Modelo BNP para datos discretos

con aplicación al rendimiento de clubes deportivos

Cristian Capetillo Constela

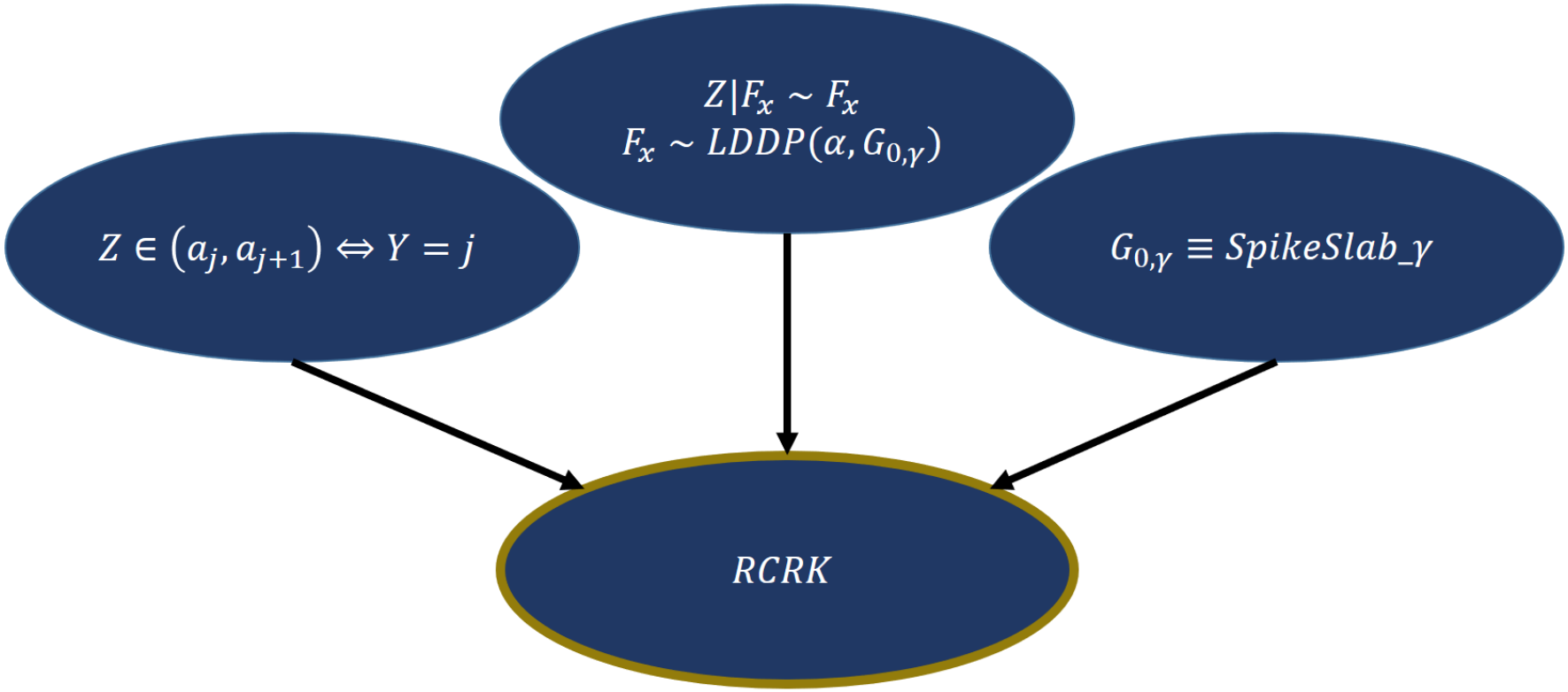
Pontificia Universidad Católica de Chile
Facultad de matemática
Departamento de Estadística

11 de Noviembre, 2024

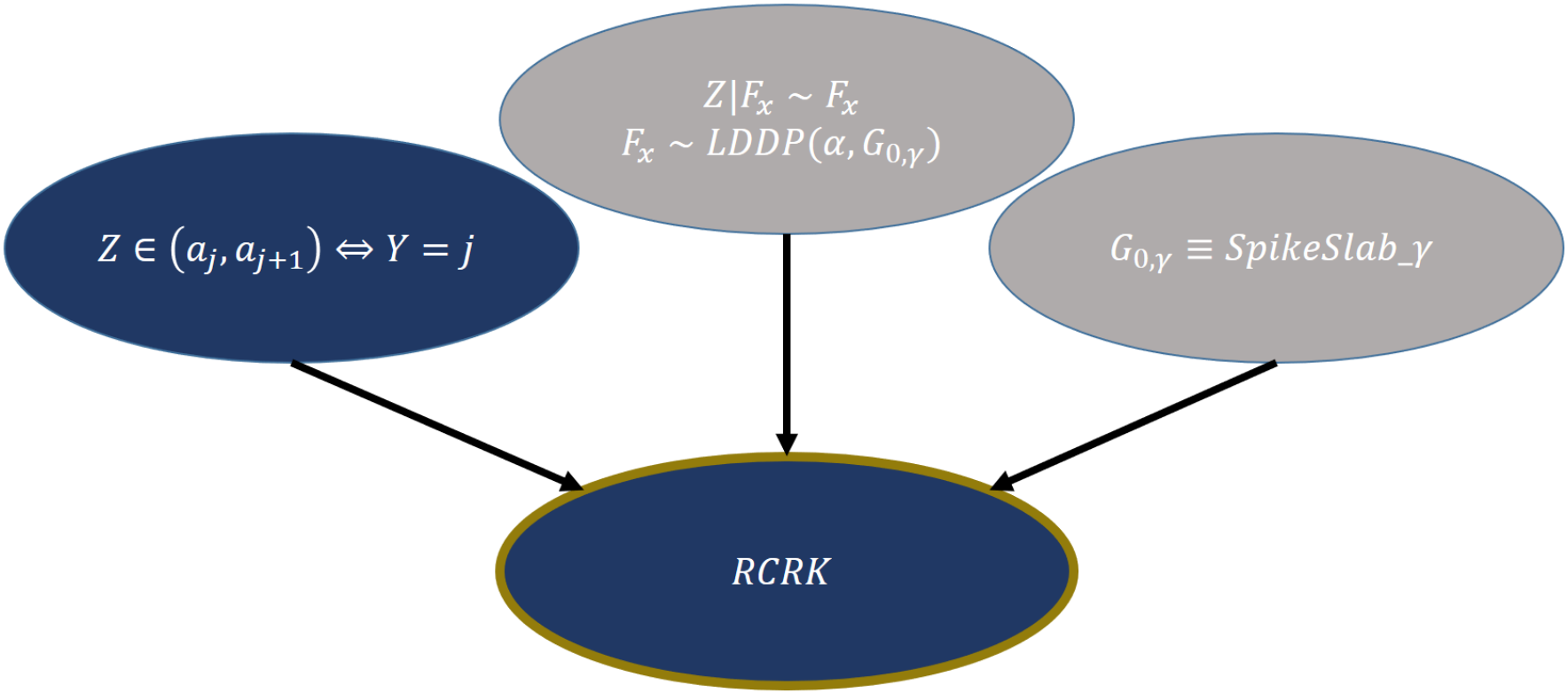
Contenidos

Contenidos

Rounded Continuous Regression Kernel Model



Rounded Continuous Regression Kernel Model



Continuidad latente para observaciones discretas

Sea Y una variable aleatoria en $\mathbb{N}$ y Z una variable aleatoria en $\mathbb{R}$. Suponemos que,

$$Z \in (a_j, a_{j+1}) \iff Y = j, \ \forall j = 0, 1, ...$$

En consecuencia,

$$P(Y = j) = P(Z \in (a_j, a_{j+1})) = F(a_{j+1}) - F(a_j).$$

Continuidad latente para observaciones discretas

Sea Y una variable aleatoria en $\mathbb{N}$ y Z una variable aleatoria en $\mathbb{R}$. Suponemos que,

$$Z \in (a_j, a_{j+1}) \iff Y = j, \ \forall j = 0, 1, ...$$

En consecuencia,

$$P(Y = j) = P(Z \in (a_j, a_{j+1})) = F(a_{j+1}) - F(a_j).$$

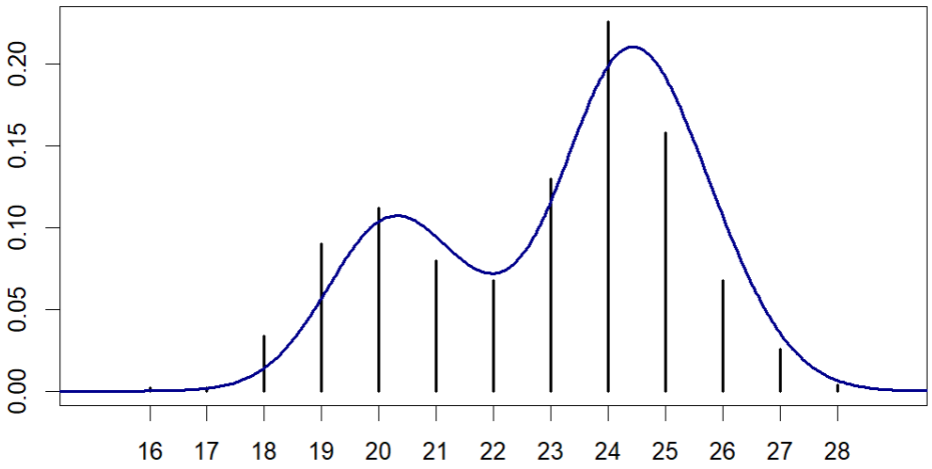


Figura 1: Cuantía empírica de Y y función de densidad de Z.

