Maximum likelihood estimation and Expectation-Maximization algorithm for controlled branching processes

M. González, C. Minuesa*, I. del Puerto

Department of Mathematics, University of Extremadura, 06006, Badajoz. Spain.

Abstract

Supplementary material

Simulated data

We consider a CBP whose offspring distribution is given by $p_0 = 0.1084$, $p_1 = 0.2709$, $p_2 = 0.3386$ and $p_3 = 0.2822$, and the control variables $\phi_n(k)$ follow binomial distributions with parameters k and $q = 0.6$. Thus, the offspring mean and variance are $m = 1.7946$ and $\sigma^2 = 0.9443$, respectively; $\theta = 1.5$, $\mu(\theta) = 0.6$ and the mean growth rate is $\tau_m = 1.0767$. We simulated the first 30 generations of this process starting with $Z_0 = 1$ individual. We denote by z^*_{30}, $\overline{z}_{30}$ and z_{30}, the samples based on the entire family tree, on the the total number of individuals and progenitors in each generation, and on the generation sizes only, respectively. The data obtained are the following:

n	Z_n	$\phi_n(Z_n)$	$Z_n(0)$	$Z_n(1)$	$Z_n(2)$	$Z_n(3)$
0	1	1	0	1	0	0
1	1	1	0	1	0	0
2	1	1	0	0	1	0
3	2	1	0	0	1	0
4	2	2	0	0	1	1
5	5	5	0	2	2	1

*Corresponding Author: Phone: +34 924289300 ext. 86820. Fax: +34 924272911. E-mail: cminuesaa@unex.es

This is the plain accepted version of the Supplementary Material of the following paper published in the journal *Computational Statistics and Data Analysis* (see the official journal website at `https://doi.org/10.1016/j.csda.2015.01.015`):

González, M., Minuesa, C., del Puerto, I., 2016. Maximum likelihood estimation and expectation-maximization algorithm for controlled branching processes. Computational Statistics & Data Analysis 93, 209–227. DOI: 10.1016/j.csda.2015.01.015

6	9	6	1	2	2	1
7	9	7	2	2	1	2
8	10	8	0	3	1	4
9	17	14	1	8	3	2
10	20	14	0	8	2	4
11	24	17	2	2	6	7
12	35	25	1	6	8	10
13	52	39	4	11	14	10
14	69	48	10	15	13	10
15	71	38	8	9	14	7
16	58	36	1	5	17	13
17	78	51	5	13	15	18
18	97	61	7	28	13	13
19	93	61	5	22	22	12
20	102	64	6	15	20	23
21	124	72	7	14	25	26
22	142	76	5	24	32	15
23	133	73	9	21	22	21
24	128	81	9	16	33	23
25	151	86	8	21	34	23
26	158	83	7	19	34	23
27	156	94	11	26	32	25
28	165	94	11	29	24	30
29	167	107	10	27	37	33
30	200	·	·	·	·	·
	z_{30}					
	$\overline{z}_{30}$					
			z^*_{30}			

Table 1: Simulated data

Analysis of the robustness of the EM algorithm based on the sample z_{30}

The EM algorithm based on the sample z_{30} is observed not to be at all robust to the choice of the initial values $(p^{(0)}, \theta^{(0)})$, with convergence to different estimates that could be local maxima or saddle points of the log-likelihood function. We detected this fact by starting the algorithm with 300 different initial values, choosing each $p^{(0)}$ randomly from a Dirichlet distribution with all the parameters equal to one (that is, uniformly from the unit simplex) and each $\theta^{(0)}$ through the equation $\theta^{(0)} = q^{(0)}(1 - q^{(0)})^{-1}$, with $q^{(0)}$ sampled from a uniform distribution on the open interval $(0, 1)$. To overcome this problem, we propose in the paper a methodological

approach in order to choose the best approximation to the MLE based on $\mathcal{Z}_n$ -called EM estimate-, that we analyze in detail below.

