), \\ \theta_i^{(2)} &= \arctan\left(\frac{z_{i2}}{z_{i1}}\right) \cap [\pi, 2\pi) \qquad r_i^{(2)} = z_{i1}\cos(\theta_i^{(2)}) + z_{i2}\text{sen}(\theta_i^{(2)}), \end{aligned}$$

para $i = 1, ..., n$, en donde $\theta_i^{(1)}$ y $\theta_i^{(2)}$ difieren en $\pi$ y $r_i^{(1)} = -r_i^{(2)}$.

Posteriormente se nota que el $100\tau$-ésimo cuantil puede expresarse como una función $g$ del ángulo $\theta$ dada por $g_\tau(\theta) : [0, 2\pi] \to \mathbb{R}$ bajo la restricción $g_\tau(0) = g_\tau(2\pi)$, con lo cual $g_\tau(\theta)$ es estimada al minimizar

$$\sum_{k=1}^{2} \sum_{i=1}^{n} \rho_{(\tau+1)/2}(r_i^{(k)} - g(\theta_i^{(k)})) \tag{3.19}$$

Con $\rho_\tau(x) = x(\tau - \mathbf{1}_{\{x<0\}})$ y sobre la familia de funciones

$$\{g(\theta) : [0, 2\pi] \to \mathbb{R} : g(0) = g(2\pi)\} \tag{3.20}$$

Wei (2008) aproxima $g$ mediante una base de funciones B-spline con restricciones. En un último paso el contorno resultante es transformado de vuelta a las unidades originales. Para el estudio del crecimiento infantil se toman como referencia los valores usuales de $\tau$ en las curvas de crecimiento: 0.50, 0.75 y 0.90, realizando así un total de 3 regresiones no paramétricas en la obtención de los cuantíles multivariados $\mathcal{C}_{0.50}$, $\mathcal{C}_{0.75}$ y $\mathcal{C}_{0.90}$.

# CAPÍTULO 4

# Construcción de curvas de crecimiento

Este capítulo desarrolla los resultados principales del trabajo: la estimación y contraste de curvas estándar de crecimiento junto a la construcción de curvas estándar de crecimiento ajustando por madurez, ambos mediante una muestra no probabilística de niños saludables de la ciudad de Bogotá. La estimación de las curvas de crecimiento es realizada siguiendo los lineamientos internacionales de la OMS para posteriormente confrontar los resultados obtenidos con curvas de crecimiento nacionales e internacionales preestablecidas. En su elaboración son utilizados los registros de única ocasión del conjunto de datos, los cuales corresponden a los pacientes que se presentan una única vez al médico especialista. La construcción de las curvas de crecimiento ajustando por madurez es realizada con aquellos individuos que tienen más de una visita registrada, las curvas son estimadas utilizando regresión cuantílica sobre los puntajes muestrales correspondientes a las primeras funciones propias estimadas. Se propone diferenciar los registros longitudinales por la madurez de los individuos de manera previa a la construcción de los cuantiles multivariados de crecimiento.

A continuación se describe el contenido de las diferentes secciones del capítulo: en la sección 4.1 se detalla la obtención y preprocesamiento de la información obtenida para la elaboración de las curvas de crecimiento. En la sección 4.2 se describe el conjunto de datos y se destacan las características principales en la relación entre la variable edad decimal y la variable talla como variable respuesta. Posteriormente, la sección 4.3 implementa el proceso transversal de estimación de las curvas de crecimiento usando los datos de única ocasión para la comparación con los estándares nacionales e internacionales. La sección 4.4 presenta los resultados del modelamiento longitudinal usando los datos con más de una visita registrada. Se selecciona el modelo SITAR con el objetivo de diferenciar la madurez de los individuos para posteriormente estimar funciones de covarianza mediante el procedimiento FACE-S. Usando las funciones de covarianza estimadas, en la sección 4.5 se resumen los registros de crecimiento mediante los puntajes muestrales obtenidos a través de análisis en componentes principales funcionales. Finalmente son estimados las curvas de crecimiento infantil diferenciando por madurez utilizando la propuesta de Wei (2008) sobre los puntajes muestrales. Las figuras y el análisis de la información en el presente documento son realizadas mediante el uso del software R versión 3.6.2 (R Core Team, 2019). Las gráficas son elaboradas mediante el uso del paquete `ggplot` (Wickham, 2016).

## 4.1. Obtención y depuración del conjunto de datos

Dada la dificultad en la obtención de la información antropométrica, es usual que la información para la construcción de las curvas provenga de ciudades específicas de un país determinado (como por ejemplo las curvas de crecimiento colombianas, que consideran las principales ciudades del país) por lo cual no se selecciona una muestra representativa nacional en la construcción de las curvas de crecimiento. En particular, para la estimación de curvas de crecimiento estándar como las curvas internacionales de la OMS, no se cuenta con información representativa para Latinomerica, ni tampoco del país seleccionado de la región para la estimación de las curvas. Para el presente trabajo, la información obtenida corresponde a una muestra no probabilística de la ciudad de Bogotá, ya que la información recolectada proviene de visitas médicas en las que los padres voluntariamente seleccionan un endocrinólogo pediatra para la evaluación y el seguimiento del crecimiento y desarrollo de sus hijos. El seguimiento del crecimiento infantil mediante las visitas médicas múltiples no obedece a una condición médica particular de los pacientes, esta información proviene de visitas voluntarias al médico especialista de una población saludable y socioeconómicamente privilegiada, que bajo los supuestos de la OMS, presentan un comportamiento homogéneo del crecimiento.

### 4.1.1. Obtención del conjunto de datos

En el conjunto de datos está conformado por registros médicos pediátricos obtenidos en consultas médicas múltiples o de única ocasión a niños y adolescentes saludables entre los 0 y los 18 años, residentes de la ciudad de Bogotá y provenientes de hogares socioeconómicamente privilegiados que pueden costear el pago de consultas privadas de pediatría. Las visitas médicas sucedieron en su mayoría entre los años 2010 y 2016, siendo así una muestra contemporánea a las curvas de crecimiento infantil de la OMS y las curvas de crecimiento colombianas. A continuación se listan los criterios de inclusión adicionales para los pacientes del conjunto de datos: nacido en parto simple, control adecuado del embarazo de la madre y del nacimiento del menor, buenas condiciones de salud (no sufrir de enfermedades clínicas), nutrición adecuada al menor, no recibir tratamientos médicos que puedan afectar el crecimiento del paciente. Todos los registros médicos presentan la fecha de nacimiento del paciente, la longitud (antes de los dos años) o talla (después de los dos años) al momento del examen medida usando un estadiómetro y la fecha del examen, mientras que para algunos de los pacientes se cuenta con covariables adicionales tales como talla materna. Las fechas dispuestas en la base de datos son transformadas a su versión decimal para una obtención de la edad decimal del paciente en la fecha del examen.

### 4.1.2. Preprocesamiento del conjunto de datos

El conjunto de datos disponible es preprocesado buscando tanto una identificación adecuada de los individuos como una consistencia en las variables de interés. Las diferentes validaciones permiten la corrección de la información dispuesta en la base de datos al ser contrastadas con las historias médicas originales de los pacientes y con un endocrinólogo pediatra en caso de ser necesario. Inicialmente, errores tipográficos en la identificación de los individuos son corregidos, este procedimiento involucró tanto una identificación manual de problemas en nombres y apellidos de los pacientes como una búsqueda de potenciales errores a partir de métricas entre secuencias de caracteres. Posteriormente, diferentes reglas

son definidas para verificar la consistencia de la información relevante para el estudio. A partir de la información disponible de talla, peso, fechas de examen y de nacimiento, estas reglas buscan identificar problemas de registros médicos repetidos, variables de interés no recolectadas en las consultas, inconsistencias en talla de crecimiento y atipicidades tanto en el índice de masa corporal (IMC) como en la velocidad lineal de crecimiento.

Para evitar la influencia de individuos con talla y peso inusual en la estimación de las curvas, las visitas de pacientes que presentan un IMC mayor a tres desviaciones estándar o menor a tres desviaciones estándar fueron retiradas utilizando como referencia las curvas desarrolladas por Durán et al. (2016). Además, velocidades lineales de crecimiento fueron calculadas para determinar crecimientos consecutivos inusuales de talla en el componente longitudinal: visitas con velocidades de crecimiento extremas por inspección visual son corregidas o eliminadas de la base de datos. Finalmente, observaciones longitudinales que presentan un decrecimiento en talla mayor al error de medición establecido (tres milímetros) son retiradas. Bajo los anteriores criterios de exclusión descritos, un pequeño número de observaciones fueron retiradas para el análisis: por IMC inusual 25 observaciones fueron eliminadas, para la velocidad lineal extrema 16 observaciones, y para decrecimiento en talla en total 29 visitas al médico especialista fueron removidas. Finalmente, las observaciones de 17 sujetos, 5 con registros longitudinales y 12 transversales, fueron excluidos del análisis al tener una talla inusual para su edad.

## 4.2. Descripción del conjunto de datos

Una vez preprocesado, el conjunto de datos cuenta con 7909 registros en la base de datos correspondientes a un total de 2000 niños. En la tabla 4.1 se presenta la distribución de frecuencias absolutas para la variable edad decimal al momento del examen la cual es discretizada en intervalos significativos para el crecimiento infantil, en donde la edad decimal corresponde a la diferencia entre la fecha decimal del examen y la fecha decimal de nacimiento del paciente. Se observa una distribución asimétrica con una alta frecuencia en niños menores de dos años y entre los 10 y 15 años, lo cual probablemente se debe a la preocupación de los padres en el crecimiento de sus hijos justo después de su nacimiento y durante la preadolescencia y adolescencia. Como es de esperarse, se cuenta con una menor cantidad de visitas al médico pediatra para las edades mayores de los pacientes.

| Edad | [0,1) | [1,2) | [2,5) | [5,10) | [10,15) | [15,18] |
|---|---|---|---|---|---|---|
| Frecuencia | 881 | 409 | 912 | 1630 | 3382 | 695 |

Tabla 4.1. Distribución de frecuencias absolutas. Edad decimal (discreta).

Por otra parte, en la tabla 4.2 se observa la distribución del número de observaciones por individuo. Se destaca una distribución sesgada a la derecha para esta variable, en particular, se nota una cantidad importante de pacientes con una única observación en el tiempo. Esto indica que es bastante común la visita única al médico especialista para determinar el estado de salud de los individuos.

| Número de observaciones | 1 | [2,5) | [5,10) | [10,15) | [15,20] | Más de 20 |
|---|---|---|---|---|---|---|
| Frecuencia | 921 | 526 | 338 | 118 | 58 | 39 |

Tabla 4.2. Distribución de frecuencias absolutas. Número de observaciones por individuo.

Las 881 observaciones en el intervalo de cero a un año corresponden a 158 individuos, de los cuales tan sólo una cuarta parte tienen observaciones únicas. Los 118 pacientes restantes en mediana tienen 8 observaciones y en promedio 7.1, similar a las 8.9 observaciones en promedio en las curvas colombianas para la misma edad (Durán et al., 2016). Para las 1290 observaciones en el intervalo de cero a dos años, se tienen en mediana 9 observaciones y en promedio 8.2 al retirar las 83 observaciones únicas. Después de los dos años se tienen 6619 observaciones correspondientes a 1913 individuos, de los cuales un poco menos de la mitad tienen observaciones únicas. En contraste con el intervalo de cero a dos años, después de los dos años se tienen en mediana 4 observaciones y 5.6 en promedio por individuo al retirar observaciones únicas. Se evidencia así un seguimiento más frecuente para los primeros años de vida, en los cuales se presenta un mayor cambio en el crecimiento infantil.

Los diagramas de dispersión de la figura 4.1 presentan la longitud/talla de los pacientes en sus edades decimales correspondientes diferenciando por la naturaleza de los datos (consultas de única ocasión o consultas consecutivas). En la figura se muestra que para las diferentes edades decimales consideradas la relación entre la variable respuesta y la edad decimal no es lineal, además, no existe una varianza homogénea a través de los valores de la edad decimal en la talla infantil. Se destaca una similitud en las observaciones para ambos escenarios: tanto las consultas múltiples como las de única ocasión presentan un marcado incremento en longitud/talla para los primeros años de vida que posteriormente disminuye. Los registros longitudinales destacan un nuevo incremento cerca a la pubertad que no es tan evidente para las observaciones únicas, sin embargo, en general ambos gráficos muestran características similares en los datos.

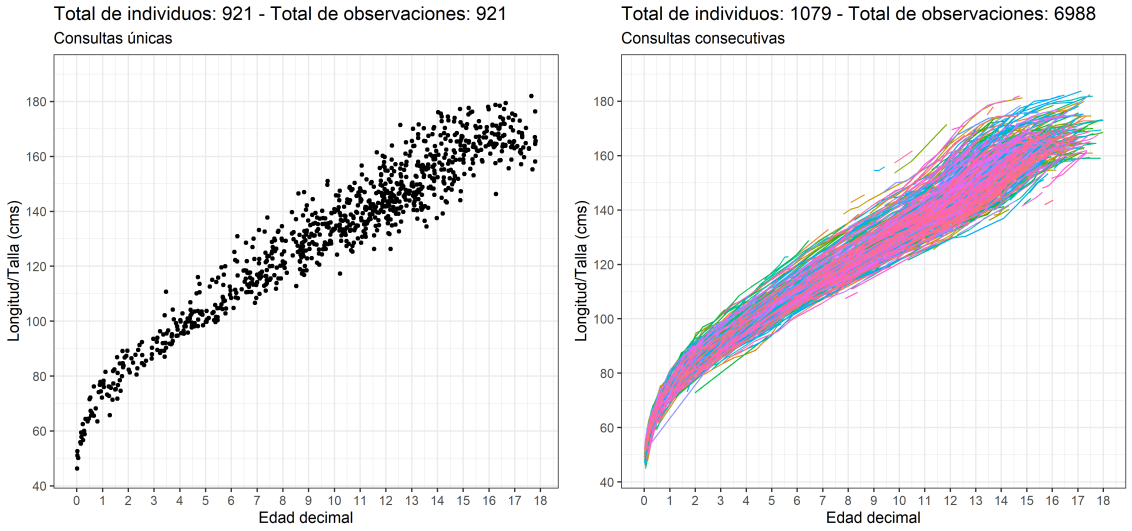



FIGURA 4.1. Longitud/talla por edad decimal diferenciando por naturaleza de los datos.

Las siguientes secciones presentan el modelamiento transversal de los datos de única ocasión y el longitudinal para las consultas consecutivas, buscando estimar curvas de crecimiento diferenciadas para pacientes de primera vez en consulta y pacientes para los que se cuenta con información antropométrica previa. Desde la perspectiva transversal, este modelamiento independiente facilita el contraste de los datos del estudio con las curvas de referencia locales y estándares internacionales, además, posibilita la evaluación clínica de individuos con observaciones únicas. Desde la perspectiva longitudinal, esta distinción permite interpretar adecuadamente las medidas repetidas en la detección de patrones inusuales de crecimiento mediante la reconstrucción de trayectorias individuales.