Continuidad latente para observaciones discretas

Sea Y una variable aleatoria en $\mathbb{N}$ y Z una variable aleatoria en $\mathbb{R}$. Suponemos que,

$$Z \in (a_j, a_{j+1}) \iff Y = j, \ \forall j = 0, 1, ...$$

En consecuencia,

$$P(Y = j) = P(Z \in (a_j, a_{j+1})) = F(a_{j+1}) - F(a_j).$$

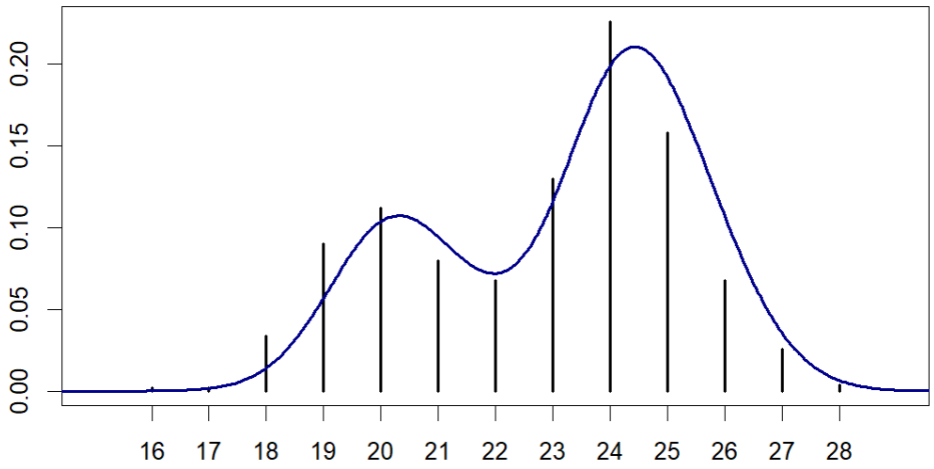


Figura 1: Cuantía empírica de Y y función de densidad de Z.

$$Y|F \sim \left\{ F(a_{j+1}) - F(a_j) \right\}_{j=0}^{\infty}$$

Continuidad latente para observaciones discretas

A considerar:

Continuidad latente para observaciones discretas

A considerar:

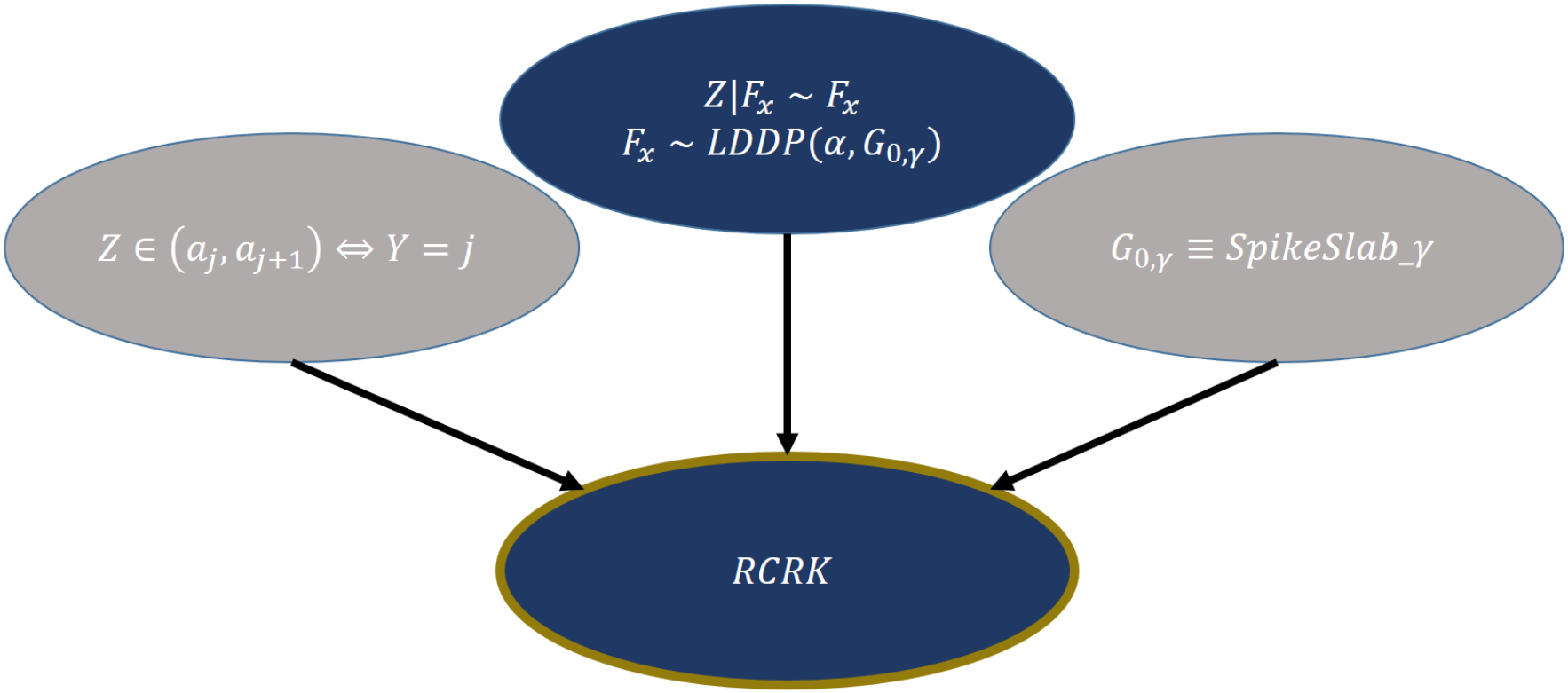
1. La distribución F es desconocida. Debemos darle una **estructura**.

Continuidad latente para observaciones discretas

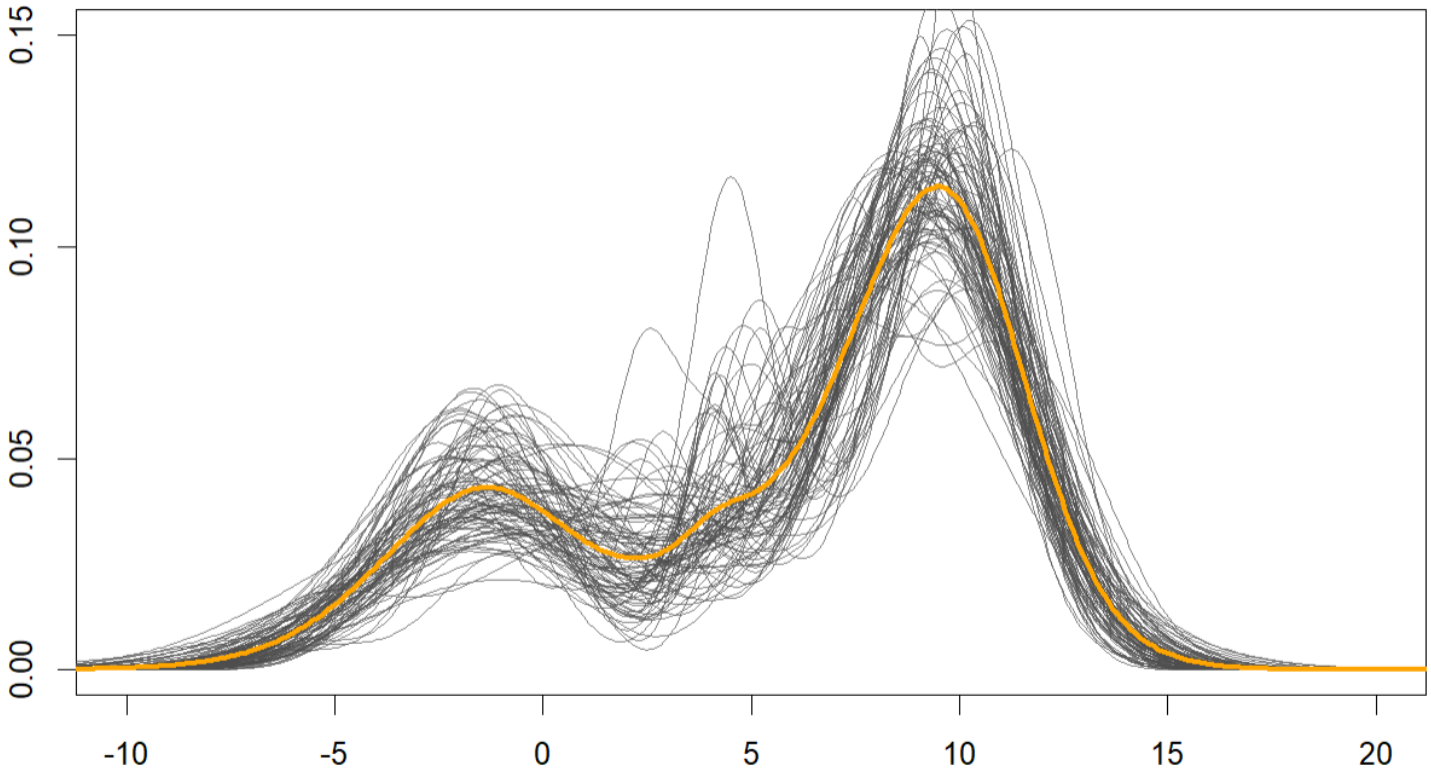
A considerar:

1. La distribución F es desconocida. Debemos darle una **estructura**.
2. Dado que nuestro objetivo es relacionar Y con covariables $\boldsymbol{x}$, F es en realidad F_x.