The log-likelihood function based on the sample $\mathcal{Z}_n = \{Z_0, \ldots, Z_n\}$, denoted by $\ell(p, \theta \mid \mathcal{Z}_n)$, is given, in the case of binomial control functions, by

$$\ell(p, \theta \mid Z_l = z_l,\ l = 0, \ldots, n) = \sum_{j=0}^{n-1} \log \left(\sum_{l=0}^{z_j} P^{*l}_{z_{j+1}} \binom{z_j}{l} \frac{\theta^l}{(1+\theta)^{z_j}} \right) \tag{1}$$

with $P^{*l}_{\cdot}$ denoting the l-fold convolution of the offspring law p. We assume, for computational purposes, that the support of this distribution is $\{0, \ldots, s_{max}\}$, with s_{max} denoting the maximum number of offspring per progenitor (in our example $s_{max} = 3$). Notice that $P^{*l}_{z_{j+1}}$ is the coefficient of the monomial of degree z_{j+1} of the polynomial $(\sum_{k=0}^{s_{max}} p_k s^k)^l$. This fact allows us to develop a computational program to evaluate the log–likelihood function on each pair (p, θ).

Thus, the methodological approach we propose consists of taking as EM estimates of the parameters those associated with the greatest log-likelihood when is evaluated at the convergence points of the EM algorithm started with different randomly chosen values of the parameters. Related to our example, in the following figures we show the exact values of the log-likelihood function versus the convergence points of the EM algorithm started with each one of the 300 different seeds, for the parameters p_0 (Figure 1, left), p_1 (Figure 1, center), p_2 (Figure 1, right), p_3 (Figure 2, left) and $\mu(\theta)$ (Figure 2, right).

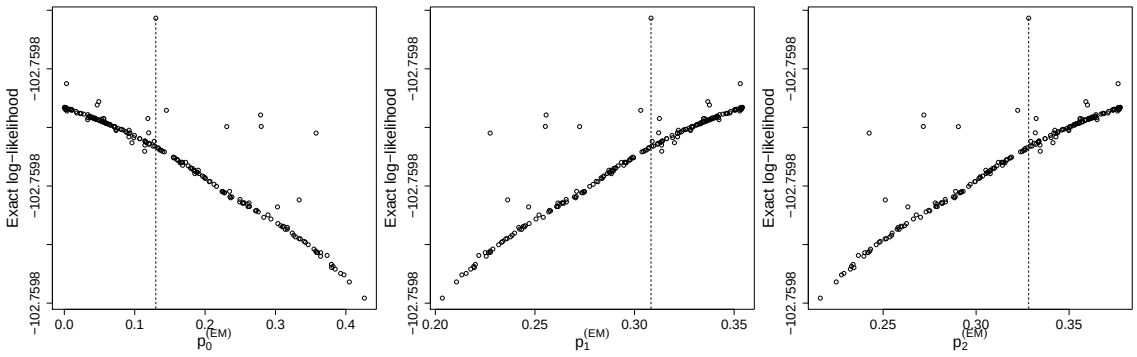

Figure 1: Exact log-likelihood function versus the convergence points of the EM algorithm for the parameters p_0, p_1 and p_2, denoted by $p_0^{(EM)}$, $p_1^{(EM)}$ and $p_2^{(EM)}$.

The values that maximizes the log-likelihood function (1) are given in Table 2, and shown in Figures 1 and 2 with vertical dashed lines.