## 4.3. Modelamiento transversal

Esta sección presenta los resultados del modelamiento transversal con y sin asimetría en los datos junto a la comparación del modelo seleccionado respecto a los estándares de crecimiento internacionales y referencias locales. El proceso de estimación es adaptado a partir del modelamiento GAMLSS de la OMS descrito en la subsección 3.2.1.2 para los datos de única ocasión. La estimación de los modelos y los gráficos diagnóstico son realizados mediante el paquete `gamlss` en R (Stasinopoulos et al., 2017)

### 4.3.1. Centralidad y dispersión

Inicialmente se encuentra el valor de $\lambda$ bajo el mismo modelo inicial para $\mu$ y $\sigma$ que el usado en el modelamiento de la longitud/talla infantil para los niños de la OMS. Luego, dada la transformación $x^{\lambda}$, los valores de $\nu$ y $\tau$ son fijados en 1 y 2 respectivamente para encontrar $gl(\mu)$ y $gl(\sigma)$ tomando todas las posibles combinaciones de $gl(\mu)$ de 5 a 15 y de $gl(\sigma)$ de 2 a 10 (WHO, 2006b) y buscando la combinación con menor $GAIC(3)$. Un siguiente paso busca el valor de $\lambda$ bajo la combinación de parámetros encontrada para $gl(\mu)$ y $gl(\sigma)$. El modelo resultante asume una distribución de Box-Cox potencia exponencial de parámetros df($\mu$)=9, df($\sigma$)=5, $\nu$=1 y $\tau$=2 para la variable respuesta en función de la edad$^{0.7}$. Finalmente, se realiza el diagnóstico del modelo GAMLSS en donde la figura 4.2 evidencia de manera aproximada un comportamiento normal en los residuales. Los gráficos de dispersión de residuales presentan un patrón aleatorio sin ninguna tendencia aparente. La estimación de la densidad de los residuales es cercana a la distribución normal con un pequeño sesgo a la derecha, también evidenciado en el gráfico QQ plot, que además exalta un residual inusualmente alto y cercano a 4. En resumen, se presenta una leve asimetría en la distribución de los residuales que sugiere un modelamiento de la simetría de los datos.

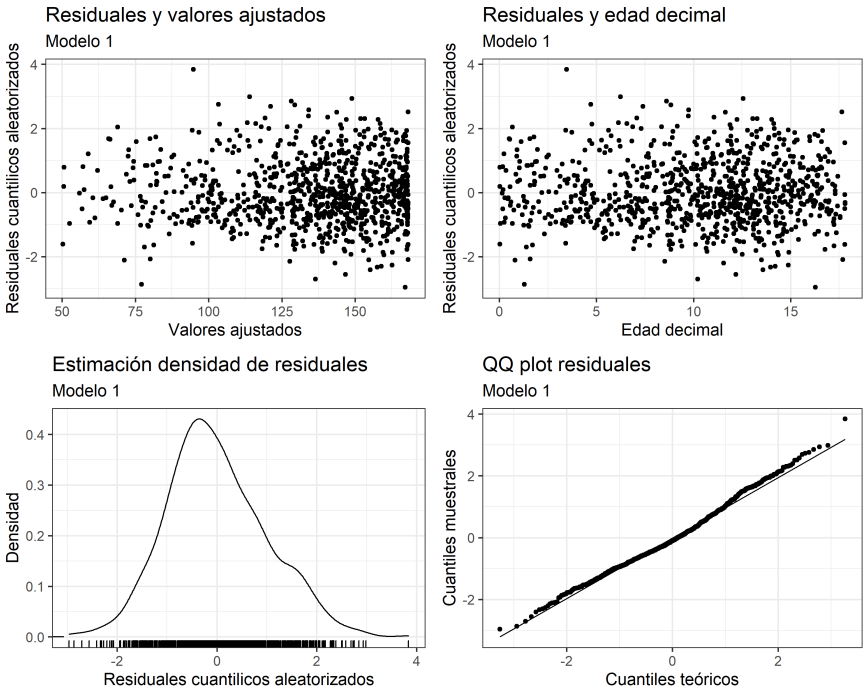
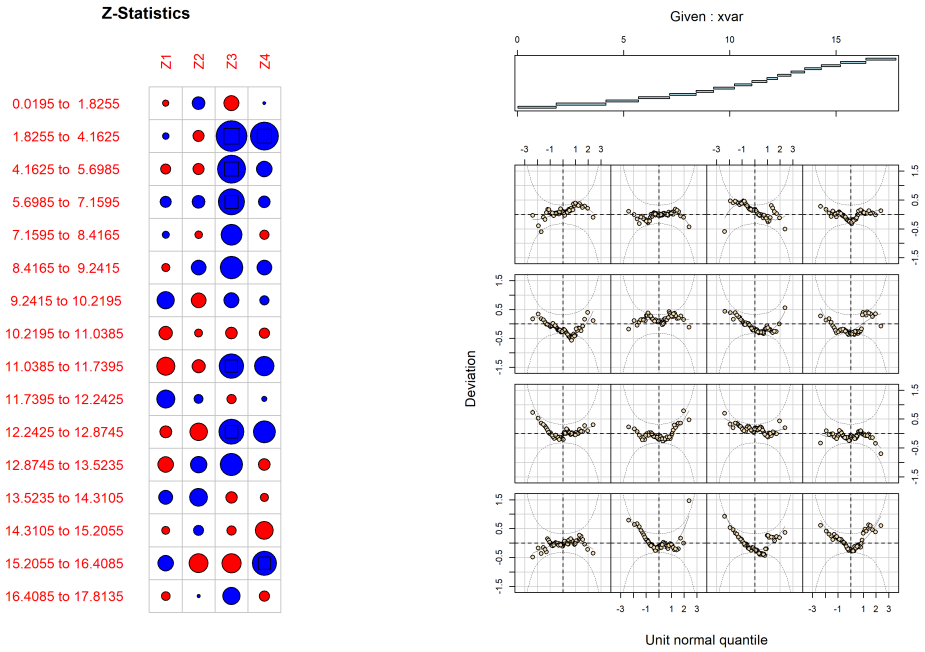



FIGURA 4.2. Descripción de residuales. Modelo GAMLSS (centralidad, dispersión)

Con el objetivo de identificar edades en las cuales el modelo no caracteriza de manera apropiada el patrón de los datos, los estadísticos $Z$ y los gráficos de oruga permiten evaluar la normalidad de los residuales en intervalos disyuntos de la variable independiente. Una descripción detallada de la interpretación de los gráficos y de los estadísticos puede ser encontrada en el capítulo 12 de obra de Stasinopoulos et al. (2017). En resumen, bajo supuestos de normalidad, los estadísticos $Z$ calculados con los residuales del intervalo de edad $g$ $(Z_{g1}, Z_{g2}, Z_{g3}, Z_{g4})$ siguen una distribución normal estándar. Stasinopoulos et al. (2017) indican que valores de $Z_{gj}$ mayores (menores) a 2 (-2) con $g = 1, 2, 3, 4$ indican una mayor (menor) media, varianza, simetría, curtosis respecto a una distribución normal estándar para el grupo de edad $g$. Adicionalmente, los gráficos de oruga presentan una figura análoga al QQ plot para cada intervalo de la variable independiente. Se espera bajo normalidad estándar que los residuales se encuentren sobre la línea horizontal del gráfico. Diferencias importantes respecto a este patrón son potenciales indicativos de inadecuación del modelo en el intervalo de edad correspondiente. Las curvas elípticas de los gráficos de oruga representan intervalos de confianza al 95 % para los residuales del modelo. Bajo un modelo adecuado, se espera que un 95 % de los puntos se encuentren comprendidos entre las elipses y un 5 % fuera de ellas. Un mayor porcentaje de puntos por fuera de la región indica problemas de ajuste en el modelo.

La figura 4.3 (A) presenta los estadísticos $Z$ para 16 intervalos de la variable edad decimal cada uno con aproximadamente el mismo número de residuales. El color del círculo representa el signo del estadístico (azul positivo y rojo negativo) mientras que su tamaño representa la magnitud del estadístico. El cuadrado del círculo indica que el estadístico correspondiente es mayor a 2 en valor absoluto. En la figura 4.3 (B) se muestran los gráficos de oruga para los mismos 16 intervalos de edad decimal, los cuales se leen de abajo a la izquierda a arriba a la derecha. Mediante los dos gráficos se concluye que el modelo requiere ajustar por la asimetría de los datos entre los 2 y los 7 años.




(A) Estadísticos Z por intervalo (B) Gráficos de oruga por intervalo

FIGURA 4.3. Descripción de residuales por intervalo. Modelo GAMLSS (centralidad, dispersión)

### 4.3.2. Centralidad, dispersión y simetría

Bajo un modelo correctamente especificado para los datos, se espera que la distribución de los residuales sea normal estándar (Stasinopoulos et al., 2017), sin embargo, el primer modelo evidencia una distribución sesgada para los residuales. Esto se debe en particular por las observaciones de talla entre los 2 los 7 años, en las cuales el modelo no ajusta la asimetría presente en los datos. Un segundo modelo es ajustado considerando la asimetría en los datos, manteniendo inicialmente los parámetros de modelo anterior y seleccionando $gl(\nu)$ entre 2 y 10 como aquel con menor $GAIC(3)$ (WHO, 2006b). Luego se actualizan los valores de $gl(\mu)$ y $gl(\sigma)$ dado el valor de $gl(\nu)$ encontrado, en un último paso se estima nuevamente el valor de $\lambda$ dados los parámetros anteriores. En la figura 4.4 se observa de manera aproximada un comportamiento normal en los residuales del modelo resultante BCPE(x=edad$^{0.7}$,df($\mu$)=10,df($\sigma$)=6,df($\nu$)=4,$\tau$=2) para la longitud/talla. Los gráficos de dispersión de los residuales evidencian un patrón aleatorio sin ninguna tendencia aparente. Se destaca respecto al modelo anterior que la estimación de la densidad de los residuales no presenta el leve sesgo observado anteriormente, además el gráfico QQ no evidencia observaciones inusuales. En síntesis, la leve asimetría en la distribución de los residuales del modelo anterior es controlada con la inclusión del parámetro correspondiente.

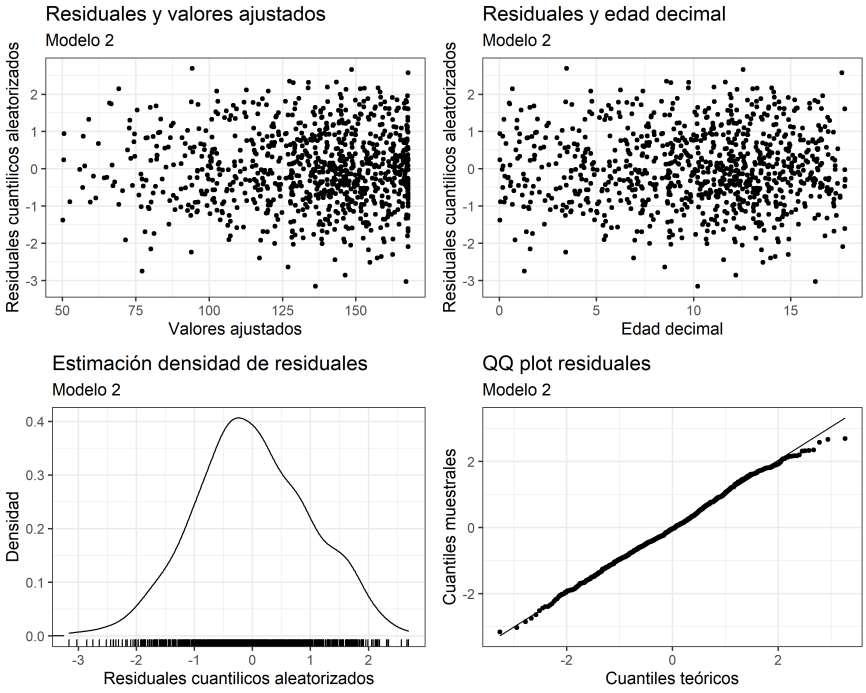
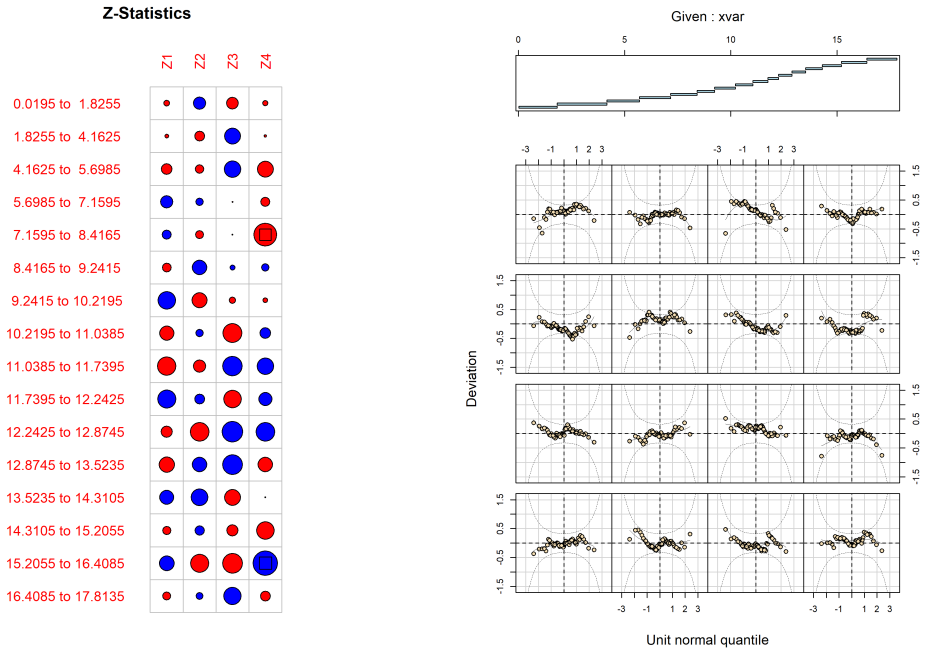



FIGURA 4.4. Descripción de residuales. Modelo GAMLSS (centralidad, dispersión, simetría)

La figura 4.5 (A) presenta los estadísticos $Z$ para los mismos 16 intervalos de la variable edad decimal del modelo anterior y en la figura 4.5 (B) se muestran los gráficos de oruga correspondientes. Para los intervalos de edad entre los 2 y los 7 años el valor del estadístico $Z_{g3}$ es ahora mucho menor que 2 en valor absoluto. Los demás gráficos de oruga se ven muy similares respecto al modelo anterior, reforzando la hipótesis que las edades de 2 a 7 años son las que presentan sesgo no despreciable de talla en función de la edad mientras que en los demás intervalos se tiene un ajuste apropiado de los datos.




(A) Estadísticos Z por intervalo (B) Gráficos de oruga por intervalo

FIGURA 4.5. Descripción de residuales por intervalo. Modelo GAMLSS (centralidad, dispersión, simetría)

En el diagnóstico de los dos modelos estimados se presenta una leve curtosis en grupos particulares de edad. Como menciona Stasinopoulos et al. (2017), en general no es posible encontrar un modelo que se ajuste de manera apropiada en todos los subintervalos de la variable independiente, además, dadas las características de la variable longitud/talla destacadas en el documento, se opta por omitir la curtosis en el modelamiento.