Rounded Continuous Regression Kernel Model



Estructura de F_x



Linear Dependent Dirichlet Process

Una medida de probabilidad aleatoria G_x se dice que sigue un Linear Dependent Dirichlet Process (LDDP) con parámetros $\alpha > 0$ y G_0 medida de probabilidad sobre $\mathbb{R}^p \times \mathbb{R}_+$ si

$$G_x(\cdot) = \sum_{h=1}^{\infty} w_h \mathcal{N}(\cdot | \boldsymbol{x}^T \boldsymbol{\beta}_h, \tau_h^{-1}),$$

donde $w_h = v_h \prod_{l<h}^{\infty}(1 - v_h)$ para $h = 2, 3, ...,$ y $w_1 = v_1$, $v_h \overset{iid}{\sim} \mathcal{B}eta(1, \alpha)$, y $(\beta_h, \tau_h) \overset{iid}{\sim} G_0$.

Linear Dependent Dirichlet Process

Una medida de probabilidad aleatoria G_x se dice que sigue un Linear Dependent Dirichlet Process (LDDP) con parámetros $\alpha > 0$ y G_0 medida de probabilidad sobre $\mathbb{R}^p \times \mathbb{R}_+$ si

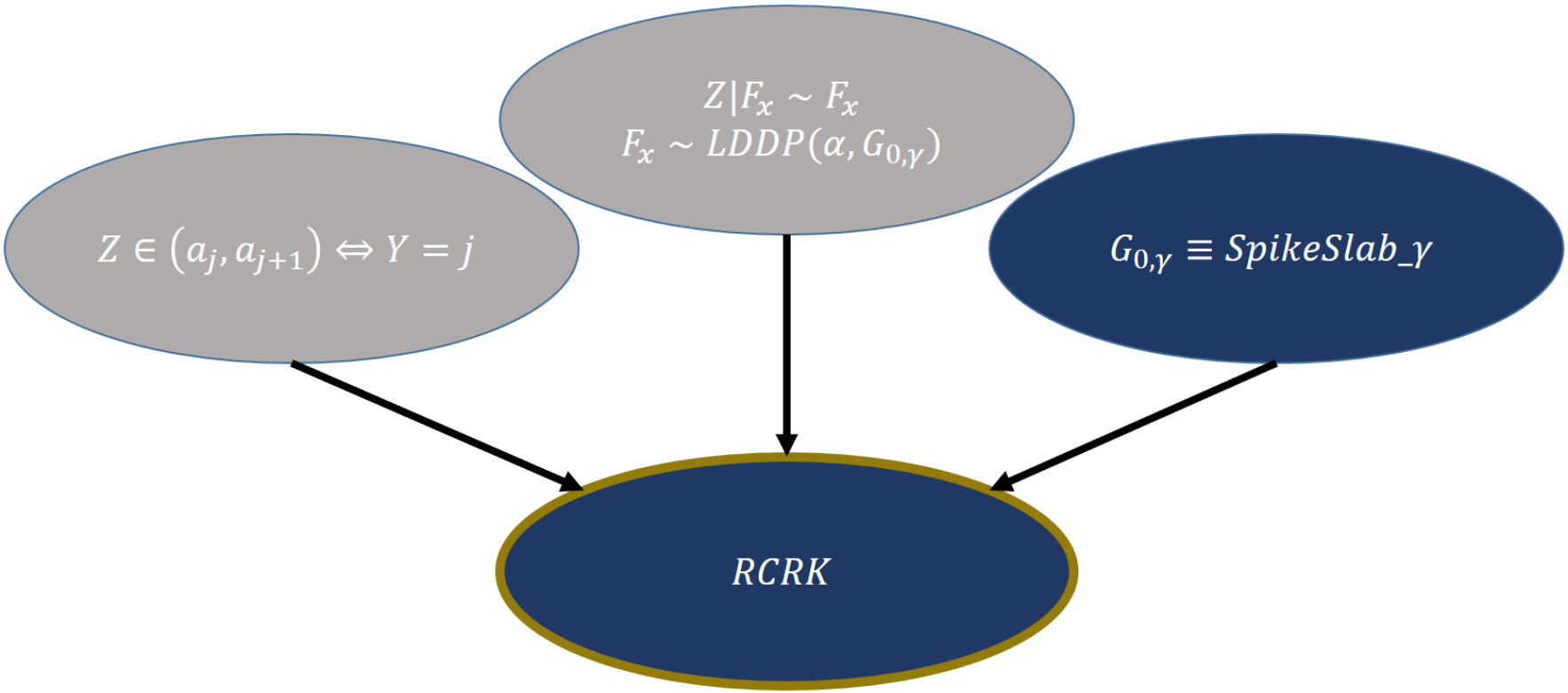
$$G_x(\cdot) = \sum_{h=1}^{\infty} w_h \mathcal{N}(\cdot | \boldsymbol{x}^T \boldsymbol{\beta}_h, \tau_h^{-1}),$$

donde $w_h = v_h \prod_{l<h}^{\infty}(1 - v_h)$ para $h = 2, 3, ...,$ y $w_1 = v_1$, $v_h \overset{iid}{\sim} \mathcal{B}eta(1, \alpha)$, y $(\beta_h, \tau_h) \overset{iid}{\sim} G_0$.

Por ejemplo, utilizando la distribución $Normal\text{-}Gamma$ como medida base, se tiene que

$$\beta_h | \tau_h \overset{iid}{\sim} \mathcal{N}_p(\mu, (\kappa\tau_h)^{-1}),$$
$$\tau_h \overset{iid}{\sim} \mathcal{G}a(a_\tau, b_\tau).$$

Rounded Continuous Regression Kernel Model



Metodología Spike-and-Slab

Para la selección de covariables, la metodología spike-and-slab consiste en **introducir una variable aleatoria** γ_j para cada coeficiente de regresión, de manera que

$$\gamma_j = \begin{cases} 1 \text{ si la covariable } j \text{ es parte del modelo,} \\ 0 \text{ si no.} \end{cases}$$

Metodología Spike-and-Slab

Para la selección de covariables, la metodología spike-and-slab consiste en **introducir una variable aleatoria** γ_j para cada coeficiente de regresión, de manera que

$$\gamma_j = \begin{cases} 1 \text{ si la covariable } j \text{ es parte del modelo,} \\ 0 \text{ si no.} \end{cases}$$

Luego, la priori Spike-and-Slab para (β, τ), denotada por $SpikeSlab_\gamma$, viene dada por

$$\beta|\tau,\gamma \sim \mathcal{N}(\beta_0|0,(\kappa\tau)^{-1}) \prod_{j=1}^{p} (\gamma_j \mathcal{N}(\beta_j|0,(\kappa\tau)^{-1}) + (1-\gamma_j)\delta_0(\beta_j)),$$

$$\tau \sim \mathcal{G}a(a_\tau, b_\tau),$$

con κ, a_τ y b_τ hiperparámetros conocidos.

Metodología Spike-and-Slab

Para la selección de covariables, la metodología spike-and-slab consiste en **introducir una variable aleatoria** γ_j para cada coeficiente de regresión, de manera que

$$\gamma_j = \begin{cases} 1 \text{ si la covariable } j \text{ es parte del modelo,} \\ 0 \text{ si no.} \end{cases}$$

Luego, la priori Spike-and-Slab para (β, τ), denotada por $SpikeSlab_\gamma$, viene dada por