Finally, it is also worth mentioning that the expectation of the log-likelihood $\ell(p, \theta \mid \mathcal{Z}_n^*,\ \mathcal{Z}_n)$ with respect to the distribution $\mathcal{Z}_n^* | (\mathcal{Z}_n, \{p^{(EM)}, \theta^{(EM)}\})$, with $(p^{(EM)}, \theta^{(EM)})$ a convergence point of the EM algorithm –this kind of expected values can be calculated in each iteration of the algorithm (see Equation (7) in the paper)–,

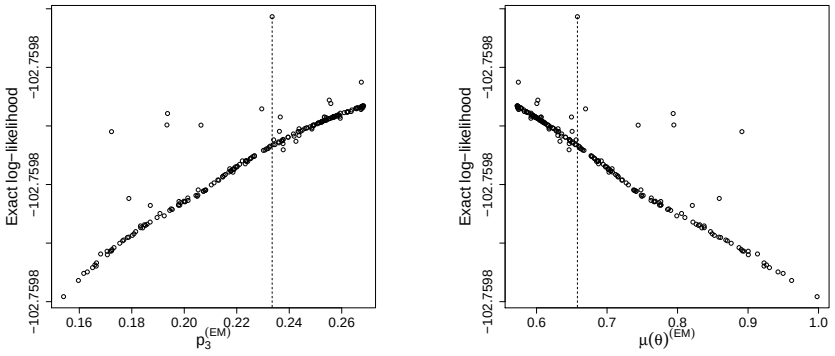

Figure 2: Exact log-likelihood function versus the convergence points of the EM algorithm for the parameters p_3 and $\mu(\theta)$, denoted by $p_3^{(EM)}$ and $\mu(\theta)^{(EM)}$.

	PARAMETERS							
SAMPLE	p_0	p_1	p_2	p_3	m	σ^2	$\mu(\theta)$	τ_m
z_{30}	.1299	.3083	.3283	.2335	1.6653	.9496	.6579	1.0957
TRUE VALUE	.1084	.2709	.3386	.2822	1.7946	.9443	.6000	1.0767

Table 2: Estimates of the parameters of interest based on the sample z_{30}.

can not be used to determine the maximum likelihood estimates, as an alternative to our proposal. This is due to the fact that it can happen that

$$E_{\mathcal{Z}_n^*|(\mathcal{Z}_n,\{p^{(EM)},\theta^{(EM)}\})}[\ell(p,\theta \mid \mathcal{Z}_n^*,\mathcal{Z}_n)] \geq E_{\mathcal{Z}_n^*|(\mathcal{Z}_n,\{\tilde{p}^{(EM)},\tilde{\theta}^{(EM)}\})}[\ell(p,\theta \mid \mathcal{Z}_n^*,\mathcal{Z}_n)]$$

and

$$\ell(p^{(EM)},\theta^{(EM)} \mid \mathcal{Z}_n) < \ell(\tilde{p}^{(EM)},\tilde{\theta}^{(EM)} \mid \mathcal{Z}_n),$$

being $(p^{(EM)},\ \theta^{(EM)})$ and $(\tilde{p}^{(EM)},\ \tilde{\theta}^{(EM)})$ two convergence points provided by the EM algorithm. Figure 3 shows this fact. We plot on it $\ell(p^{(EM)},\theta^{(EM)}\ |z_{30})$ versus $E_{\mathcal{Z}_{30}^*|(z_{30},\{p^{(EM)},\theta^{(EM)}\})}[\ell(p,\theta \mid \mathcal{Z}_{30}^*,z_{30})]$, with $(p^{(EM)},\theta^{(EM)})$ the convergence points of the EM algorithm started with the 300 randomly chosen seeds.

Computational complexity

Focussing our interest in the case of binomial control functions, in order to determine upper bounds of the values b_l and b_l^*, let us introduce the following functions. Let $b(z_l,\phi_l^*,z_{l+1},s_{max})$ the function that provides the number of possible vectors $(z_l(0),\ldots,z_l(s_{max}))$ such that $\sum_{k=0}^{s_{max}} z_l(k)=\phi_l^*$ and $\sum_{k=0}^{s_{max}} kz_l(k)=z_{l+1}$, with z_l, ϕ_l^* and z_{l+1} whatever possible values of the variables Z_l, $\phi_l(Z_l)$ and Z_{l+1}, respectively, and s_{max} with the maximum number of offspring per progenitor. Notice that given the sample $\overline{\mathcal{Z}}_n$, then $b_l=b(Z_l,\phi_l(Z_l),Z_{l+1},s_{max})$, $l=0,1,\ldots,n-1$. It is also remarkable that b_l depends on Z_l through $\phi_l(Z_l)$. Analogously, let $b^*(z_l,z_{l+1},s_{max})$ the