### 4.3.3. Selección del modelo y comparación de resultados

Los resultados bajo los dos modelos ajustados son presentados en la figura 4.6, la cual muestra las funciones percentilicas 5, 25, 50, 75 y 95 de talla en función de la edad junto a los datos originales. Se resalta que la metodología GAMLSS no permite restricciones de monotonicidad en la estimación de las funciones percentílicas, esto justifica el leve decrecimiento en los percentiles estimados más altos bajo ambos modelos, que sumado a la menor cantidad de información para las edades mayores resulta en estimaciones menos confiables para estas edades. Dado que la variable talla en función de la edad es monótona creciente, los percentiles 75 y 95 se mantienen constantes a partir de la edad del primer decrecimiento en posteriores figuras. Se menciona además que previo al modelamiento 0.7 centímetros fueron añadidos a las tallas posteriores a los dos años dado que la diferencia entre las mediciones antropométricas de talla y longitud infantil es de aproximadamente 7 milímetros (WHO, 2006b). Una vez ajustado el modelo, esta corrección es invertida para estas edades con el objetivo de obtener una única curva estándar de crecimiento para talla y longitud infantil, de manera semejante a las curvas de la OMS, lo cual explica la discontinuidad a los dos años en la figura 4.6. Esta característica propia de los datos de crecimiento infantil no se menciona en el modelamiento GAMLSS de Durán et al. (2016).

Ambos modelos capturan las no linealidades de la talla en función de la edad, además, en las primeras edades se nota un rápido crecimiento que es modelado a través de la transformación de potencia. Se destaca en las curvas percentilicas ajustadas de ambos modelos el aumento de la variabilidad en función de la edad, acorde con los datos originales. En general se evidencia una alta concordancia entre los datos y los percentiles estimados bajo los dos modelos a través de la edad, a excepción del intervalo de 2 a 7 años, en el cual el primer modelo presenta una preponderancia de información en los percentiles bajos, lo cual implica una asimetría para los datos la cual es controlada en el segundo modelo. Esta asimetría para la variable longitud/talla no es justificada en la literatura (Cole, 1988; WHO, 2006b) y además en este grupo de edad se cuenta con una menor cantidad de información para caracterizar el comportamiento de los datos, por lo cual se considera que la asimetría presente es artificial y debida a la falta de información en el intervalo de edad. Se opta por seleccionar el modelo menos complejo para representar el comportamiento de la muestra, de manera similar a las curvas de la OMS.

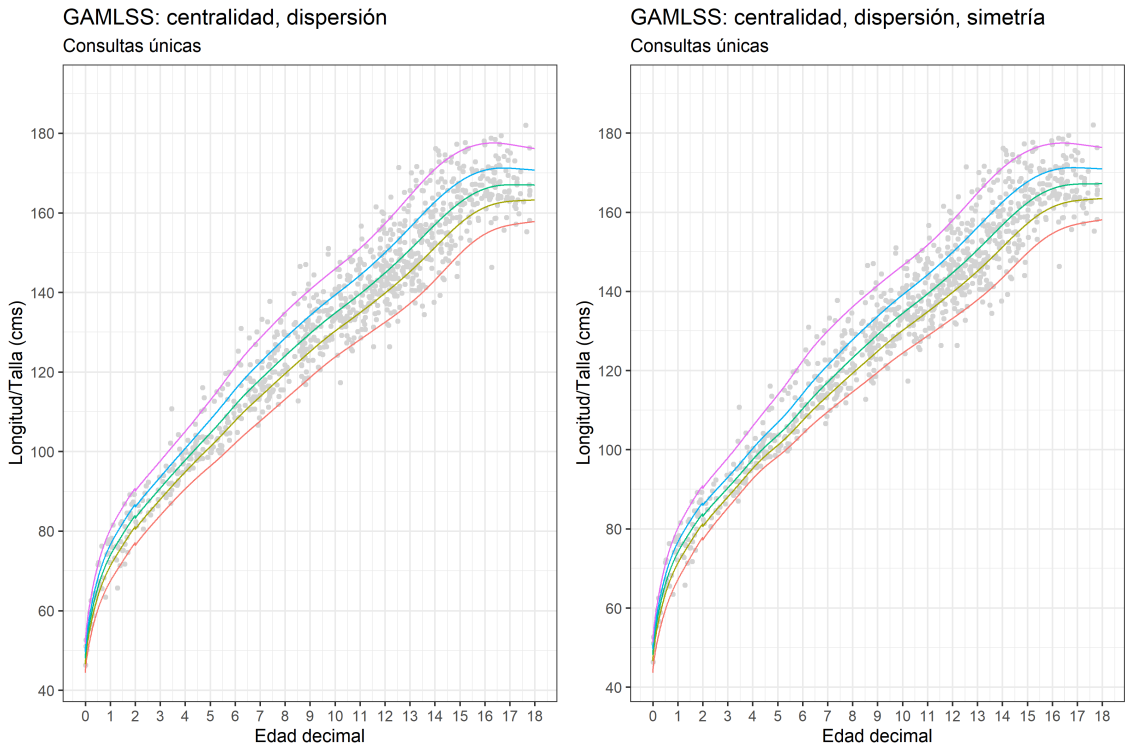



FIGURA 4.6. Talla por edad decimal junto a curvas de crecimiento estimadas de 0 a 18 años: percentiles 5, 25, 50, 75, 95. Modelamiento GAMLSS con (derecha) y sin (izquierda) simetría.

Se contrastan los resultados obtenidos bajo el modelo seleccionado con las curvas estándar de crecimiento de la OMS de manera diferenciada para la longitud y la talla. En la figura 4.7 se observa la comparación de las curvas estimadas respecto a los estándares internacionales de la OMS, en la cual las curvas de crecimiento calculadas son resaltadas y además son presentadas como líneas continuas, el estándar internacional de la OMS se presenta mediante líneas delgadas discontinuas. En ambos gráficos se destacan diferencias entre las curvas calculadas y el estándar internacional de la OMS: en general los niños del estándar internacional son más grandes y en los primeros dos años las curvas calculadas presentan mayor variabilidad.

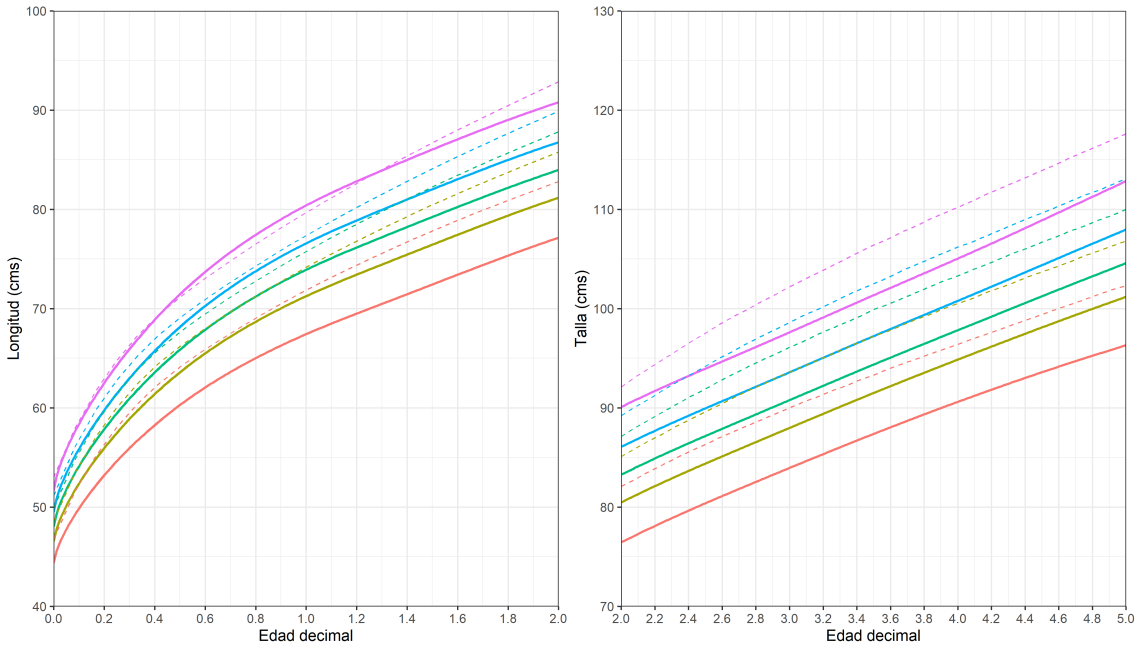



FIGURA 4.7. Curvas de crecimiento estimadas (continuas) y estándar de la OMS (discontinuas) de 0 a 5 años: percentiles 5, 25, 50, 75, 95. Modelamiento GAMLSS.

El aumento de variabilidad observado entre los 0 y 2 años se debe principalmente al diseño longitudinal utilizado en la construcción de las curvas de la OMS, diferente a la información de única ocasión recolectada en las curvas calculadas: en la estimación de las curvas de crecimiento de la OMS se garantizó representatividad para edades de particular importancia en la caracterización del crecimiento infantil, lo cual no es alcanzado con la información retrospectiva de una sola ocasión recolectada en el cálculo de las curvas, lo que se traduce en mayor variabilidad en los percentiles estimados. Para curvas de la OMS, en el componente longitudinal se realizaron un total de 21 mediciones: en las semanas 1, 2, 4 y 6; de manera mensual desde los 2 hasta los 12 meses y bimestralmente en el segundo año (WHO, 2006b). Además, a diferencia del conjunto actual de datos, los niños de la OMS fueron predominantemente amamantados en su infancia, lo cual representa una menor variabilidad en las observaciones de talla del estándar (Marques et al., 2004). La diferencia entre los percentiles calculados y su contraparte de la OMS incrementa en función de la edad para los primeros dos años, pero parece estabilizarse desde los dos hasta los cinco años, en los que se mantiene de manera aproximada la distancia entre los percentiles. A partir de los dos años se evidencia claramente que en general los niños de la muestra son más pequeños respecto a los niños de la OMS.

La diferencia en el tamaño puede deberse a diferentes factores: una tendencia secular de crecimiento presente en los datos, en la cual las medias antropométricas de una población dada evolucionan en el tiempo debido a factores genéticos y ambientales (Van Wieringen, 1978), es descartada dado que un 88 % de la información usada en la construcción de las curvas corresponde a la década de 2008 a 2018, haciendo los datos obtenidos comparables con los estándares internacionales. Otro posible factor es la talla materna, que al estar asociada de manera positiva con la talla infantil (Addo et al., 2013), pueda explicar que la menor talla en los individuos de la muestra se deba a una preponderancia de madres bajas respecto a las madres del estándar internacional. Dado que cerca del 90 % de las observaciones de los pacientes tienen su talla materna correspondiente, estos datos son

comparados con la distribución esperada para las mujeres de 18 años bajo las curvas de la OMS, las cuales fueron posteriormente complementadas para edades mayores (Onis et al., 2007). La figura 4.8 (izquierda) presenta la distribución de los datos de única ocasión con talla materna disponible diferenciando por los cuartiles de talla de la OMS para mujeres de 18 años.

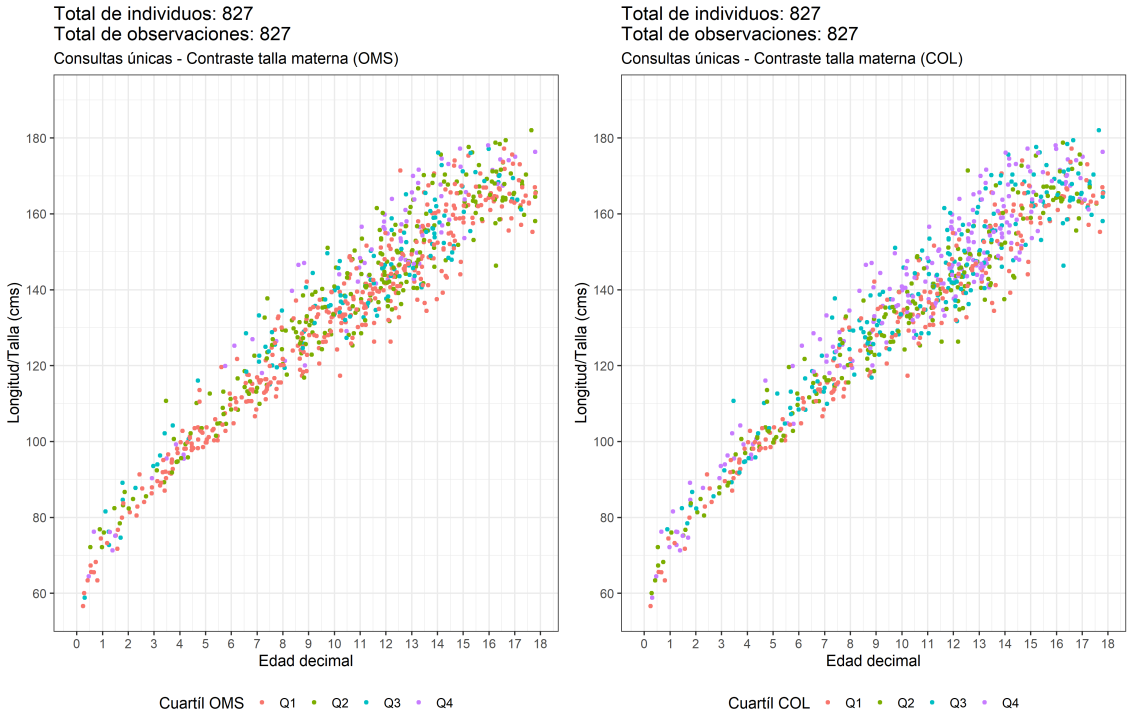



FIGURA 4.8. Talla por edad decimal de 0 a 18 años para niños con talla materna disponible y diferenciando por cuartiles maternos nacionales (derecha) e internacionales (izquierda).

La figura 4.8 (izquierda) evidencia que cerca de la mitad de las madres (45 %) están ubicadas en el primer cuartil de talla de la OMS mientras que una pequeña proporción se encuentra en el cuarto cuartíl (10 %). Esto se mantiene a través de las diferentes edades, evidenciando que las madres de los niños en la muestra son en general pequeñas respecto al estándar de la OMS para mujeres de 18 años. Este comportamiento puede explicar la diferencia en magnitud de los percentiles estimados, ya que las observaciones de talla infantil provienen de madres pequeñas respecto a la talla del estándar internacional con la cual están siendo comparadas. En la figura 4.8 (derecha) se contrasta la talla materna con la distribución esperada para las mujeres de 18 años bajo las curvas colombianas, la gráfica evidencia un comportamiento mucho más homogéneo para los diferentes percentiles a través de la edad. Por lo cual, las madres en la muestra de única ocasión tienen una estatura similar respecto a las curvas colombianas para mujeres de 18 años.