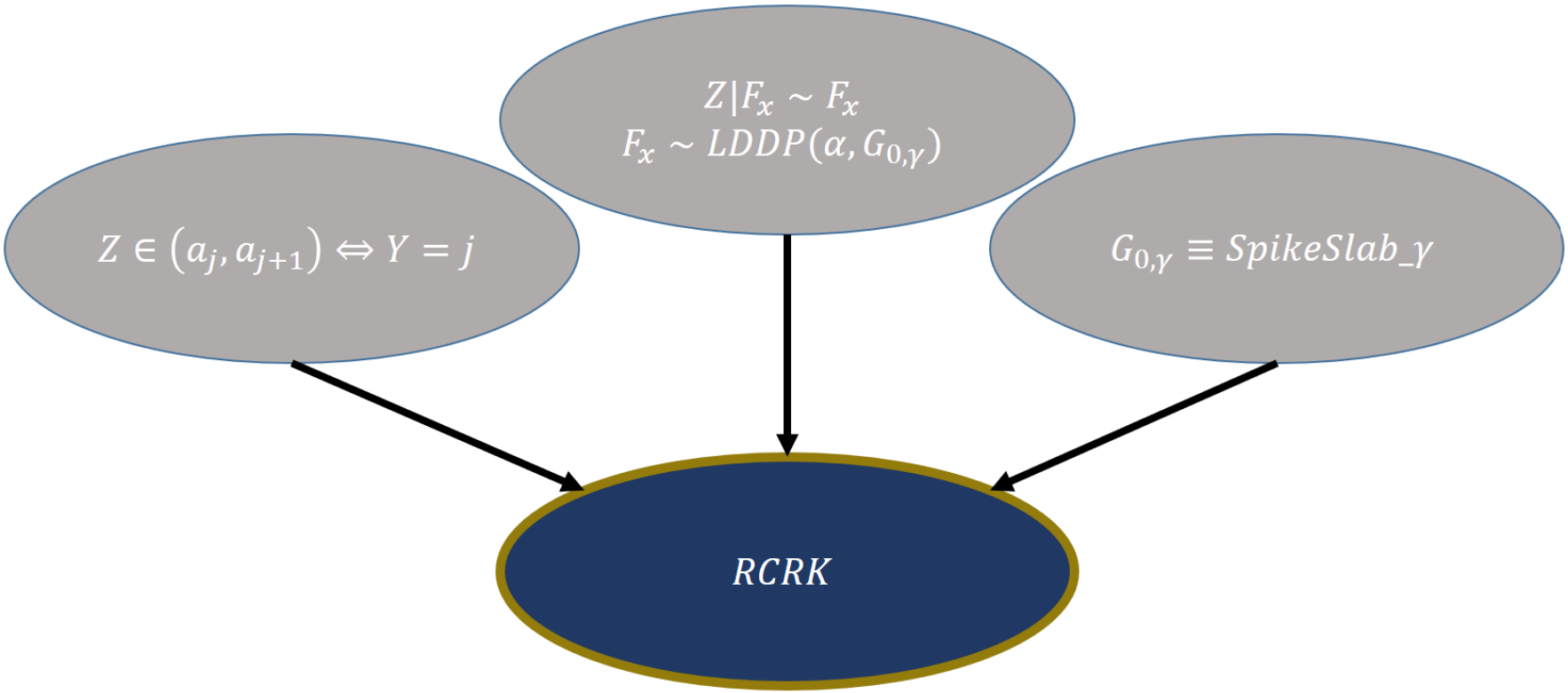
$$\beta|\tau,\gamma \sim \mathcal{N}(\beta_0|0,(\kappa\tau)^{-1}) \prod_{j=1}^{p} (\gamma_j \mathcal{N}(\beta_j|0,(\kappa\tau)^{-1}) + (1-\gamma_j)\delta_0(\beta_j)),$$

$$\tau \sim \mathcal{G}a(a_\tau, b_\tau),$$

con κ, a_τ y b_τ hiperparámetros conocidos.

Una priori para el vector γ puede ser la distribución Binomial, Beta-Binomial, o la que denominamos **priori de Womack** (Womack et al., 2015).

Rounded Continuous Regression Kernel Model



Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

$$Y_i | x_i, F_{x_i} \overset{ind}{\sim} \left\{ F_{x_i}(a_{j+1}) - F_{x_i}(a_j) \right\}_{j=0}^{\infty},$$

Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

$$Y_i | x_i, F_{x_i} \overset{ind}{\sim} \left\{ F_{x_i}(a_{j+1}) - F_{x_i}(a_j) \right\}_{j=0}^{\infty},$$

$$F_{x_i} \sim LDDP(\alpha, G_{0,\gamma}),$$

Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

$$Y_i | x_i, F_{x_i} \overset{ind}{\sim} \left\{ F_{x_i}(a_{j+1}) - F_{x_i}(a_j) \right\}_{j=0}^{\infty},$$

$$F_{x_i} \sim LDDP(\alpha, G_{0,\gamma}),$$

$$G_{0,\gamma} \equiv SpikeSlab_{\gamma},$$

Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

$$Y_i | x_i, F_{x_i} \overset{ind}{\sim} \left\{ F_{x_i}(a_{j+1}) - F_{x_i}(a_j) \right\}_{j=0}^{\infty},$$

$$F_{x_i} \sim LDDP(\alpha, G_{0,\gamma}),$$

$$G_{0,\gamma} \equiv SpikeSlab_{\gamma},$$

$$\gamma \sim Womack(\rho),$$

Rounded Continuous Regression Kernel Model

El **modelo RCRK** para la muestra $Y_1, ..., Y_n$ está dado por

$$Y_i | x_i, F_{x_i} \overset{ind}{\sim} \left\{ F_{x_i}(a_{j+1}) - F_{x_i}(a_j) \right\}_{j=0}^{\infty},$$

$$F_{x_i} \sim LDDP(\alpha, G_{0,\gamma}),$$

$$G_{0,\gamma} \equiv SpikeSlab_{\gamma},$$

$$\gamma \sim Womack(\rho),$$

$$\alpha \sim \mathcal{G}a(a_\alpha, b_\alpha).$$

Esquema MCMC

Realizando una **aumentación de datos**, el modelo RCRK puede ser reescrito jerárquicamente como

$$
\begin{aligned}
Y_i | Z_i \in (a_j, a_{j+1}) &\overset{ind}{\sim} \delta_j, \ \forall j, \\
Z_i | x_i, F_{x_i,\gamma}, S_i &\sim \mathcal{N}(x_i^T \beta_{S_i,\gamma}, \tau_{S_i}^{-1}), \\
S_i | w &\sim \{w_h\}_{h=1}^{\infty}, \\
v_h &\sim \mathcal{B}eta(1, \alpha), \\
(\beta_h, \tau_h) &\sim G_{0,\gamma}, \\
\gamma &\sim Womack(\rho), \\
\alpha &\sim \mathcal{G}a(a_\alpha, b_\alpha).
\end{aligned}
$$

Ishwaran y James (2001) propusieron truncar la suma hasta un número H lo suficientemente grande para una buena aproximación.

Contenidos

Estudio de simulación preliminar

Evaluamos el modelo a datos sintéticos. En particular, se generan 5 escenarios: N-cat, Pois-cat, NB-cat, ZIP-cat y ZINB-cat.

Estudio de simulación preliminar

Evaluamos el modelo a datos sintéticos. En particular, se generan 5 escenarios: N-cat, Pois-cat, NB-cat, ZIP-cat y ZINB-cat.

Los hiperparámetros escogidos son $a_\tau = b_\tau = 1$, $\kappa = 0.001$, $\rho = 1$, $a_\alpha = b_\alpha = 2$. Además, $H = 30$.

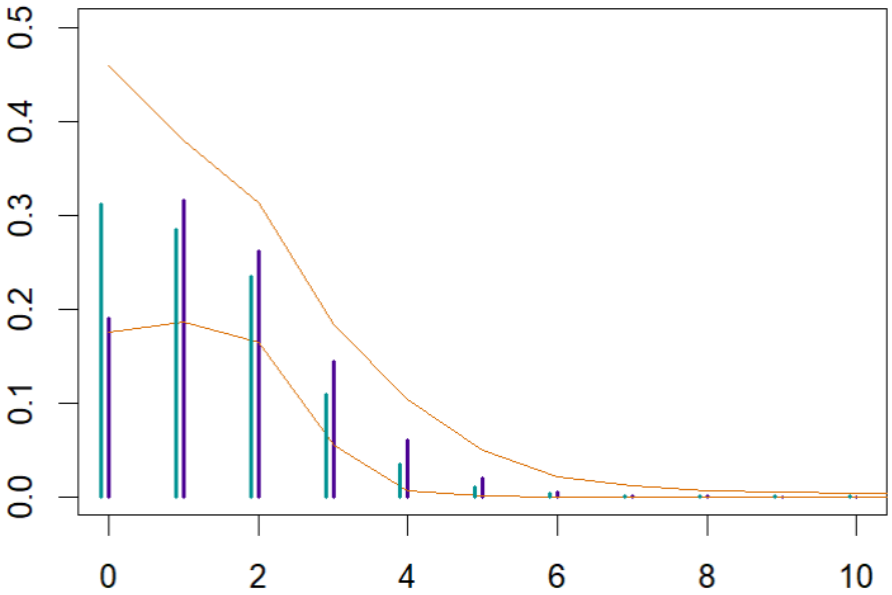
Estudio de simulación preliminar

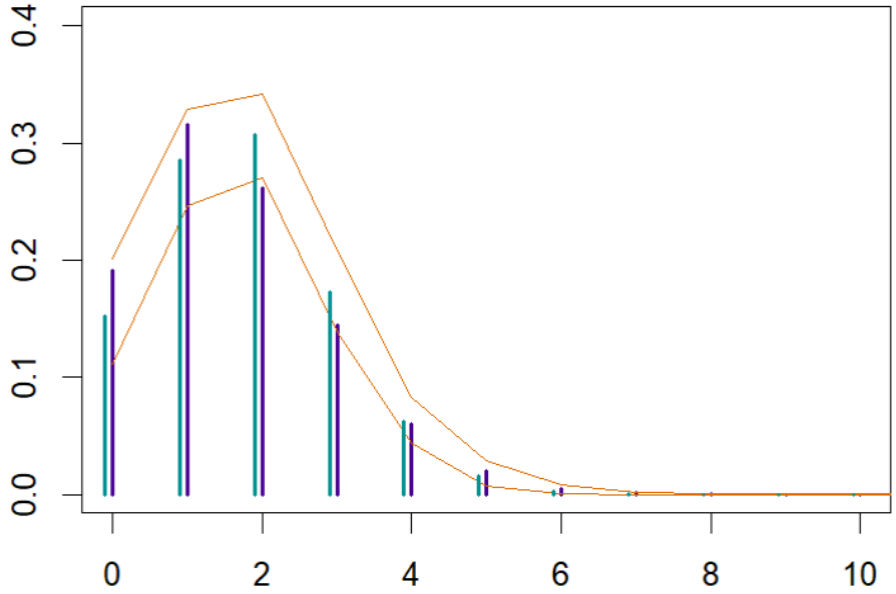
Evaluamos el modelo a datos sintéticos. En particular, se generan 5 escenarios: N-cat, Pois-cat, NB-cat, ZIP-cat y ZINB-cat.