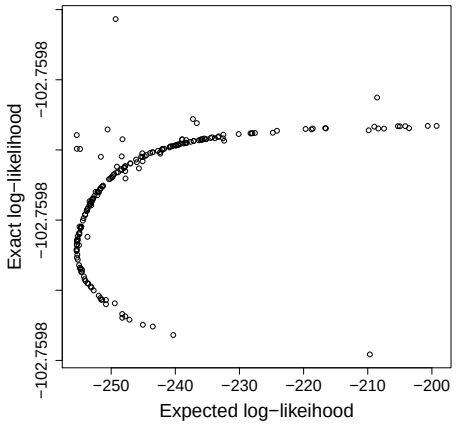

Figure 3: Exact log-likelihood function versus expected log-likelihood.

function that provides the number of possible vectors $(z_l(0), \ldots, z_l(s_{max}))$ such that $1 \leq \sum_{k=0}^{s_{max}} z_l(k) \leq z_l$ and $\sum_{k=0}^{s_{max}} k z_l(k) = z_{l+1}$, with z_l, and z_{l+1} whatever possible values of the variables Z_l and Z_{l+1}, respectively (we have not considered the case $\sum_{k=0}^{s_{max}} z_l(k) = 0$ because it does not lead to any factible value in the case $z_{l+1} = 0$ or to the null vector if $z_{l+1} = 0$; in any case its contribution is irrelevant for our purpose). Notice that $b^*(z_l, z_{l+1}, s_{max}) = \sum_{1 \leq \phi_l^* \leq z_l} b(z_l, \phi_l^*, z_{l+1}, s_{max})$. Assuming the sample $\mathcal{Z}_n$, $b_l^* = b^*(Z_l, Z_{l+1}, s_{max})$, $l = 0, 1, \ldots, n-1$.

Thus, to obtain upper bounds of the functions b and b^* in terms of z_l and s_{max}, we determined the functions

$$\begin{aligned} b_{max} &= b_{max}(z_l, s_{max}) = \max_{\substack{1 \leq \phi_l^* \leq z_l \\ 0 \leq z_{l+1} \leq s_{max} \cdot \phi_l^*}} b(z_l, \phi_l^*, z_{l+1}, s_{max}) \\ &= \max_{0 \leq z_{l+1} \leq s_{max} \cdot z_l} b(z_l, z_l, z_{l+1}, s_{max}) \end{aligned} \tag{2}$$

and

$$b_{max}^* = b_{max}^*(z_l, s_{max}) = \max_{0 \leq z_{l+1} \leq s_{max} \cdot z_l} b^*(z_l, z_{l+1}, s_{max}). \tag{3}$$

To get these maximum values, we have considered three possible values of $s_{max} = 3, 4, 5$, and for each one we have obtained the values of the function b for ϕ_l^* going from 1 to 167 (this is the maximum value of $\phi_l(Z_l)$ in our simulated sample -see Table 1) and for z_{l+1} going from 0 to $167 \cdot s_{max}$. The values obtained have been stored in matrices of dimension $(167 \cdot s_{max} + 1) \times 167$. Each column corresponds to a possible value of ϕ_l^*, going from 1 to 167, and each row to one of z_{l+1}, from 0 to $167 \cdot s_{max}$. Notice that the non-null values for the column corresponding to ϕ_l^* are the $\phi_l^* \cdot s_{max} + 1$ first elements. The matrices obtained are stored in the files `tree-max-3.cvs` (for $s_{max} = 3$), `tree-max-4.cvs` (for $s_{max} = 4$) and `tree-max-5.cvs` (for $s_{max} = 5$). From them, and taking into account (2) and (3), it