La figura 4.9 presenta la comparación de las curvas de crecimiento estimadas con las curvas de referencia de Colombia para las edades de 0 a 18 años. El contraste es similar al observado para las curvas de crecimiento de la OMS: mayor variabilidad para los primeros dos años y mayor concordancia entre los percentiles a partir de los tres años, esta afinidad se destaca particularmente en los percentiles centrales. La variabilidad en el primer año de vida nuevamente puede verse justificada por el componente longitudinal en las curvas colombianas, las cuales realizaron un seguimiento mensual al crecimiento infantil durante

el primer año de vida (Durán et al., 2016). Respecto al cambio en el tamaño observado por los individuos, por una parte, las curvas Colombianas provienen de las ciudades de Medellín, Cali y Barranquilla, además de la cuidad de Bogotá, estas ciudades se encuentran a una altitud geográfica significativamente menor a la de Bogotá. Esto es relevante porque se espera en ciudades de mayor altitud una menor talla respecto a lugares de menor altitud geográfica, aún para poblaciones privilegiadas (Stinson, 1982). Por otra parte, las curvas colombianas proveen un comportamiento de crecimiento esperado sin controlar por la madurez del individuo y en este caso, la diferencia observada entre las curvas puede deberse a diferencias en la maduración de los niños de la muestra de única ocasión.

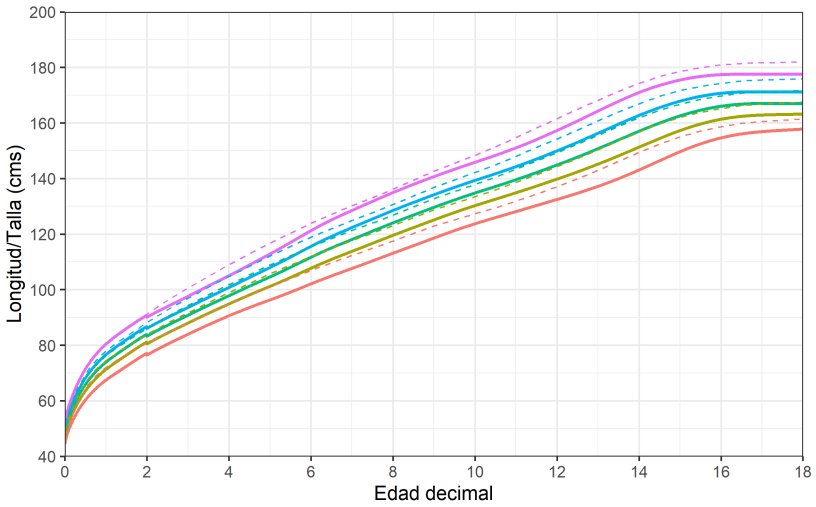



FIGURA 4.9. Curvas de crecimiento estimadas (continuas) y curvas de referencia Colombianas (discontinuas) de 0 a 18 años: percentiles 5, 25, 50, 75, 95. Modelamiento GAMLSS.

Se resalta nuevamente que las funciones percentílicas de la figura 4.9 utilizan como insumo los datos de única ocasión visualizados en la figura 4.1 bajo un enfoque transversal en el modelamiento. Estas curvas estimadas son apropiadas en la evaluación médica del crecimiento en talla para individuos con observaciónes únicas. Para las demás observaciones se hace un modelamiento longitudinal el cual es descrito e implementado a continuación.

## 4.4. Modelamiento longitudinal

En la siguiente sección se presentan los resultados del modelamiento longitudinal usando la información correspondiente a los pacientes con múltiples consultas, el objetivo es diferenciar el análisis funcional de la variable talla por la madurez de los individuos en la muestra. Se selecciona el modelo SITAR para definir grupos de individuos homogéneos en su maduración para posteriormente estimar en cada grupo funciones de covarianza mediante el procedimiento FACE-S. A diferencia de FACE-S, bajo SITAR es posible controlar el cambio de fase en el crecimiento infantil dado por el reloj biológico de cada paciente. Por otra parte, FACE-S, a diferencia de SITAR, es apropiado para un modelamiento en todo el dominio de interés y no sólo en la pubertad. Además, la obtención de la función de covarianza permite una posterior implementación del análisis en componentes principales funcionales para determinar las fuentes de variación más importantes en los datos respecto a la función media.

La primera derivada de las curvas individuales bajo el modelo SITAR es usada para determinar la edad en la que es alcanzada una máxima velocidad de crecimiento para cada individuo (es decir, el APHV). Posteriormente, el APHV es utilizado para segmentar el conjunto de datos con el objetivo de controlar efectos de confusión por la madurez del paciente en su evaluación clínica y de diferenciar los procesos aleatorios subyacentes en la estimación de los parámetros funcionales. Para cada madurez establecida por el modelo SITAR es obtenida la estimación de la función media y del operador de covarianza bajo el modelo FACE-S. Usando la tanto la función media como la función de covarianza obtenida mediante FACE-S, los resultados elaborados en la subsección 3.3.1.3 bajo el modelo PACE son implementados para la predicción de las curvas y de los puntajes muestrales para cada grupo de maduración. Son los resultados de la función media y de la función de covarianza bajo FACE-S, y no los obtenidos mediante PACE, los usados en la estimación de las curvas y de los puntajes muestrales mediante esperanza condicionada.

### 4.4.1. Modelamiento longitudinal para talla

Como destaca Kelly et al. (2014), variantes normales de maduración en la pubertad pueden resultar en desviaciones significativas respecto a percentiles de talla, por lo cual es importante caracterizar la madurez individual en la determinación del estado de salud del paciente. La ecuación 3.6 presenta un modelo longitudinal para talla en la pubertad, el cual permite resumir las desviaciones individuales respecto a la función media $h$ mediante tres efectos aleatorios: $\alpha_i$ (tamaño), $\beta_i$ (madurez) y $\gamma_i$ (velocidad). Para las observaciones longitudinales en el intervalo de 8 a 18 años el modelo es ajustado mediante el uso del paquete `sitar` en el software R con el objetivo de diferenciar el patrón de crecimiento en función de la madurez del individuo. La figura 4.10 (izquierda) presenta la función media estimada para talla, respecto a la cual los efectos aleatorios del modelo cuantifican las desviaciones individuales para cada uno de los pacientes. La figura 4.10 (derecha) presenta la velocidad media de crecimiento de la variable talla, estimada mediante la primera derivada de la función media, en la gráfica se destaca que la APHV para la función media estimada es de 13.3 años aproximadamente.

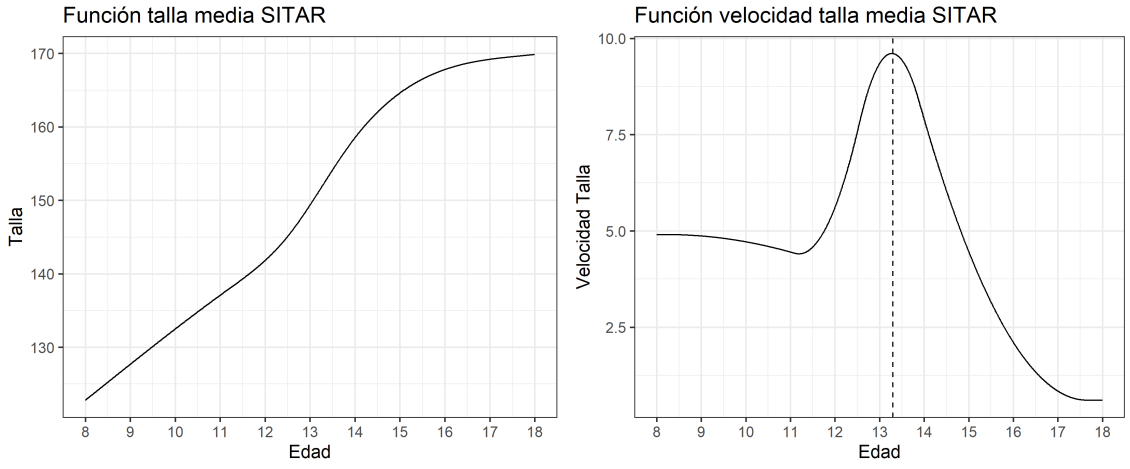



FIGURA 4.10. Función talla media (izquierda) y velocidad de talla media junto a línea punteada de APHV media (derecha) bajo modelo SITAR.

Bajo el escenario escaso del presente trabajo, las curvas individuales en la estimación del APHV son obtenidas mediante SITAR, ya que a diferencia del procedimiento de la subsección 2.1.2, no es posible reconstruir las trayectorias individuales sin combinar la información muestral disponible. Por otra parte, estimar las curvas individuales mediante componentes principales asume que los individuos comparten un mismo proceso aleatorio independiente del reloj biológico del paciente. Se opta así por una segmentación homogénea de los individuos por madurez previo al modelamiento mediante componentes principales.

Respecto a la obtención del APHV, dado que el efecto aleatorio $\beta_i$ induce una traslación horizontal de $i$-ésimo individuo respecto a la función media, este permite determinar el APHV de manera individual para cada paciente. Los individuos con mayor APHV (mayor $\beta_i$) presentan una madurez tardía respecto al individuo promedio mientras que individuos con APHV baja (menor $\beta_i$) tienen una madurez temprana respecto al individuo promedio. Por otra parte, Simpkin et al. (2017) obtienen el APHV mediante la primera derivada de las funciones individuales reconstruidas bajo el modelo estimado. Por facilidad en la interpretación y comparabilidad con otros trabajos, la alternativa de Simpkin et al. (2017) es adoptada.

Para el ajuste del modelo SITAR, son seleccionados 4 $gl$ en la definición del spline natural y se considera una estructura general en la correlación de los efectos aleatorios[1]. Una vez ajustado el modelo, se tiene que el APHV mediano para la muestra es 13.35 años, el primer y tercer cuartil de APHV es $Q_1 = 12.59$ años y $Q_3 = 13.97$ años respectivamente. Cada uno de los 839 pacientes con los que el modelo es ajustado son clasificados como maduradores tempranos (APHV menor a $Q_1$, $n = 213$), promedio (APHV entre $Q_1$ y $Q_3$, $n = 416$) o tardío (APHV mayor a $Q_1$, $n = 210$). La figura 4.11 muestra los registros longitudinales diferenciando por la madurez de los individuos. Se distingue que la madurez de los individuos induce una variación de fase, el desplazamiento es particularmente evidenciado para los maduradores tempranos.

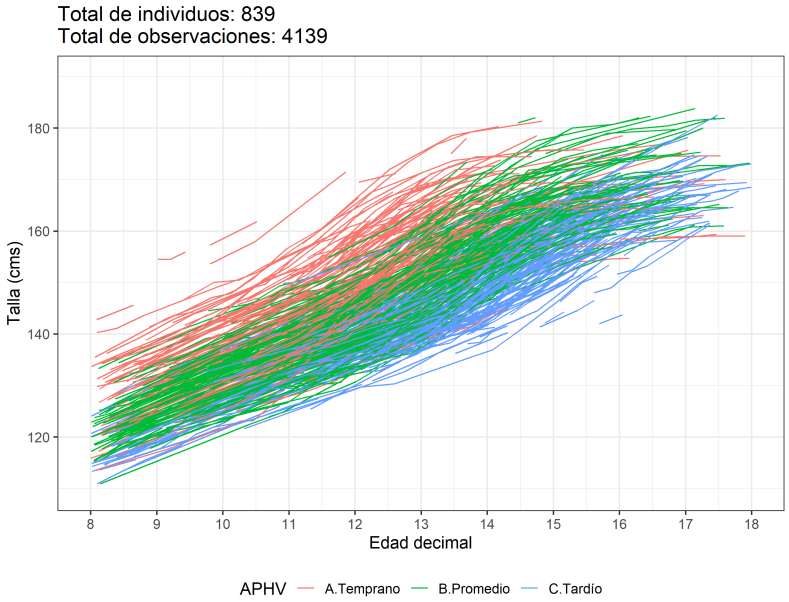



FIGURA 4.11. Observaciones longitudinales de talla para niños de 8 a 18 años por madurez.

[1]El modelo fue ajustado transformando a escala logarítmica la edad decimal. Como los resultados bajo el cambio de escala proveen un ajuste similar, el modelo sin transformar fue seleccionado para el análisis.

### 4.4.2. Modelamiento funcional

El conjunto longitudinal de datos es segmentado a partir de los 8 años por la madurez estimada de los individuos mediante el modelo SITAR. Motivado por el principio del modelo FACE-S, en cada subgrupo se obtiene una estimación de la función media mediante suavizamiento p-spline univariado, que a diferencia de la regresión local, no requiere la selección del ancho de banda ni de una función kernel. La elección del parámetro de suavizamiento es realizada mediante validación cruzada dejando un individuo afuera. En las funciones medias estimadas se detectó que algunas presentaban un comportamiento decreciente para las edades mayores, atribuido a que el procedimiento de estimación no captura la monotonicidad de la variable, por lo cual, se mantiene constante la talla a partir de la primera edad de crecimiento inusual. Las funciones medias son obtenidas mediante el paquete `face` del software R y se presentan en la figura 4.12, se destaca una diferencia entre estas tanto en escala como en fase, sugiriendo que las observaciones escasas provienen de procesos aleatorios diferentes y dependientes de la madurez de los individuos.

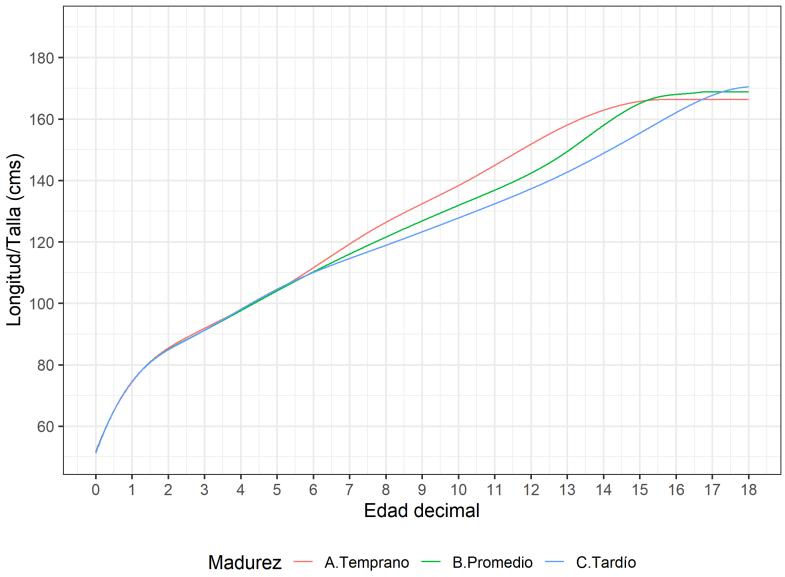



FIGURA 4.12. Funciones medias estimadas de talla para niños de 0 a 18 años por madurez.

Dadas las funciones medias estimadas, los datos funcionales escasos son centrados y la función de covarianza es estimada para cada madurez mediante el procedimiento FACE-S, el cual es implementado en el paquete `face` del software R. Las funciones de correlación son representadas mediante gráficos de contorno en la figura 4.13. Se evidencia a partir de estas figuras una alta correlación positiva entre las observaciones funcionales cercanas en el tiempo, lo cual es esperado para un conjunto de datos de crecimiento humano, además se nota en general una correlación positiva para todos los pares de edades. Maduradores tempranos (derecha superior) evidencian altas correlaciones positivas para sus diferentes edades decimales, mientras que maduradores tardíos (derecha inferior) muestran una alta correlación estimada en edades cercanas. Instantes más distanciados en el tiempo presentan una correlación mucho menor respecto a los maduradores tempranos. Los maduradores promedios (izquierda) exhiben un patrón intermedio entre las correlaciones observadas para los maduradores tempranos y los tardíos, además, resaltan algunas correlaciones negativas en el gráfico de contorno respectivo.