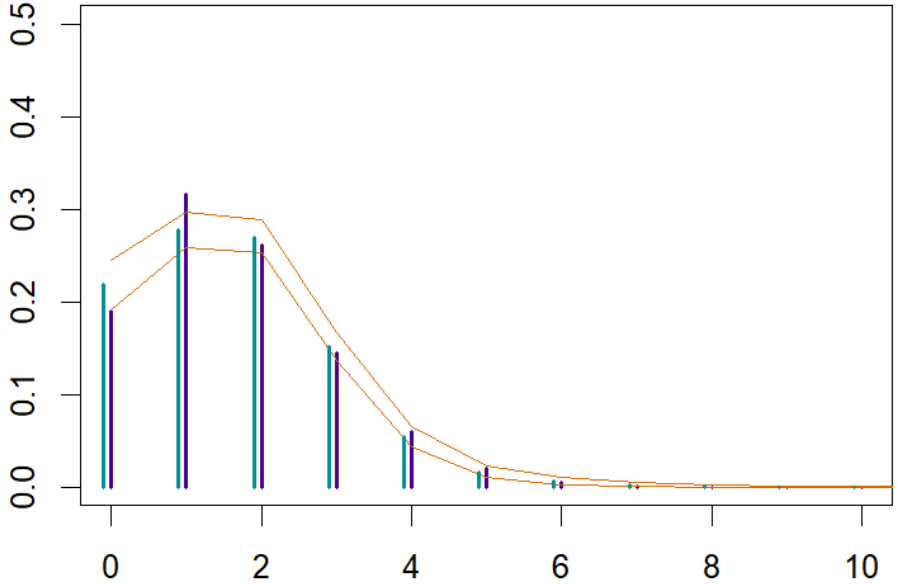
Los hiperparámetros escogidos son $a_\tau = b_\tau = 1$, $\kappa = 0.001$, $\rho = 1$, $a_\alpha = b_\alpha = 2$. Además, $H = 30$.

En cuanto a la simulación MCMC, el número de iteraciones es de 20000, con un periodo de quemado de 4000 y un submuestreo cada 10 iteraciones.