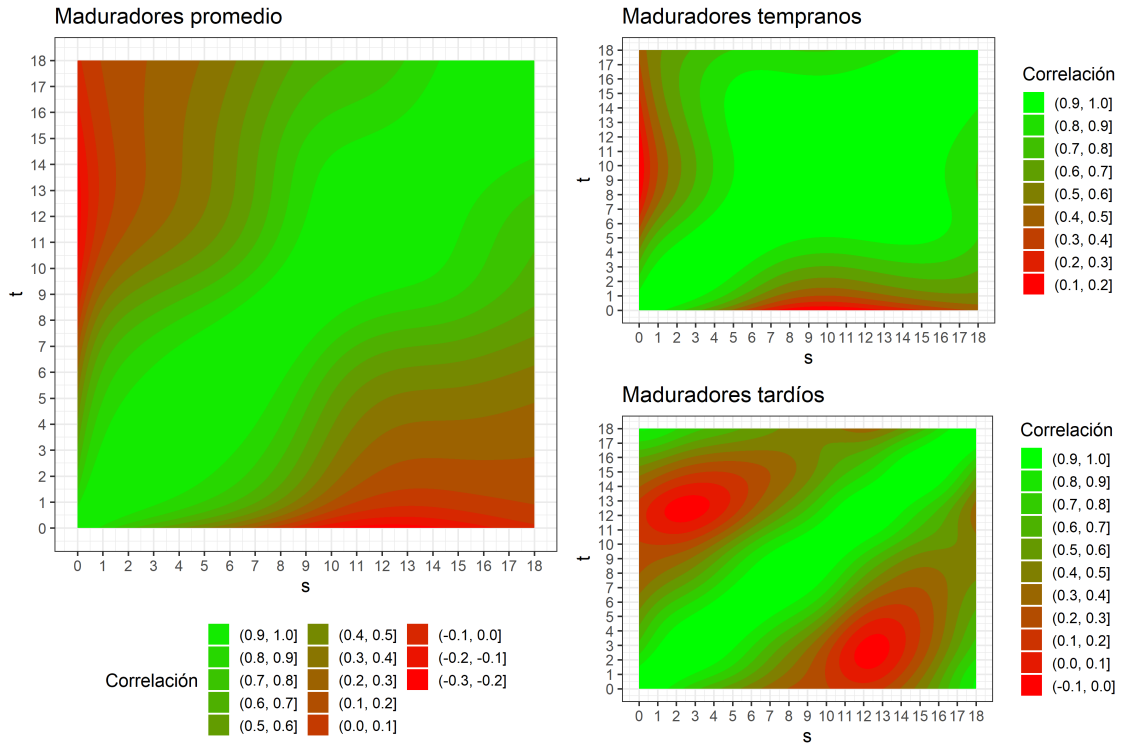



FIGURA 4.13. Contornos de funciones de correlación estimadas de 0 a 18 años por madurez.

Dada la caracterización diferenciada en cada madurez de las propiedades funcionales de primer y segundo orden para los procesos aleatorios funcionales de crecimiento infantil, son encontrados para cada subgrupo los componentes principales funcionales junto a los puntajes muestrales correspondientes. El método utilizado en la estimación de los puntajes sobre los componentes principales es el descrito en la subsección 3.3.1.3, implementado en el paquete `fdapace` en R. La tabla 4.3 presenta los valores propios estimados a partir de las funciones de covarianza junto al porcentaje de varianza acumulado explicado por las funciones propias correspondientes, diferenciando por la madurez de los individuos en la muestra. Los primeros tres componentes para los tres subgrupos explican más del 95 % de la variabilidad respecto a la función media estimada. En particular, el primer componente para la madurez temprana por si solo representa más del 92 % de la variabilidad en la muestra, un porcentaje similar es encontrado para la madurez promedio con los primeros dos componentes, mientras que la madurez tardía requiere los primeros tres componentes para alcanzar un porcentaje de varianza explicada mayor al 90 %.

| | Primeros tres valores propios | | |
|---|---|---|---|
| Madurez | $\hat{\lambda}_1$ | $\hat{\lambda}_2$ | $\hat{\lambda}_3$ |
| Temprano | 749.27 | 37.00 | 22.16 |
| % Acum | 92.62 % | 97.20 % | 99.94 % |
| Promedio | 341.21 | 98.93 | 26.09 |
| % Acum | 72.34 % | 93.31 % | 98.84 % |
| Tardío | 331.08 | 119.42 | 83.37 |
| % Acum | 60.49 % | 82.31 % | 97.55 % |

TABLA 4.3. Valores propios estimados por madurez junto a porcentaje de varianza explicada

La figura 4.14 muestra para cada madurez los primeros tres componentes principales estimados. El primer componente denota un aumento/disminución de la función media estimada, este difiere de manera importante para los tres subgrupos, destacando la diferencia existente entre las medias de los procesos funcionales. El segundo componente en los tres conjuntos de datos presenta un patrón similar, indicando un aumento/disminución en la velocidad de crecimiento, con un cambio en fase evidenciado en el valor mínimo que asume cada una de las funciones para maduradores tempranos y tardíos. La tercera función propia parece diferenciar los individuos con un acelerado crecimiento en la preadolescencia, sin embargo su interpretación no es clara y únicamente es marginalmente relevante para maduradores tardíos. Dado el porcentaje de varianza explicada por los primeros dos componentes en los tres subgrupos y por uniformidad en la selección del número de componentes para cada madurez, se consideran los dos primeros componentes para la estimación de las curvas en los tres grupos de maduración.

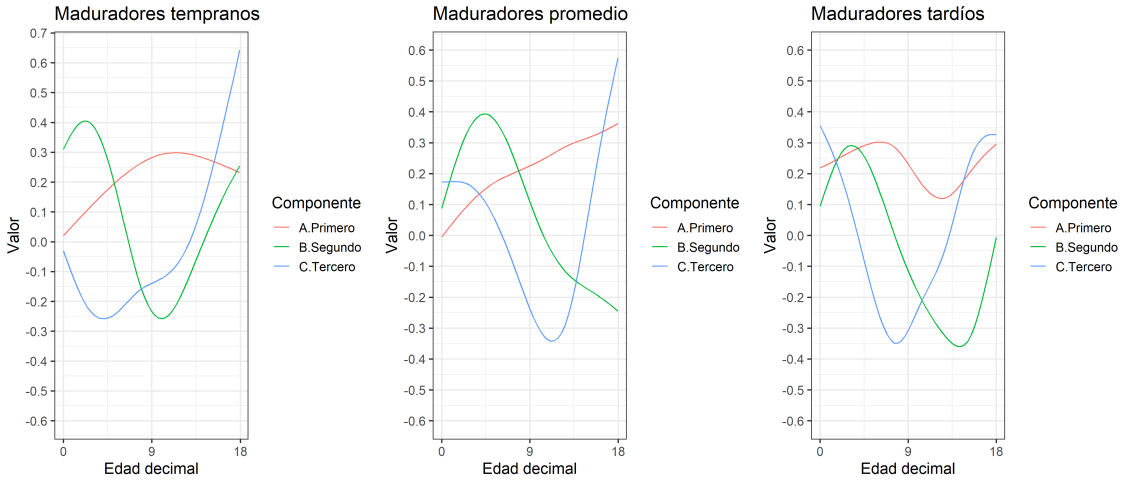



FIGURA 4.14. Primeros componentes principales estimados para niños de 0 a 18 años por madurez

Una vez seleccionado el número de componentes principales y estimados los puntajes muestrales, es posible realizar la estimación de las curvas en el dominio de interés. Con el objetivo de ilustrar el efecto de los componentes principales en la reconstrucción de las curvas, el Apéndice A del trabajo presenta diferentes curvas reconstruidas para cada madurez bajo valores extremos en la distribución de los puntajes muestrales.

La siguiente sección presenta la elaboración de curvas de crecimiento infantil para datos longitudinales mediante un enfoque funcional escaso basado en cuantiles multivariados sobre los puntajes estimados mediante componentes principales. La construcción de curvas de crecimiento infantil para la detección de observaciones inusuales mediante cuantiles multivariados, en contraste a metodologías funcionales para detección de curvas atípicas como el boxplot funcional (Sun & Genton, 2011), permite diferenciar y cuantificar patrones de crecimiento inusuales conjuntos o marginales. Es decir, para un individuo atípico dado el procedimiento permite detectar si el comportamiento inusual se debe a la contribución individual o conjunta de los componentes considerados en la estimación de las curvas. Medir la atipicidad del individuo usando puntajes muestrales permite determinar en el contexto del trabajo actual qué tan atípico es un individuo y si es inusual en su talla general (puntaje sobre el primer componente), en su velocidad de crecimiento (puntaje sobre el segundo componente) o en ambos.

## 4.5. Cuantiles multivariados

Con el objetivo de determinar patrones inusuales en las curvas reconstruidas usando los puntajes muestrales, diferentes propuestas pueden implementarse en la evaluación de la atipicidad de observaciones multivariadas. El bagplot (Rousseeuw et al., 1999), por ejemplo, es una generalización bivariada del boxplot que aplica la profundidad de Tukey (Tukey, 1975) para determinar el ordenamiento multivariado de las observaciones y así extender la definición de la caja y las vallas del boxplot tradicional. Por otra parte, la regresión cuantílica descrita en la sección 3.4 define contornos percentílicos usando los puntajes muestrales, extendiendo así las curvas de crecimiento transversales. El bagplot, a diferencia de la regresión cuantílica, no permite definir los percentiles usuales utilizados en la construcción de curvas de crecimiento, además de estar limitado al caso bidimensional. La noción de profundidad, sin embargo, puede ser explorada en la elaboración de curvas de crecimiento longitudinales. No obstante, al considerar covariables en la construcción de los cuantiles, como por ejemplo la talla materna, para el modelo de regresión su incorporación es inmediata a diferencia de los métodos de profundidad multivariada.

La estimación de los cuantiles distribucionales de la variable talla para la elaboración de las curvas de crecimiento longitudinales es elaborada mediante regresión cuantílica sobre los puntajes muestrales de los componentes principales. El procedimiento de Wei (2008) es implementado en el software R y el código para su estimación es adaptado a partir del material suplementario de Zhang et al. (2015). En la figura 4.15 se observan los cuantiles multivariados concéntricos de talla para niños de 0 a 18 años por madurez. Se representan los cuantiles $\mathcal{C}_{0.50}$, $\mathcal{C}_{0.75}$ y $\mathcal{C}_{0.90}$ de color negro, azul oscuro y azul claro, respectivamente. Estos contornos permiten determinar las regiones centrales de los puntajes muestrales, se espera que cerca del 50 % se encuentren comprendidos en $\mathcal{R}_{0.50}$, 75 % en $\mathcal{R}_{0.75}$ y 90 % en $\mathcal{R}_{0.90}$, encontrando así los análogos funcionales a las curvas de crecimiento convencionales.

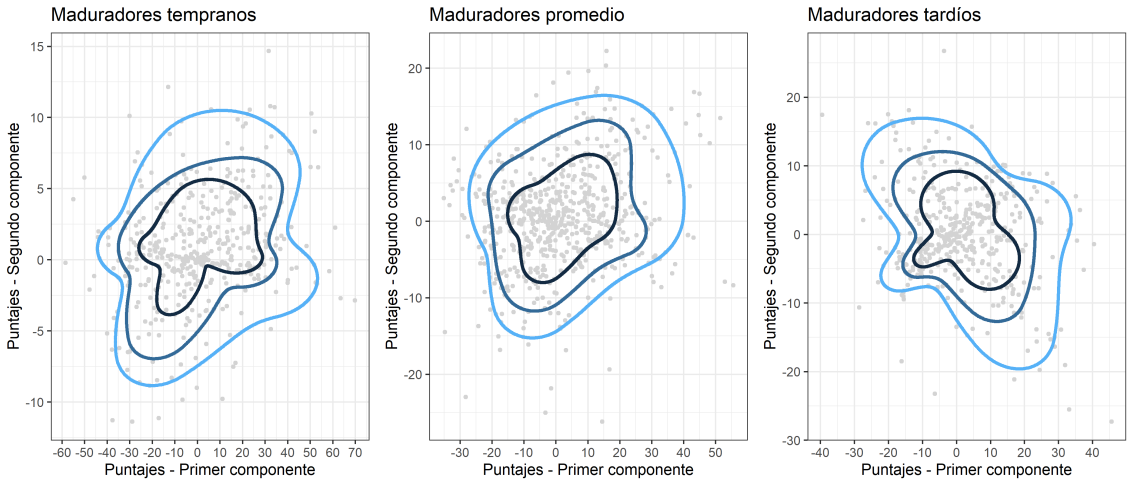



FIGURA 4.15. Cuantiles multivariados concéntricos de talla para niños de 0 a 18 años por madurez: $\mathcal{C}_{0.50}$ (negro), $\mathcal{C}_{0.75}$ (azul oscuro) y $\mathcal{C}_{0.90}$ (azul claro).

Como fue mencionado en la subsección 3.2.2, las curvas de crecimiento convencionales son estimadas mediante estudios de corte transversal que proveen información relevante para mediciones simples, sin embargo, en la práctica clínica se realiza un seguimiento de las múltiples observaciones de un individuo mediante estas curvas. Esta práctica desconoce la estructura longitudinal en los datos y puede identificar erróneamente pacientes sanos como

inusuales y también omitir el seguimiento médico de pacientes con posibles problemas de salud. La figura 4.16 presenta las curvas de crecimiento de única ocasión encontradas en la sección 4.3, percentiles 5, 25, 50, 75 y 95, junto a observaciones longitudinales de algunos de los individuos detectados como inusuales en la distribución conjunta de los puntajes muestrales para cada subgrupo de maduración. En cada madurez se observa que las curvas de crecimiento de única ocasión pueden fallar en detectar patrones inusuales de crecimiento, en particular, los individuos de color verde evidencian patrones anormales de crecimiento que las curvas de única ocasión no logran determinar. Esto de debe a la naturaleza transversal de la metodología subyacente en la construcción de las curvas transversales junto a la marginalización de la madurez de los individuos en su construcción.

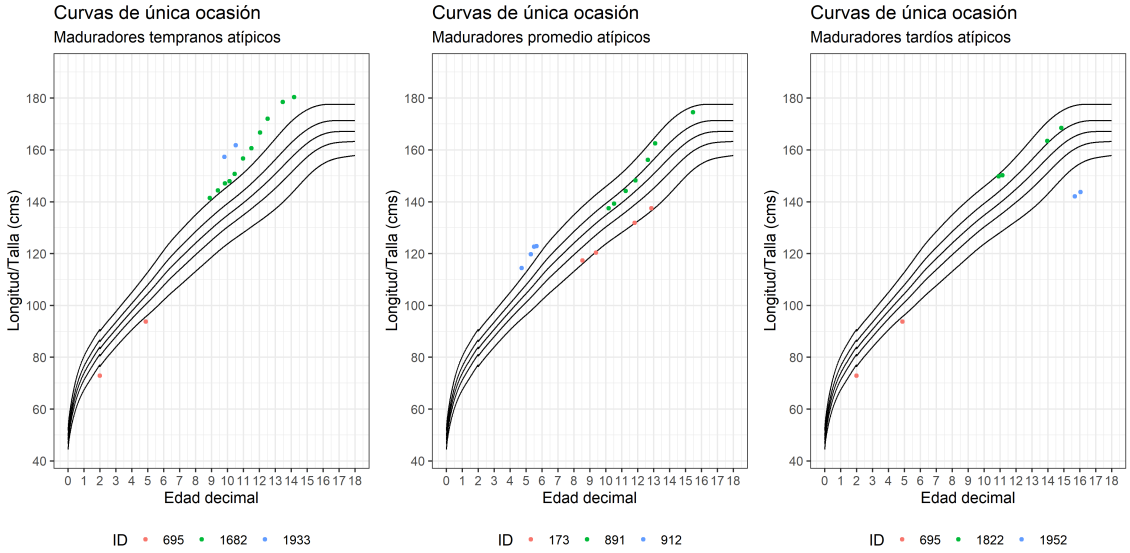



FIGURA 4.16. Observaciones longitudinales atípicas mediante cuantiles multivariados junto a curvas de crecimiento de única ocasión.

Las observaciones correspondientes al individuo 1682 se ubican encima del percentil 95, indicando un paciente bastante alto respecto a la población correspondiente. A pesar de ello, estas curvas no identifican el rápido crecimiento que presenta este sujeto, el cual es un posible indicativo de crecimiento anormal. Al contrastar el comportamiento del paciente 891 con las curvas de única ocasión, se muestra un incremento constante en talla mayor al presentado por los percentiles cercanos, implicando para este sujeto un cruzamiento de centiles que aún así se mantiene dentro de los percentiles estimados. De manera similar sucede para el sujeto 1822, el cual presenta un comportamiento que no es detectado como inusual por las curvas de única ocasión, sin embargo es identificado por las curvas de crecimiento longitudinales como anormal. Se tiene entonces que observaciones de única ocasión provenientes de sujetos con patrones de crecimiento inusuales pueden ser clasificadas como normales al ser contrastadas con curvas de crecimiento tradicionales.

Las curvas de crecimiento construidas mediante cuantiles multivariados sobre los puntajes muestrales permiten identificar patrones de crecimiento anormales al obtener una estimación de la atipicidad del paciente en todo el dominio de interés, aún sin contar con información antropométrica completa. Además, el procedimiento funcional evita la interpretación errónea de las curvas de única ocasión al contar con registros longitudinales de crecimiento infantil (Cole, 1994). Finalmente, este método permite aumentar la especificidad de las curvas al controlar por la madurez de los individuos.

Para las mediciones longitudinales de la figura 4.16, se presenta a continuación la reconstrucción funcional de las observaciones escasas en la nube de curvas estimadas para la madurez correspondiente. Se muestra además la ubicación de los puntajes muestrales de estas observaciones respecto a los cuantiles multivariados estimados mediante regresión cuantílica.

### 4.5.1. Maduradores tempranos atípicos

Para los puntajes atípicos destacados en la figura 4.17 (izquierda) se evidencia que los individuos resaltados como inusuales usando cuantiles multivariados son consistentes con las curvas de única ocasión presentadas anteriormente. Sin embargo, el procedimiento funcional escaso permite además detectar que, por ejemplo, el individuo 1682, al ser atípico tanto en el primer como en el segundo componente, presenta una talla inusualmente alta junto a una alta velocidad de crecimiento en su infancia y en su adolescencia, se evidencia también un desaceleramiento en su preadolescencia que es visualizado en la curva correspondiente de la figura 4.17 (derecha). En resumen, este es un individuo con un patrón de crecimiento inusual respecto a los pacientes maduradores tempranos en el dominio de interés tanto en talla como en velocidad de crecimiento.

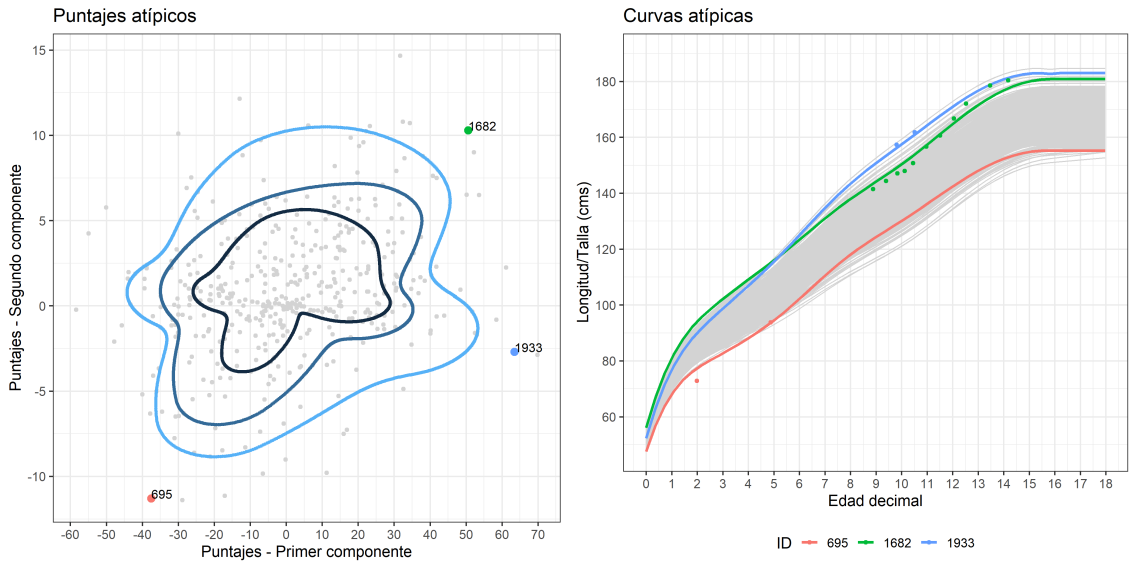



FIGURA 4.17. Puntajes atípicos seleccionados (izquierda) y funciones estimadas correspondientes en la nube de curvas (derecha). Maduradores tempranos.

Por otra parte, aunque los pacientes identificados como 695 y 1933 en el gráfico son inusuales, presentan un comportamiento diferencial por la ubicación de los puntajes estimados correspondientes. Al ser atípico principalmente en el primer componente, el paciente 1933 no presenta anomalías importantes en su velocidad de crecimiento a pesar de ser inusualmente alto, esto se observa en su alto primer puntaje. Por otra parte, el paciente 695 es particularmente bajo, y además presenta una baja velocidad de crecimiento en su infancia seguida de una velocidad alta e inusual en su preadolescencia, lo cual es evidenciado por su bajo primer y segundo puntaje.

### 4.5.2. Maduradores promedio atípicos

De manera similar a los maduradores tempranos son visualizadas algunas observaciones atípicas para maduradores promedio en la figura 4.18. En la figura izquierda se evidencia que los individuos 912 y 891 son atípicos y difieren principalmente en su segundo componente, por lo cual, la talla alcanzada por los pacientes es similar, pero su comportamiento a través de la edad es diferente. El individuo 912 presenta una alta velocidad de crecimiento en su infancia posteriormente aplacada en su preadolescencia, mientras que el individuo 891 presenta una menor velocidad de crecimiento que incrementa en su adolescencia.

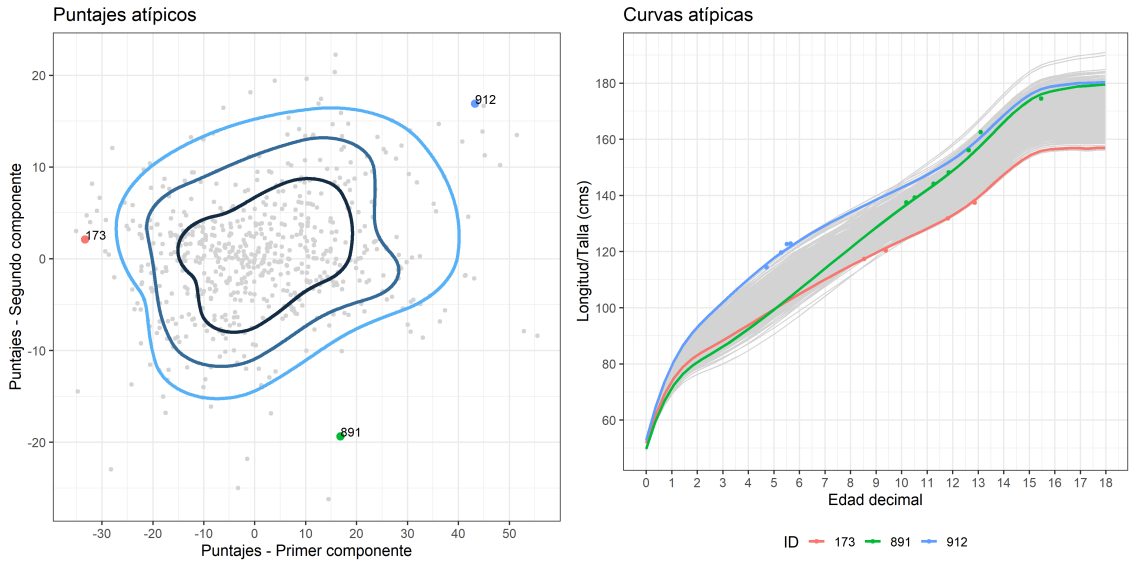



FIGURA 4.18. Puntajes atípicos seleccionados (izquierda) y funciones estimadas correspondientes en la nube de curvas (derecha). Maduradores promedio.

Al controlar por la madurez de los individuos, este procedimiento permite detectar patrones de crecimiento inusuales respecto a la población determinada. Por ejemplo, a diferencia de los pacientes 891 y 912, el paciente 173 se mantiene constante en el quinto percentil de las curvas de única ocasión a través de la edad (ver figura 4.16), lo cual podría indicar un sujeto bajo pero saludable en la población. No obstante, los cuantiles multivariados detectan que este comportamiento es inusual al ser un madurador promedio. Por tanto, al controlar por la madurez, el comportamiento del individuo es detectado por las curvas de crecimiento longitudinales.

### 4.5.3. Maduradores tardíos atípicos

Para los maduradores tardíos, la figura 4.19 presenta los puntajes atípicos junto a la reconstrucción funcional de las observaciones escasas. El individuo 1952 es bastante bajo para ser un madurador tardío, presenta además una velocidad de crecimiento en su adolescencia aproximadamente constante. Esto es dado por el valor negativo sobre el primer componente y el positivo, pero menor en valor absoluto, sobre el segundo. Por el contrario, el sujeto 1822 es en general bastante alto, al tener un primer puntaje mayor a 40. Esto también indica que el paciente presenta una alta velocidad de crecimiento en su infancia, seguida por una desaceleración en su adolescencia, que sin embargo se ve contrarrestada por puntaje sobre el segundo componente.

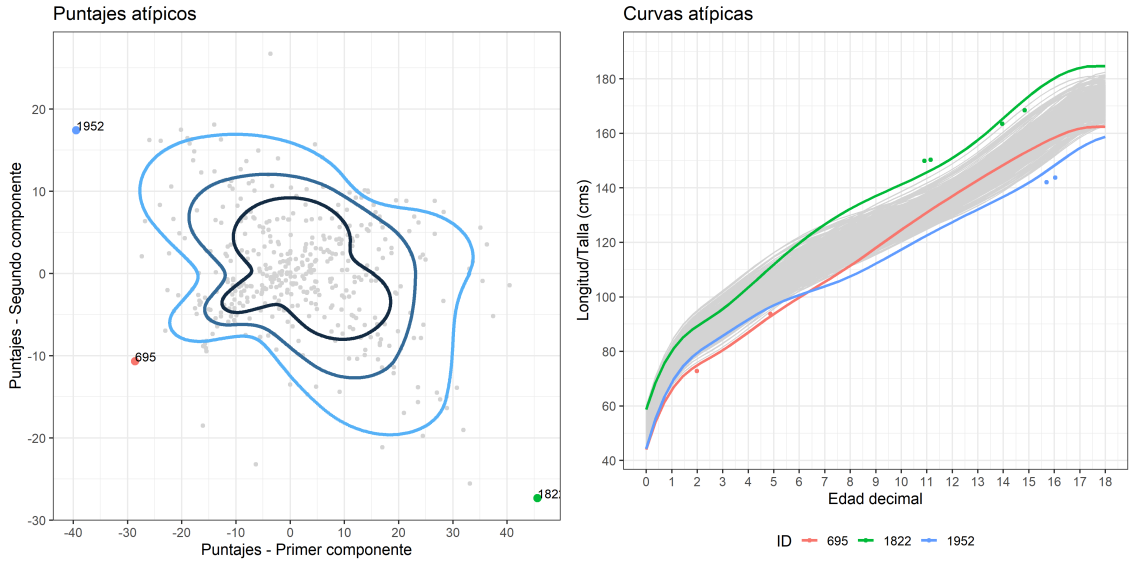



FIGURA 4.19. Puntajes atípicos seleccionados (izquierda) y funciones estimadas correspondientes en la nube de curvas (derecha). Maduradores tardíos.

Dado el procedimiento para la construcción de las curvas, los sujetos con observaciones longitudinales anteriores a los 8 años hacen parte de las tres poblaciones, pues es desconocida su madurez al ser determinada mediante el modelo SITAR. Por ejemplo, el paciente 695 también es representado en la figura 4.17, y en ambos escenarios es detectado como atípico, por lo cual, respecto a ambas subpoblaciones (maduradores tempranos y maduradores tardíos) presenta un comportamiento anormal de crecimiento.

# APÉNDICE A

## Interpretación de componentes principales funcionales

Para cada individuo se tienen observaciones escasas en el dominio de interés, y por medio del análisis componentes principales funcionales, estas son estandarizadas mediante puntajes muestrales y trayectorias individuales son estimadas. Esta reconstrucción bajo modelamiento funcional escaso no permite restricciones en la estimación de las curvas, como las presentadas en la subsección 2.1.2 bajo el escenario funcional denso, por lo cual no se garantiza que las curvas reconstruidas sean crecientes. Esto es particularmente problemático en el modelamiento de la talla infantil, en el cual las trayectorias estimadas pueden indicar un decrecimiento en talla.

Vía correo electrónico fue contactado al profesor Hans-Georg Müller, referente mundial en el análisis de datos funcionales y coautor de la propuesta PACE (Yao et al., 2005), dado que no solo PACE sino además S-FACE y la reconstrucción propuesta por Zhang et al. (2015) presentan el mismo inconveniente en el conjunto actual de datos (los resultados de los últimos dos modelos no son mostrados en el presente trabajo). El doctor Müller confirma en su respuesta que PACE y sus variantes no garantizan monotonicidad en las curvas y sugiere diferenciar las curvas decrecientes estimadas, igualar a cero la derivada donde sea negativa e integrar de vuelta para obtener la función talla, incurriendo así en un sesgo en la estimación para garantizar monotonicidad. Este fenómeno es observado en las edades mayores, dónde se tiene menor cantidad de información para caracterizar el proceso aleatorio funcional y además un menor cambio en talla, por lo cual se opta por mantener la curva constante a partir del instante en el cual decrece la trayectoria.