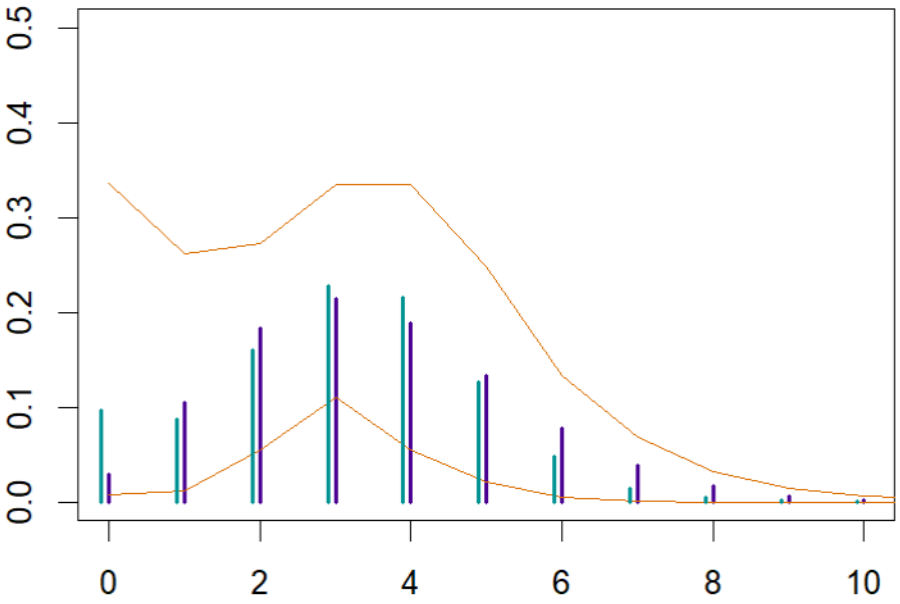
Resultados: Pois-cat

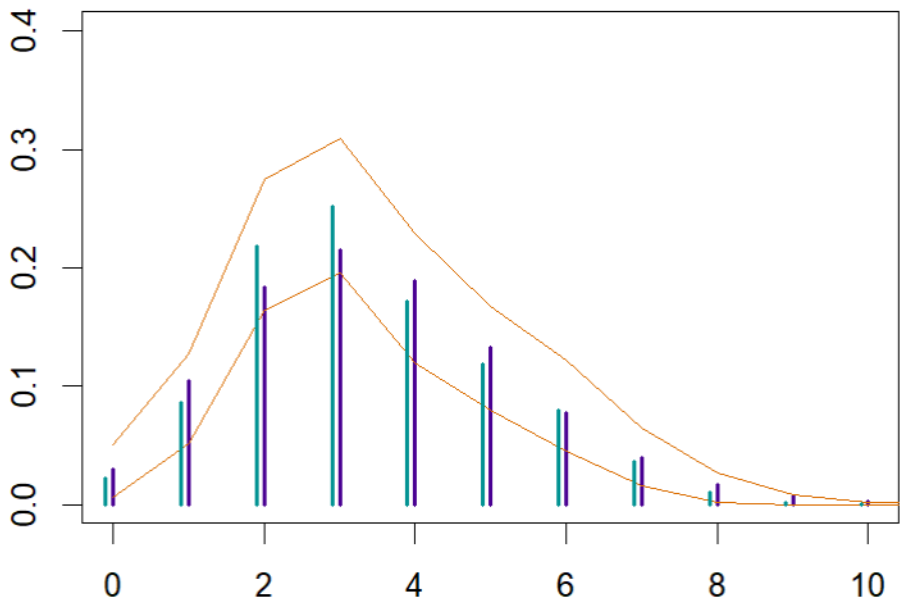


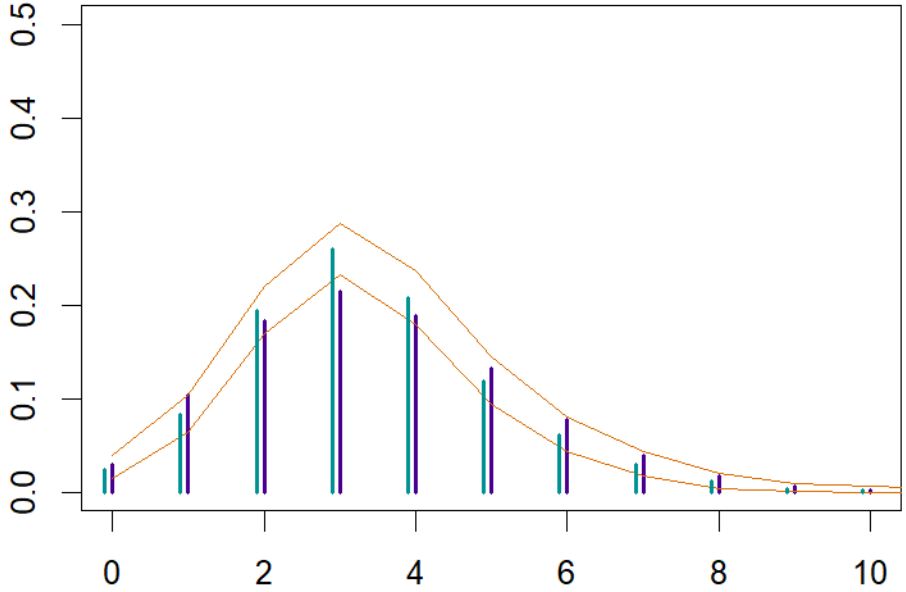




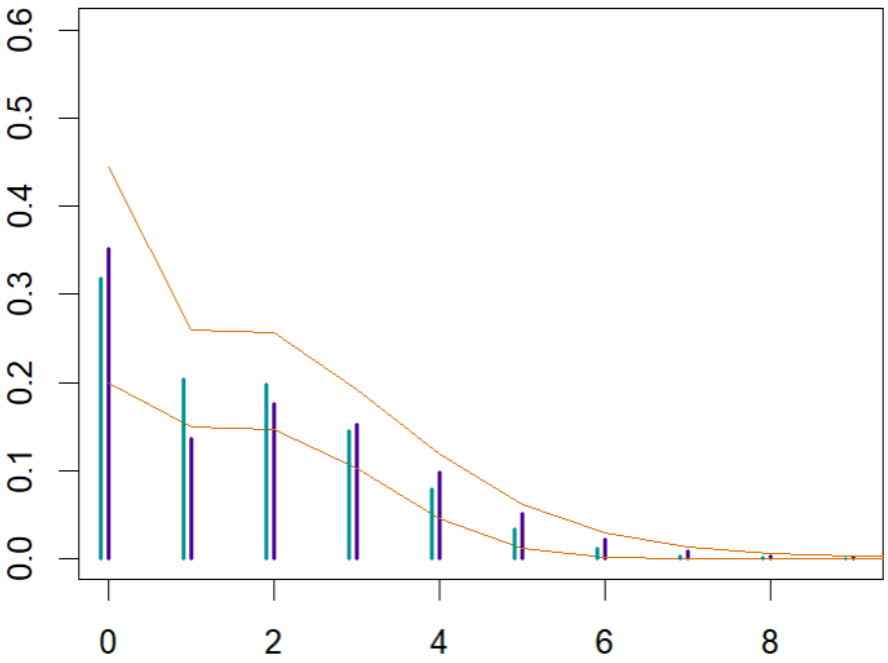
Resultados: Pois-cat

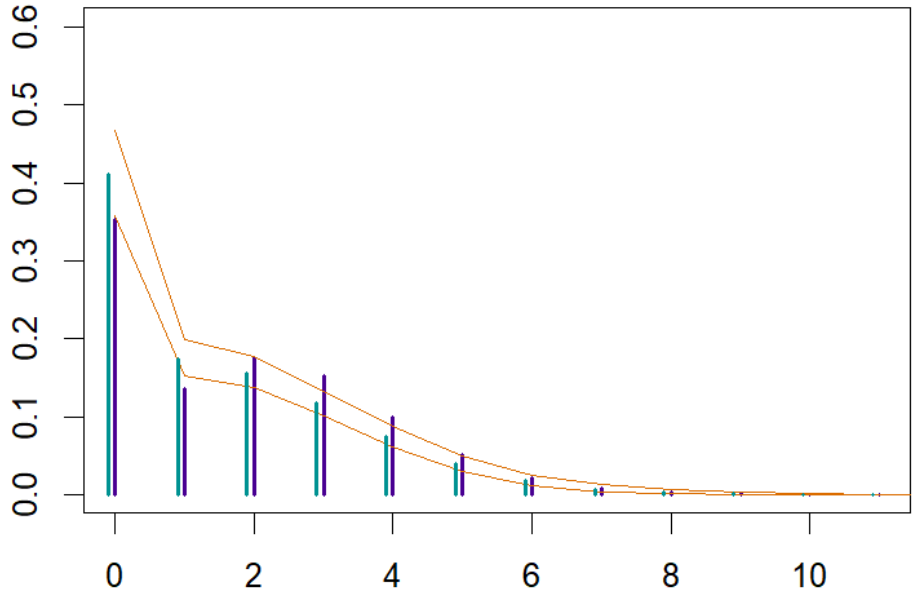


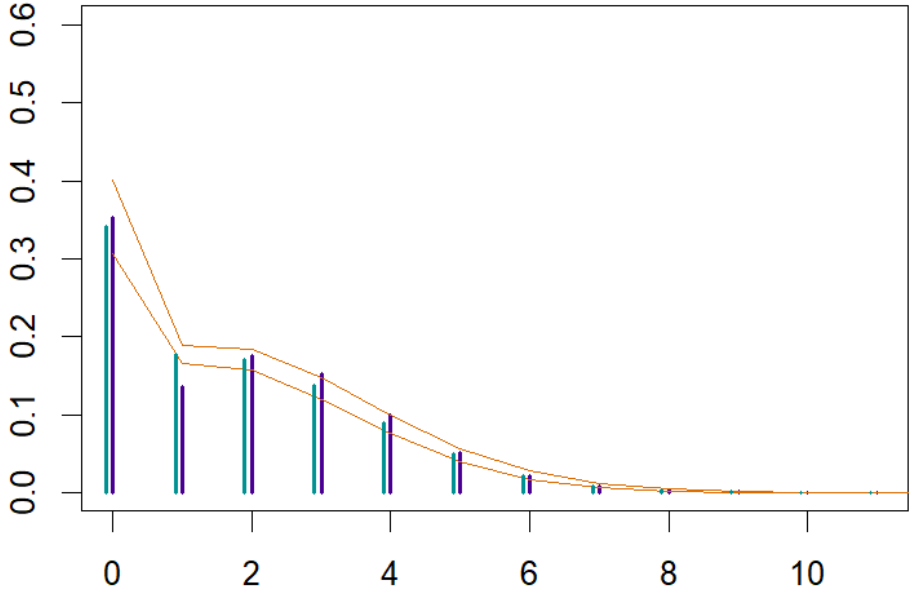




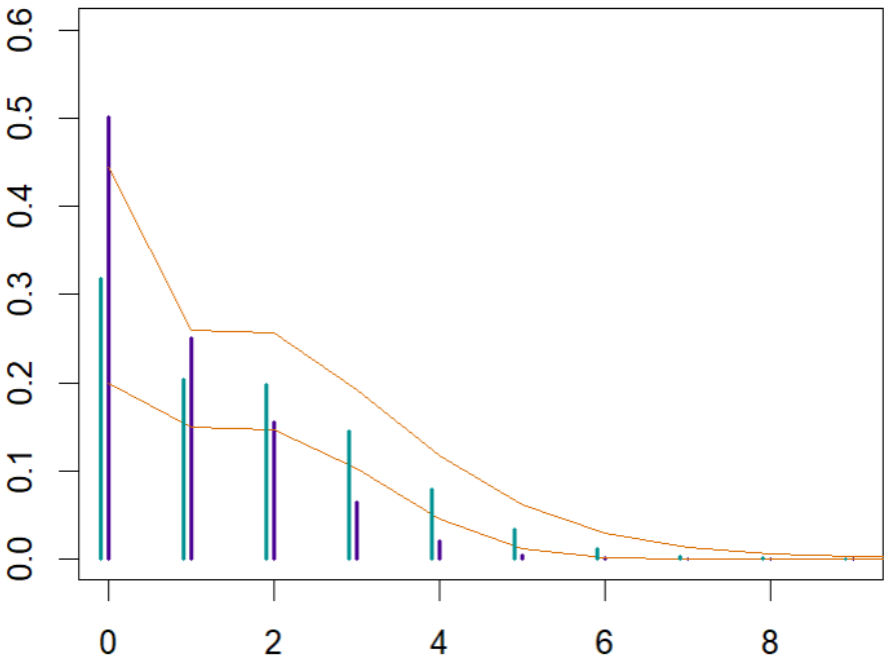
Resultados: ZIP-cat

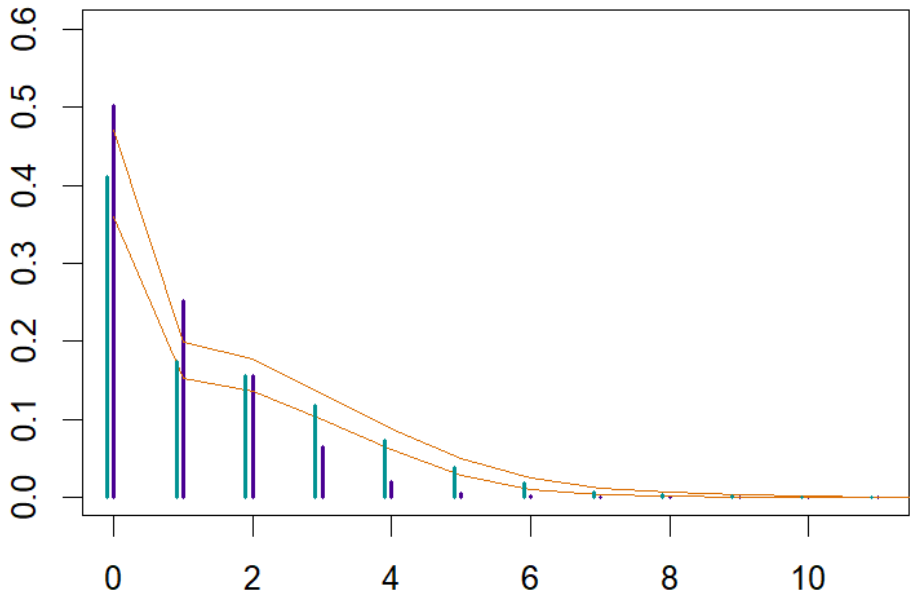


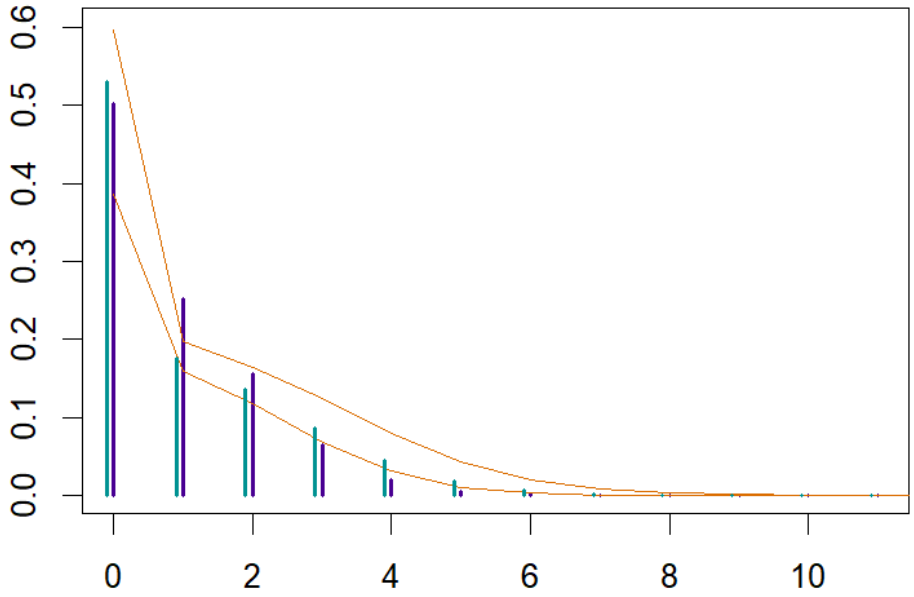




Resultados: ZIP-cat







Contenidos

Aplicación al rendimiento deportivo de clubes

Año	Liga	Ganados	Empatados	Perdidos	GF	GC	PorcVictorias	Champions	Europa	Copa
2023	7	14	11	13	44	40	36.84	Octavos de final	No participa	Segunda ronda
2024	3	20	8	10	53	41	52.63	No participa	Campeón	Cuartos de final
2025	4	20	10	8	57	35	52.63	Final	No participa	Campeón
2026	6	17	9	12	49	31	44.74	Cuartos de final	No participa	Campeón
2027	5	17	14	7	50	32	44.74	No participa	Campeón	Cuartos de final
2028	3	21	9	8	55	28	55.26	No participa	Primera ronda	Campeón
2029	7	16	12	10	48	40	42.11	Cuartos de final	No participa	Segunda ronda