Buscando ilustrar el efecto de los primeros dos componentes principales estimados y graficados en la figura 4.14, a continuación son presentadas diferentes curvas reconstruidas para cada madurez bajo valores extremos en la distribución condicional de los puntajes muestrales. En caso de obtener curvas decrecientes en las edades mayores, el procedimiento descrito anteriormente es implementado. Las figuras presentan un diagrama de dispersión con los puntajes muestrales destacando las parejas ordenadas de las curvas estimadas correspondientes, son seleccionados valores extremos arbitrarios de los puntajes para cada componente junto a un valor central para contrastar el efecto correspondiente. Esto es realizado para cada uno de los tres conjuntos de datos, añadiendo además las observaciones originales de cada individuo.

En la figura A.1 se evidencia que el primer componente representa un cambio general en el tamaño del paciente madurador temprano. Las curvas estimadas muestran que altos valores en este componente corresponden a un aumento general en talla mientras que valores bajos implican una disminución. Esto hace el primer componente análogo al parámetro $\alpha_i$ del modelamiento SITAR, sin embargo este es generalizable a todo el dominio de interés y no sólo al periodo de la adolescencia. Un alto valor para el segundo componente indica una mayor velocidad de crecimiento en los primeros años de vida, seguida por un desaceleramiento en la preadolescencia que concluye con un posterior aumento de la velocidad de crecimiento en la adolescencia. Valores bajos en este componente caracterizan individuos con menor velocidad de crecimiento en en los primeros años de vida seguida por un aceleramiento en la preadolescencia y una disminución en la adolescencia.

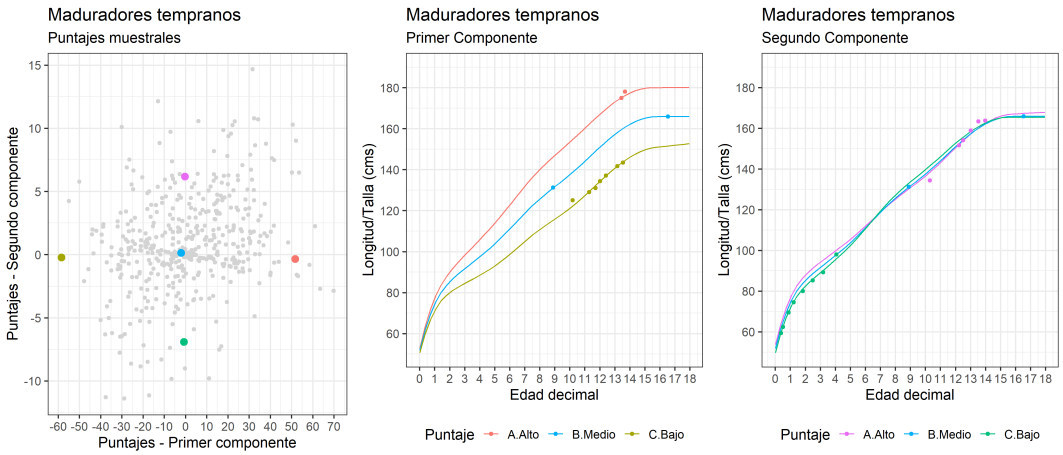



FIGURA A.1. Diagrama de dispersión de puntajes muestales junto a curvas reconstruidas extremas por componente. Maduradores tempranos

Los gráficos para maduradores promedio en la figura A.2 muestran un comportamiento similar al observado previamente: un primer componente que denota una traslación en talla y un segundo que resume la velocidad de crecimiento. Sin embargo, valores positivos (negativos) en el segundo componente distinguen a los pacientes con un rápido (lento) crecimiento en su infancia seguido por una baja (alta) talla en su adolescencia.

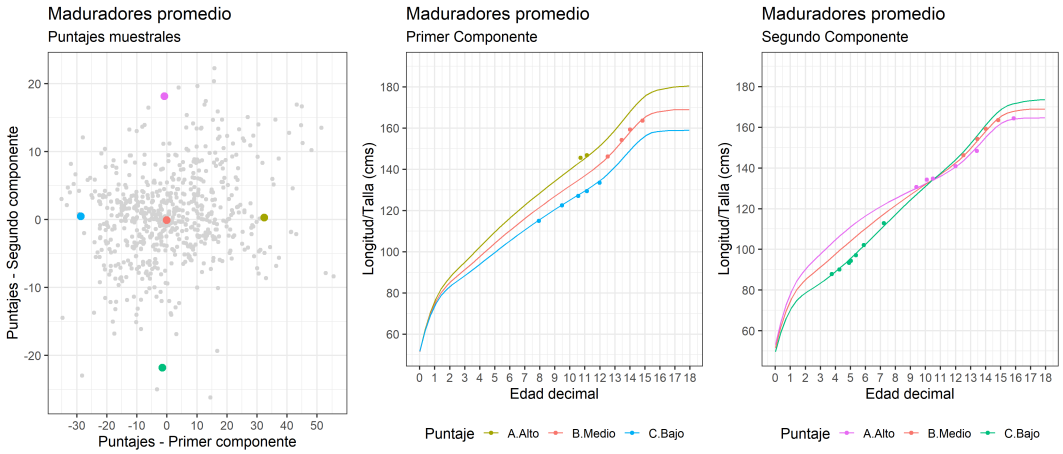



FIGURA A.2. Diagrama de dispersión de puntajes muestales junto a curvas reconstruidas extremas por componente. Maduradores promedio

En la figura A.3 se nota nuevamente un comportamiento similar para los dos primeros componentes principales de los maduradores tardíos respecto a los demás grupos de maduración. El primer componente, sin embargo, además de explicar un aumento general en talla, replica de manera aproximada el comportamiento del segundo componente para maduradores tempranos. Por lo cual, este primer componente para maduradores tardíos diferencia pacientes con un aumento en talla particularmente acentuado en los primeros años de vida que posteriormente es atenuado en la preadolescencia. Para individuos con puntajes negativos sobre este componente, se aplaca el efecto de la velocidad de crecimiento en la infancia y también se disminuye la desaceleración en la preadolescencia.

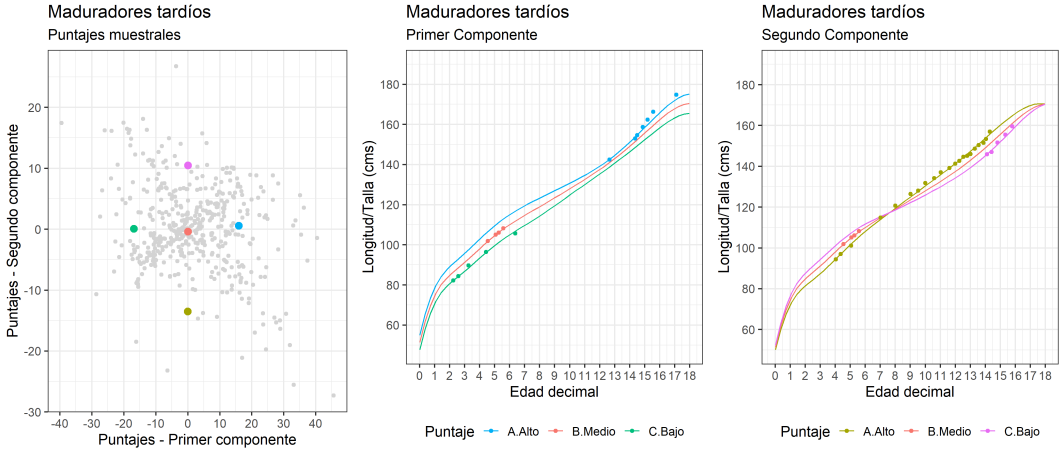



FIGURA A.3. Diagrama de dispersión de puntajes muestales junto a curvas reconstruidas extremas por componente. Maduradores tardíos

En resumen, maduradores tempranos, promedio y tardíos tienen un primer componente por encima de cero para todas las edades, denotando así un aumento (disminución) general en talla para valores positivos (negativos) sobre el componente. Su comportamiento, sin embargo, es diferencial para cada madurez: maduradores tempranos presentan una disminución en la adolescencia, maduradores promedio continúan creciendo hasta los 18 años de edad, maduradores tardíos presentan fluctuaciones que asemejan la velocidad de crecimiento infantil.

Para maduradores tempranos y tardíos respecto al segundo componente se observan dos momentos importantes en el patrón de crecimiento. Dado un puntaje positivo (negativo) sobre el componente, un instante de máximo (mínimo) crecimiento en la infancia es seguido por un momento de mínimo (máximo) crecimiento en la preadolescencia. Para maduradores tempranos esto concluye con un posterior aumento (disminución) de la velocidad de crecimiento en la adolescencia. Para maduradores tardíos en un decremento de la baja (alta) velocidad presentada en la preadolescencia. El primer momento ocurre cerca de los 2-3 años para tanto maduradores tempranos como tardíos, mientras que el segundo momento sucede cerca a los 10 años para maduradores tempranos y hacia los 13 años para los maduradores tardíos. Los maduradores promedio en el segundo componente presentan el mismo primer momento que los maduradores tempranos y tardíos. Sin embargo, el segundo momento descrito anteriormente no presenta una inflexión en la preadolescencia. Es decir, dado un puntaje positivo (negativo) sobre el componente se tiene una velocidad en talla alta (baja) en la infancia seguida de una baja (alta) en la adolescencia, sin una diferenciación entre estos dos momentos de la niñez (a diferencia de los maduradores tempranos y tardíos).

# Conclusiones

A partir de un estudio de caso para la ciudad de Bogotá, el presente trabajo final de maestría presenta la construcción de curvas de crecimiento infantil mediante métodos estadísticos funcionales. Este primer esfuerzo en la construcción de curvas estándar locales evidencia la importancia de la recolección de información antropométrica en el seguimiento y control del crecimiento infantil en una población determinada. El trabajo adicionalmente compara los estándares de crecimiento nacionales e internacionales con las curvas locales estimadas logrando con esto contrastar las características de la información disponible con referentes de crecimiento existentes. En la comparación de las curvas son aplicados modelos de regresión GAMLSS siguiendo lineamientos internacionales mientras que la metodología de estimación propuesta combina modelamiento longitudinal del crecimiento infantil junto a análisis de datos funcionales escasos y métodos de regresión cuantilica.

Además de ser una herramienta usada en pediatría para la evaluación de potenciales problemas de salud en niños y adolescentes, las curvas de crecimiento permiten una comparación de que tan similares o atípicas son las observaciones antropométricas respecto a un grupo determinado (Cole, 2012). A partir de las mediciones de consulta única el presente trabajo evidencia una diferencia general en talla para los pacientes del conjunto de datos estudiado, siendo estos en general más bajos respecto a los referentes nacionales e internacionales con los cuales son comparados. No se evidencia un sesgo de selección en la talla de las madres para los niños en el estudio respecto al patrón nacional de crecimiento ni tampoco se explica la diferencia observada por la tendencia secular de crecimiento.

De manera concordante a otros estudios de crecimiento infantil (Habicht et al., 1974; Natale & Rajagopalan, 2014; Silva et al., 2010; Orden & Apezteguía, 2016), este trabajo evidencia diferencias importantes respecto a los estándares internacionales, los cuales priorizan factores socioeconómicos y nutricionales respecto a la diversidad genética, étnica y cultural. Este resultado esta además alineado con hipótesis de dinámicas de crecimiento infantil diferenciales por altitud geográfica, en las cuales se espera en ciudades de mayor altitud como Bogotá una menor talla respecto a lugares de menor altitud geográfica, aún para poblaciones privilegiadas (Stinson, 1982). Sin embargo, la información recolectada no permite inferir que este comportamiento sea característico en la población infantil de la ciudad.

Es conocido que tanto la regularidad como la recurrencia de las visitas a pediatría obedece principalmente a cambios esperados en el crecimiento, por esta razón en los primeros años de vida se hace en general un seguimiento frecuente mientras que en edades mayores el seguimiento es más esporádico. Naturalmente, la caracterización del patrón de crecimiento en edades de rápido crecimiento requiere observaciones regulares en cortos intervalos de tiempo, por esta razón resulta más conveniente el seguimiento longitudinal

para estas edades en la construcción de las curvas de crecimiento. La revisión de literatura permite concluir que esta decisión práctica en general no se ve reflejada en las metodologías aplicadas en la estimación de las curvas de talla, que desde un punto de vista técnico son restringidas a una muestra aleatoria de observaciones. En las curvas colombianas elaboradas por Durán et al. (2016) no se discute este aspecto en el modelamiento. En las curvas de la OMS se construyen intervalos de tiempo para el diagnóstico del modelo de forma tal que eviten observaciones repetidas del mismo niño (WHO, 2006b) y omiten pruebas de bondad de ajuste que impliquen intervalos de edad con observaciones repetidas, como la implementada en Pan & Cole (2004). Las curvas de la OMS, sin embargo, al contemplar métodos que incorporan la estructura de correlación de las medidas repetidas, indican que esto tiene poco efecto en la estimación de los percentiles distribucionales (Borghi et al., 2006).

El análisis funcional de datos escasos permite caracterizar el proceso aleatorio para la variable de interés de manera análoga al caso multivariado mediante las propiedades de primer y segundo orden. Al complementar el estudio funcional con el análisis en componentes principales son esclarecidas características funcionales inherentes del conjunto de datos. Se evidenció que la principal fuente de variación en los datos corresponde a un incremento o decremento general de talla mientras que la segunda fuente de variación esta relacionada principalmente con la velocidad de crecimiento infantil. Esto para maduradores tempranos, promedio y tardíos.

Los contornos multivariados estimados con base en los puntajes muestrales sobre los componentes principales permiten detectar potenciales problemas de crecimiento al estimar el comportamiento funcional en el dominio de interés. Esto lo diferencia del acercamiento retrospectivo inherente de las curvas de crecimiento usuales, en las cuales no se hace uso de la información longitudinal contenida en la muestra. Los parámetros estimados permiten además la reconstrucción de nuevas observaciones para contrastar el patrón de crecimiento con la población estudiada en el presente trabajo. Se destaca, sin embargo, que los supuestos distribucionales en la estimación de los puntajes pueden ser restrictivos para variables diferentes a la talla infantil.

# Trabajo futuro

Respecto a la disponibilidad de información, se destacan las siguientes posibilidades de trabajo futuro:

- Existe una mayor cantidad de información disponible que cumple con los criterios de inclusión definidos en el presente estudio que una vez depurada permitirá actualizar los resultados presentados. Se cuenta además con un conjunto de datos de niñas con el cual se puede replicar el trabajo y evaluar las diferencias existentes del crecimiento desde un punto de vista funcional.
- Una mayor especificidad de las curvas puede ser alcanzada al controlar por variables adicionales en el crecimiento, variables estáticas tales como talla materna y variables dinámicas tales como edad osea o el peso pueden afectar de manera significativa el comportamiento de los percentiles de la variable antropométrica. Esto se puede ver reflejado tanto en el modelamiento transversal como en el longitudinal en la elaboración de las curvas de crecimiento.

Desde el modelamiento funcional, se destacan tres opciones de trabajo futuro: modelamiento de datos escasos con variación en fase, monotonicidad de curvas ajustadas en el escenario escaso y modelamiento de datos de corto rango. Estas son detallados a continuación:

- Los métodos utilizados para el análisis de datos funcionales escasos, como PACE y sus variantes, no permiten restricciones en la estimación de las curvas funcionales. Para el estudio del crecimiento infantil, la variable talla presenta un comportamiento monótono creciente que no es considerado en el modelamiento. El sesgo incurrido para garantizar monotonicidad en la estimación de las curvas presenta una oportunidad de mejora importante en la reconstrucción de las trayectorias funcionales con datos escasos.
- El cambio de fase propio en el estudio de la talla infantil fue modelado previamente al análisis funcional escaso mediante el uso de un modelo longitudinal independiente. Esto es requerido para la estimación apropiada de la función media y la función de covarianza en el análisis de componentes principales, también para diferenciar el patrón de crecimiento saludable por la madurez del paciente. Un modelamiento funcional conjunto de la variable talla y del reloj biológico del paciente permitirá caracterizar las diferentes fuentes de variación en los datos escasos de manera más apropiada.

- Los datos longitudinales recolectados en el estudio presentan diferentes rangos en los cuales son observados. Es común encontrar observaciones longitudinales muy cortas, propias de sujetos que visitan al médico especialista de manera recurrente en un intervalo corto de tiempo. La utilización de datos longitudinales muy cortos, o *snippets* (Dawson & Müller, 2018), mediante metodologías funcionales escasas no permite caracterizar de manera apropiada la función de covarianza y la estimación de la función subyacente no es siempre satisfactoria. Métodos para el análisis longitudinal de datos *snippets* presentan una gran oportunidad para el análisis de la información médica como las curvas longitudinales de crecimiento infantil.

# Bibliografía


Addo, O. Y., Stein, A. D., Fall, C. H., Gigante, D. P., Guntupalli, A. M., Horta, B. L., Kuzawa, C. W., Lee, N., Norris, S. A., Prabhakaran, P. et al. (2013). Maternal height and child growth patterns, *The Journal of pediatrics* **163**(2): 549–554.

Aguilera, A. M. & Aguilera-Morillo, M. (2013). Comparative study of different b-spline approaches for functional data, *Mathematical and Computer Modelling* **58**(7-8): 1568–1579.

Anderson, C., Hafen, R., Sofrygin, O., Ryan, L. & Community, H. (2019). Comparing predictive abilities of longitudinal child growth models, *Statistics in medicine* **38**(19): 3555–3570.

Araújo, C. L., Albernaz, E., Tomasi, E. & Victora, C. G. (2004). Implementation of the who multicentre growth reference study in brazil, *Food and nutrition bulletin* **25**(1_suppl_1): S53–S59.

Borghi, E., de Onis, M., Garza, C., Van den Broeck, J., Frongillo, E. A., Grummer-Strawn, L., Van Buuren, S., Pan, H., Molinari, L., Martorell, R., Onyango, A. W., Martines, J. C., Pinol, A., Siyam, A., Victoria, C. G., Bhan, M. K., Araújo, C. L., Lartey, A., Owusu, W. B., Bhandari, N., Norum, K. R., Bjoerneboe, G. E. A., Mohamed, A. J., Dewey, K. G., Belbase, K., Chumlea, C., Cole, T., Shrimpton, R., Albernaz, E., Tomasi, E., de Cássia Fossati da Silveira, R., Nader, G., Sagoe-Moses, I., Gomez, V., Sagoe-Moses, C., Taneja, S., Rongsen, T., Chetia, J., Sharma, P., Bahl, R., Baerug, A., Tufte, E., Alasfoor, D., Prakash, N. S., Mabry, R. M., Al Rajab, H. J., Helmi, S. A., Nommsen-Rivers, L. A., Cohen, R. J. & Heinig, M. J. (2006). Construction of the World Health Organization child growth standards: Selection of methods for attained growth curves, *Statistics in Medicine* **25**(2): 247–265.

Briceño, G., Durán, P., Colón, E., Line, D., Merker, A., Abad, V., Chahín, S., Del Toro, K., Matallana, A., Llano, M., Lema, A., Soder, O., Céspedes, J. & Hagenäs, L. (2012). Protocolo del estudio para establecer estándares normativos de crecimiento de niños colombianos sanos, *Pediatría* **45**(4): 235–242.

Buuren, S. v. & Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves, *Statistics in medicine* **20**(8): 1259–1277.

Cameron, N. (2002). *Human growth and development*, Academic Press.

Cole, T. J. (1988). Fitting Smoothed Centile Curves to Reference Data, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **151**(3): 385.

Cole, T. J. (1994). Growth charts for both cross-sectional and longitudinal data, *Statistics in medicine* **13**(23-24): 2477–2492.

Cole, T. J. (2012). The development of growth references and growth charts, *Annals of Human Biology* **39**(5): 382–394.

Cole, T. J., Donaldson, M. D. & Ben-Shlomo, Y. (2010). Sitar - a useful instrument for growth curve analysis, *International journal of epidemiology* **39**(6): 1558–1566.

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood, *Statistics in Medicine* **11**(10): 1305–1319.

Dawson, M. & Müller, H.-G. (2018). Dynamic modeling of conditional quantile trajectories, with application to longitudinal snippet data, *Journal of the American Statistical Association* **113**(524): 1612–1624.

De Blanco, M. L., de Espinoza, I. I. & de Tomei, C. M. (2013). *Crecimiento y maduración física: bases para el diagnóstico y seguimiento clínico*, Editorial Médica Panamericana.

De Onis, M., Garza, C., Onyango, A. W. & Martorell, R. (2006). WHO Child Growth Standards, *Acta Paediatrica* **95**(450): 106.

De Onis, M., Garza, C., Victora, C. G., Onyango, A. W., Frongillo, E. A. & Martines, J. (2004). The who multicentre growth reference study: planning, study design, and methodology, *Food and nutrition bulletin* **25**(1_suppl_1): S15–S26.

De Onis, M., Onyango, A. W., Borghi, E., Garza, C. & Yang, H. (2006). Comparison of the World Health Organization (WHO) Child Growth Standards and the National Center for Health Statistics/WHO international growth reference: Implications for child health programmes, *Public Health Nutrition* **9**(7): 942–947.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics* **5**(3): 236–244.

Durán, P., Merker, A., Briceño, G., Colón, E., Line, D., Abad, V., Del Toro, K., Chahín, S., Matallana, A. M., Lema, A., Llano, M., Céspedes, J. & Hagenäs, L. (2016). Colombian reference growth curves for height, weight, body mass index and head circumference, *Acta Paediatrica* **105**(3).

Eilers, P. & Brian, M. (1996). Flexible smoothing with b-splines and penalties, *Statistical science* pp. 89–102.

Freeman, J., Cole, T., Chinn, S., Jones, P., White, E. & Preece, M. (1995). Cross sectional stature and weight reference curves for the uk, 1990., *Archives of disease in childhood* **73**(1): 17–24.

Grummer-Strawn, L. M., Reinold, C., Krebs, N. F. & Centers for Disease Control and Prevention (CDC) (2010). Use of World Health Organization and CDC growth charts for children aged 0-59 months in the United States., *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports* **59**(RR-9): 1–15.

Habicht, J.-P., Yarbrough, C., Martorell, R., Malina, R. & Klein, R. (1974). Height and weight standards for preschool children: how relevant are ethnic differences in growth potential?, *The Lancet* **303**(7858): 611–615.

Hamill, P. V., Drizd, T. A., Johnson, C. L., Reed, R. B. & Roche, A. F. (1977). Nchs growth curves for children birth-18 years, *Technical report*, Department of Health Education and Welfare Washington DC.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Horváth, L. & Kokoszka, P. (2012). *Inference for functional data with applications*, Springer Science & Business Media.

James, G. M., Hastie, T. J. & Sugar, C. A. (2000). Principal component models for sparse functional data, *Biometrika* **87**(3): 587–602.

Kelly, A., Winer, K. K., Kalkwarf, H., Oberfield, S. E., Lappe, J., Gilsanz, V. & Zemel, B. S. (2014). Age-based reference ranges for annual height velocity in us children, *The Journal of Clinical Endocrinology & Metabolism* **99**(6): 2104–2112.

Kelnar, C., Savage, M., Saenger, P. & Cowell, C. (2007). *Growth Disorders 2E*, CRC Press.

Kokoszka, P. & Reimherr, M. (2017). *Introduction to functional data analysis*, CRC Press.

Kuczmarski, R. J. (2002). *2000 CDC Growth Charts for the United States: methods and development*, number 246, Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Kuczmarski, R. J., Ogden, C. L., Grummer-Strawn, L. M., Flegal, K. M., Guo, S. S., Wei, R., Mei, Z., Curtin, L. R., Roche, A. F. & Johnson, C. L. (2000). Cdc growth charts: United states advance data from vital and health statistics, no. 314, *Hyattsville, MD: National Center for Health Statistics* .

Lejarraga, H., del Pino, M., Fano, V., Caino, S. & Cole, T. J. (2009). Referencias de peso y estatura desde el nacimiento hasta la madurez para niñas y niños argentinos: Incorporación de datos de la oms de 0 a 2 años, recálculo de percentilos para obtención de valores lms, *Archivos argentinos de pediatría* **107**(2): 126–133.

Lejarraga, H. & Orfila, G. (1987). Estándares de peso y estatura para niñas y niños argentinos desde el nacimiento hasta la madurez, *Arch Argent Pediatr* **85**(4): 209–22.

Leroux, A., Xiao, L., Crainiceanu, C. & Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth, *Statistics in medicine* **37**(8): 1376–1388.

Lindstrom, M. J. & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* pp. 673–687.

Madrigal, P., Dai, X. & Hadjipantelis, P. Z. (2018). Sparse functional data analysis accounts for missing information in single-cell epigenomics, *bioRxiv* p. 504365.

Marques, R., Lopez, F. A. & Braga, J. (2004). Growth of exclusively breastfed infants in the first 6 months of life, *J Pediatr (Rio J)* **80**(2): 99–105.

Natale, V. & Rajagopalan, A. (2014). Worldwide variation in human growth and the world health organization growth standards: a systematic review, *BMJ open* **4**(1): e003735.

Onis, M. d., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C. & Siekmann, J. (2007). Development of a who growth reference for school-aged children and adolescents, *Bulletin of the World health Organization* **85**: 660–667.

Onyango, A. W., de Onis, M., Caroli, M., Shah, U., Sguassero, Y., Redondo, N. & Carroli, B. (2007). Field-testing the who child growth standards in four countries, *The Journal of nutrition* **137**(1): 149–152.

Orden, A. B. & Apezteguía, M. C. (2016). Weight and height centiles of argentinian children and adolescents: a comparison with who and national growth references, *Annals of human biology* **43**(1): 9–17.

Pan, H. & Cole, T. (2004). A comparison of goodness of fit tests for age-related reference ranges, *Statistics in medicine* **23**(11): 1749–1765.

Pinheiro, J. & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*, Springer Science & Business Media.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*, STATA press.

Ramsay, J. & Silverman, B. (2005). *Functional Data Analysis*, Springer Science.

Rice, J. A. & Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics* **57**(1): 253–259.

Rigby, R. A. & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution, *Statistics in medicine* **23**(19): 3053–3076.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3): 507–554.

Rigby, R. A. & Stasinopoulos, D. M. (2006). Using the box-cox t distribution in gamlss to model skewness and kurtosis, *Statistical Modelling* **6**(3): 209–229.

Rousseeuw, P. J., Ruts, I. & Tukey, J. W. (1999). The bagplot: a bivariate boxplot, *The American Statistician* **53**(4): 382–387.

Royston, P. & Wright, E. (2000). Goodness-of-fit statistics for age-specific reference intervals, *Statistics in medicine* **19**(21): 2943–2962.

Silva, D. A. S., Pelegrini, A., Petroski, E. L. & Gaya, A. C. A. (2010). Comparison between the growth of brazilian children and adolescents and the reference growth charts: data from a brazilian project, *Jornal de pediatria* **86**(2): 115–120.

Simpkin, A. J., Sayers, A., Gilthorpe, M. S., Heron, J. & Tilling, K. (2017). Modelling height in adolescence: a comparison of methods for estimating the age at peak height velocity, *Annals of human biology* **44**(8): 715–722.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. & De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*, CRC Press.

Stinson, S. (1982). The effect of high altitude on the growth of children of high socioeconomic status in bolivia, *American Journal of Physical Anthropology* **59**(1): 61–71.

Sun, Y. & Genton, M. G. (2011). Functional boxplots, *Journal of Computational and Graphical Statistics* **20**(2): 316–334.

Tanner, J. M. & Whitehouse, R. H. (1976). Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty., *Archives of disease in childhood* **51**(3): 170–179.

Tanner, J. M., Whitehouse, R. & Takaishi, M. (1966). Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. i., *Archives of disease in childhood* **41**(219): 454.

Tarupi, W., Lepage, Y., Felix, M. L., Monnier, C., Hauspie, R., Roelants, M., Hidalgo, R. & Vercauteren, M. (2020). Growth references for weight, height, and body mass index for ecuadorian children and adolescents aged 5-19 years, *Arch Argent Pediatr* **118**(2): 117–124.

Tuddenham, R. D. (1954). Physical growth of california boys and girls from birth to eighteen years, *University of California publications in child development* **1**: 183–364.

Tukey, J. W. (1975). Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, Vol. 2, pp. 523–531.

Van Wieringen, J. (1978). Secular growth changes, *Human growth*, Springer, pp. 445–473.

Wei, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts, *Journal of the American Statistical Association* **103**(481): 397–409.

Wei, Y., He, X. et al. (2006). Conditional growth charts, *The Annals of Statistics* **34**(5): 2069–2097.

WHO (2006a). Breastfeeding in the who multicentre growth reference study., *Acta paediatrica (Oslo, Norway: 1992). Supplement* **450**: 16.

WHO (2006b). Who child growth standards based on length/height, weight and age., *Acta paediatrica (Oslo, Norway: 1992). Supplement* **450**: 76.

WHO (2006c). Who child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
**URL:** *https://ggplot2.tidyverse.org*

Wright, C. M., Williams, A. F., Elliman, D., Bedford, H., Birks, E., Butler, G., Sachs, M., Moy, R. J. & Cole, T. J. (2010). Using the new uk-who growth charts, *Bmj* **340**: c1140.

Wright, E. M. & Royston, P. (1997). A Comparison of Statistical Methods for Age-related Reference Intervals, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **160**(1): 47–69.

Xiao, L., Li, C., Checkley, W. & Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data, *Statistics and computing* **28**(3): 511–522.

Xiao, L., Zipunnikov, V., Ruppert, D. & Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data, *Statistics and computing* **26**(1-2): 409–421.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A. & Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate, *Biometrics* **59**(3): 676–685.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association* **100**(470): 577–590.

Zhang, W., Wei, Y. et al. (2015). Regression based principal component analysis for sparse functional data with applications to screening growth paths, *The Annals of Applied Statistics* **9**(2): 597–620.