2030	3	23	7	8	59	31	60.53	No participa	Primera ronda	Semifinal
2031	1	23	8	7	56	28	60.53	Octavos de final	No participa	Segunda ronda
2032	6	17	10	11	45	37	44.74	Semifinal	No participa	Segunda ronda

Figura 2: 10 observaciones de un total de 24 sobre el rendimiento de un club de fútbol ficticio temporada a temporada (2020-2043).

Resultados: Posición $\sim$ GF, GC, %Victorias

GF, GC, %V	GF, GC	GF, %V	GC, %V	GF	GC	%V	---
0	0	0.0069	0.0081	0	0	0.985	0

Resultados: Posición $\sim$ GF, GC, %Victorias

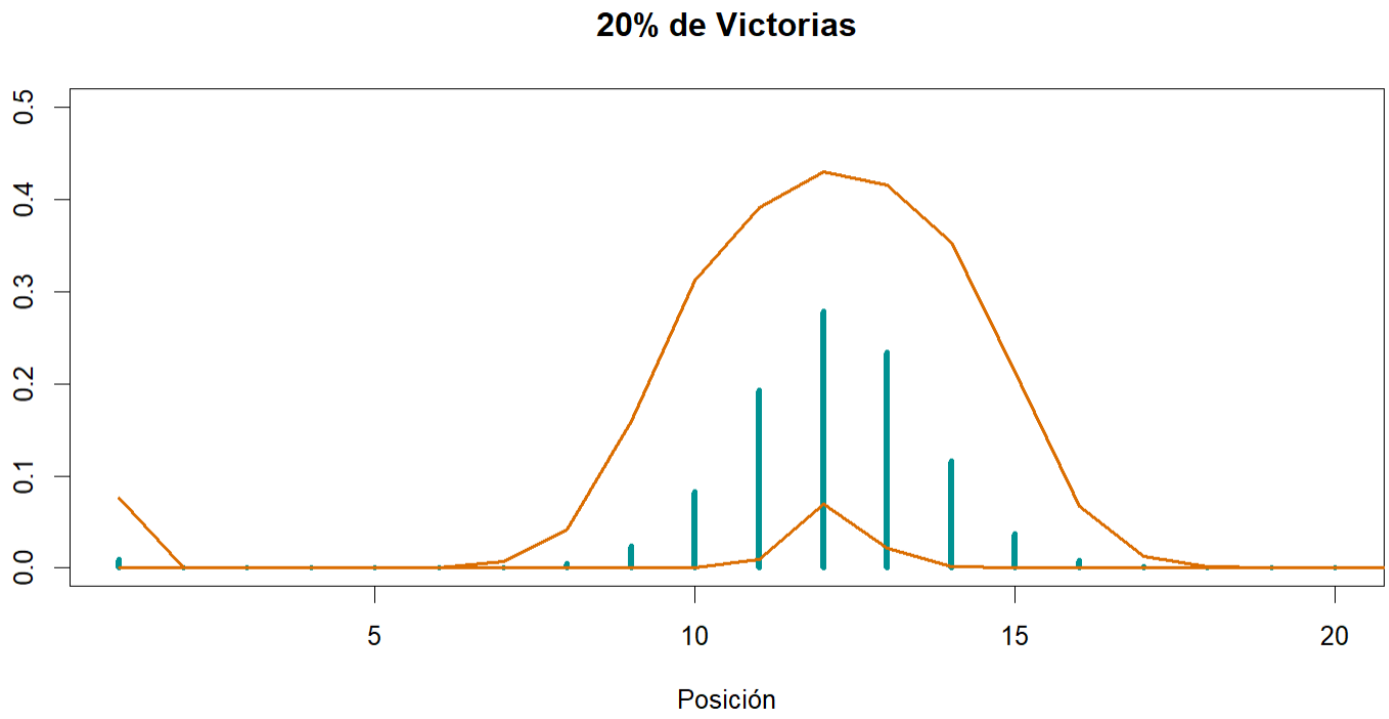


Figura 3: Estimación de la distribución de la posición del equipo si éste posee 50 goles a favor, 50 goles en contra y un porcentaje de victoria de un $20\,\%$, junto a una región de $95\,\%$ de credibilidad.

Resultados: Posición $\sim$ GF, GC, %Victorias

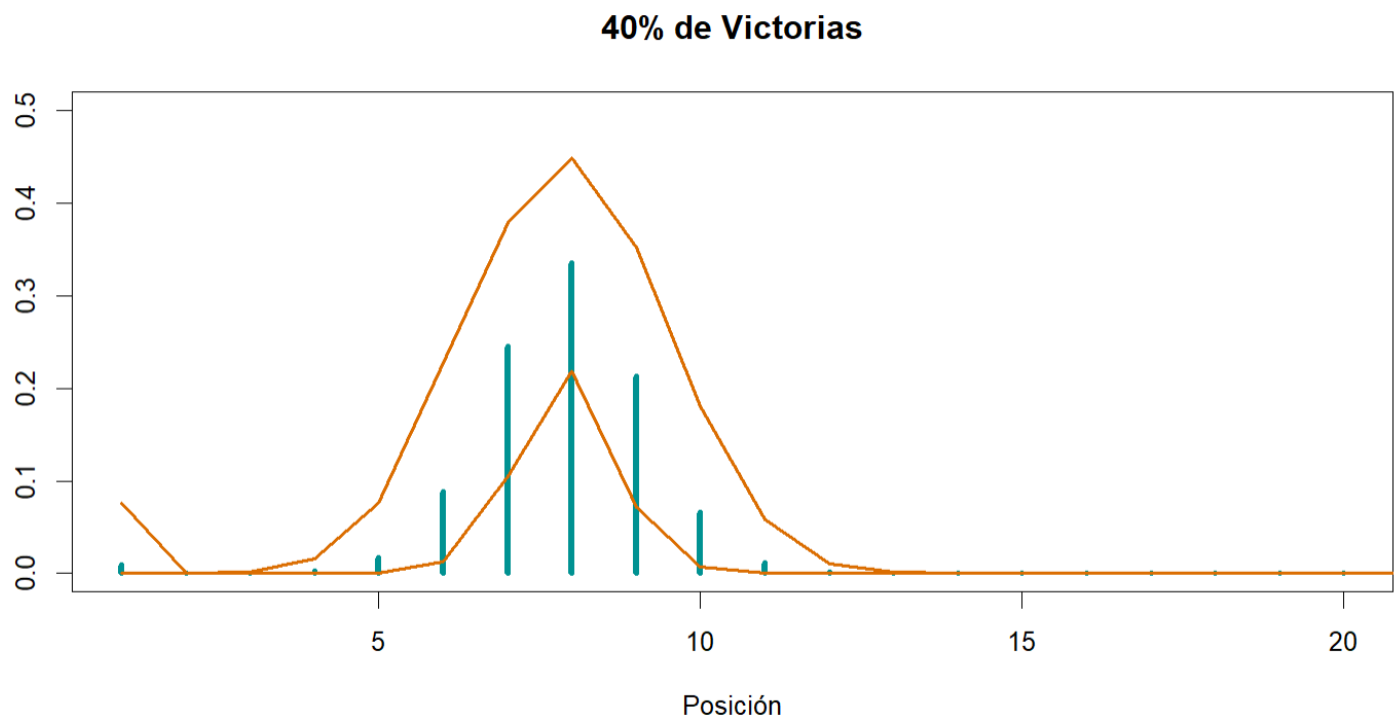


Figura 4: Estimación de la distribución de la posición del equipo si éste posee 50 goles a favor, 50 goles en contra y un porcentaje de victoria de un $40\,\%$, junto a una región de $95\,\%$ de credibilidad.

Resultados: Posición $\sim$ GF, GC, %Victorias

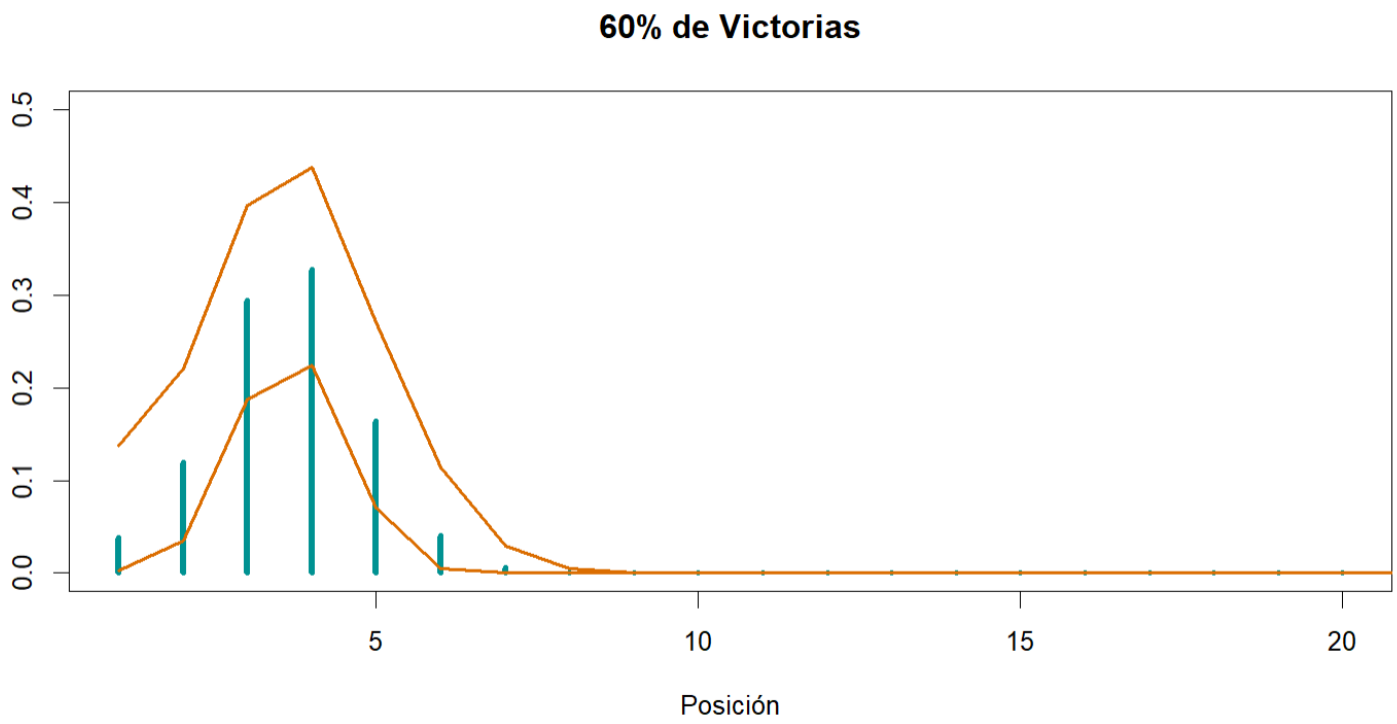


Figura 5: Estimación de la distribución de la posición del equipo si éste posee 50 goles a favor, 50 goles en contra y un porcentaje de victoria de un $60\,\%$, junto a una región de $95\,\%$ de credibilidad.

Resultados: Posición $\sim$ GF, GC, %Victorias

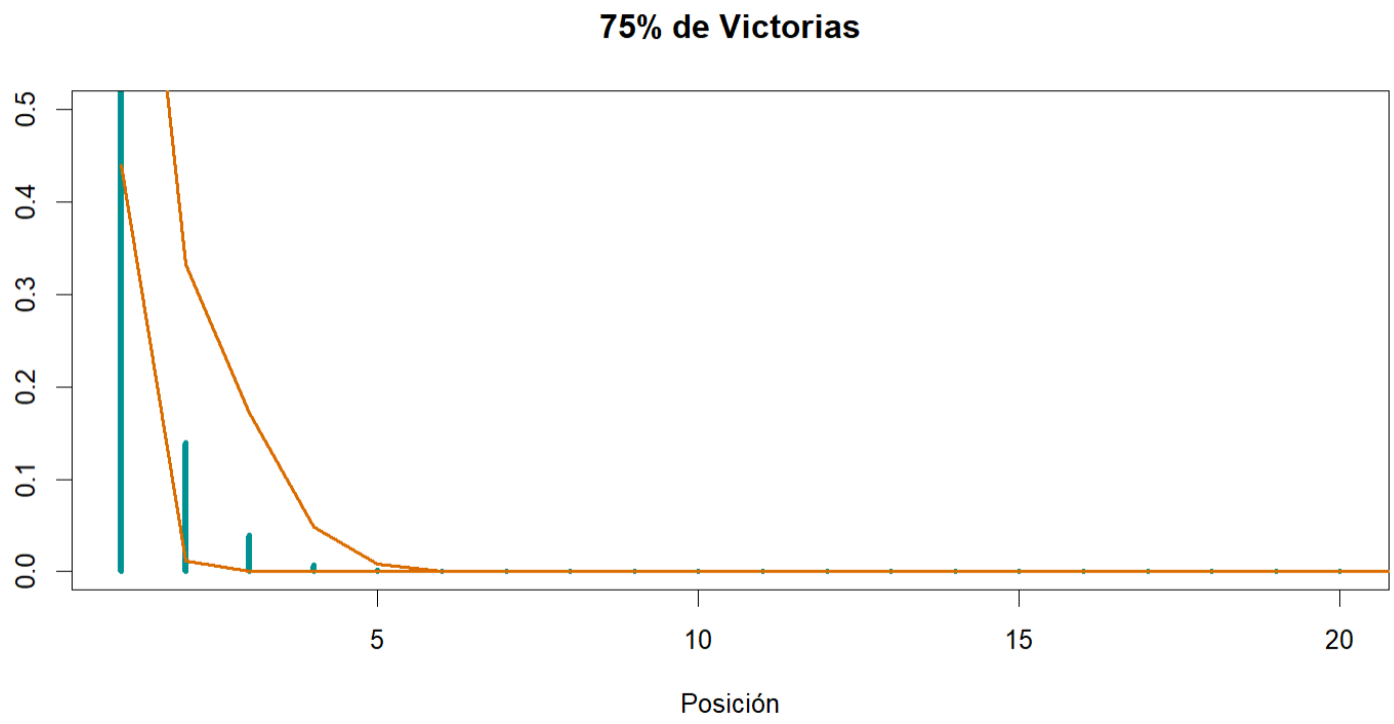


Figura 6: Estimación de la distribución de la posición del equipo si éste posee 50 goles a favor, 50 goles en contra y un porcentaje de victoria de un $75\,\%$, junto a una región de $95\,\%$ de credibilidad.

Contenidos

Discusión

Discusión

- El modelo RCRK permite modelar datos discretos de diversas naturalezas.

Discusión

- El modelo RCRK permite modelar datos discretos de diversas naturalezas.
- Extención mediante el uso de un proceso más general que el LDDP.

Discusión

- El modelo RCRK permite modelar datos discretos de diversas naturalezas.
- Extención mediante el uso de un proceso más general que el LDDP.
- Simulación MCMC exacta.

Referencias

Canale, A., and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. Journal of the American Statistical Association, 106(496), 1528-1539.

Ishwaran, H., and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American statistical Association, 96(453), 161-173.

Womack, A. J., Fuentes, C., and Taylor-Rodriguez, D. (2015). Model Space Priors for Objective Sparse Bayesian Regression. arXiv preprint arXiv:1511.04745. 651, 654, 657.

¡Gracias por su atención!

Modelo BNP para datos discretos

con aplicación al rendimiento de clubes deportivos

Cristian Capetillo Constela

Pontificia Universidad Católica de Chile
Facultad de matemática
Departamento de Estadística

11 de Noviembre, 2024

Resultados

Data generator	Data size	$(1,1,1,1)$	$(1,1,1,0)$	$(1,1,0,1)$	$(1,0,1,1)$	$(1,1,0,0)$	$(1,0,1,0)$	$(1,0,0,1)$	$(1,0,0,0)$
N-cat	$n=50$	0	0	0	0	**0.0081**	0.0006	0.0006	0.9906
	$n=300$	0	0.0006	0.0019	0	**0.9956**	0	0	0.0019
	$n=1000$	0	0.0013	0.0006	0	**0.9981**	0	0	0
Pois-cat	$n=50$	0	0.0025	0	0.0006	0	**0.8494**	0.0019	0.1456
	$n=300$	0	0	0	0	0	**1**	0	0
	$n=1000$	0	0	0	0	0	**1**	0	0
NB-cat	$n=50$	0	0	0.0038	0.0044	0	0	**0.9456**	0.0463
	$n=300$	0	0	0	0	0	0	**1**	0
	$n=1000$	0	0	0	0	0	0	**1**	0
ZIP-cat	$n=50$	0	0	0	0	0.0013	0.0031	**0.0019**	0.9938
	$n=300$	0	0	0	0	0	0.0069	**0.0069**	0.9863
	$n=1000$	0	0	0	0	0	0	**0.9163**	0.0838
ZINB-cat	$n=50$	0	0.0031	0.0005	0	**0.6581**	0.0006	0	0.3331
	$n=300$	0	0	0.0006	0	**0.9994**	0	0	0
	$n=1000$	0	0	0	0	**1**	0	0	0