is easy to obtain the values of $b_{max}(z_l, s_{max})$ and $b^*_{max}(z_l, s_{max})$, which are given in Table 3. For each value of s_{max}, analysing the relationship between z_l and $b_{max}(z_l, s_{max})$ and z_l and $b^*_{max}(z_l, s_{max})$, using polynomial regression methods, it can be concluded that $b_{max}(z_l, s_{max}) = O(z_l^{s_{max}-1})$ and $b^*_{max}(z_l, s_{max}) = O(z_l^{s_{max}})$. Consequently, we infer that $b_l = O(Z_l^{s_{max}-1})$ and $b^*_l = O(Z_l^{s_{max}})$.

	$s_{max}=3$		$s_{max}=4$		$s_{max}=5$	
z_l	b_{max}	b^*_{max}	b_{max}	b^*_{max}	b_{max}	b^*_{max}
1	1	1	1	1	1	1
2	2	3	3	4	3	4
3	3	6	5	8	6	9
4	5	9	8	14	12	19
5	6	15	12	24	20	36
6	8	22	18	37	32	63
7	10	29	24	58	49	103
8	13	39	33	85	73	164
9	15	51	43	117	102	249
10	18	66	55	164	141	369
11	21	84	69	218	190	525
12	25	103	86	287	252	736
13	28	124	104	372	325	1006
14	32	150	126	473	414	1355
15	36	178	150	598	521	1790
16	41	213	177	736	649	2332
17	45	249	207	906	795	3000
18	50	286	241	1102	967	3809
19	55	331	277	1326	1165	4789
20	61	378	318	1585	1394	5953
21	66	433	362	1875	1651	7337
22	72	492	410	2210	1944	8965
23	78	552	462	2586	2275	10873
24	85	618	519	3002	2649	13091
25	91	691	579	3478	3061	15653
26	98	769	645	3997	3523	18603
27	105	856	715	4575	4035	21982
28	113	945	790	5217	4604	25833
29	120	1036	870	5923	5225	30213
30	128	1140	956	6706	5910	35153
31	136	1246	1046	7545	6660	40728
32	145	1366	1143	8475	7483	46986
33	153	1489	1245	9486	8372	54003
34	162	1614	1353	10585	9343	61824
35	171	1750	1467	11779	10395	70533
36	181	1893	1588	13062	11538	80195
37	190	2046	1714	14456	12764	90880
38	200	2209	1848	15956	14090	102681
39	210	2374	1988	17565	15516	115675
40	221	2545	2135	19309	17053	129965
41	231	2731	2289	21161	18691	145621
42	242	2921	2451	23146	20451	162758
43	253	3129	2619	25271	22330	181469
44	265	3340	2796	27544	24342	201853
45	276	3553	2980	29976	26476	224027
46	288	3784	3172	32537	28754	248116
47	300	4021	3372	35277	31174	274220
48	313	4274	3581	38181	33751	302490
49	325	4536	3797	41269	36471	333023
50	338	4801	4023	44542	39361	365983
51	351	5077	4257	47991	42416	401493
52	365	5368	4500	51647	45654	439716
53	378	5668	4752	55500	49060	480793
54	392	5987	5014	59569	52662	524907
55	406	6309	5284	63877	56455	572201
56	421	6634	5565	68388	60459	622833
57	435	6985	5855	73135	64656	676982
58	450	7339	6155	78124	69079	734837
59	465	7717	6465	83383	73720	796574
60	481	8102	6786	88913	78602	862397
61	496	8490	7116	94674	83705	932496
62	512	8896	7458	100732	89064	1007118
63	528	9316	7810	107061	94671	1086429
64	545	9751	8173	113710	100551	1170652
65	561	10204	8547	120663	106681	1260022
66	578	10661	8933	127909	113101	1354759
67	595	11126	9329	135487	119799	1455114
68	613	11617	9738	143377	126804	1561303
69	630	12113	10158	151630	134091	1673615
70	648	12640	10590	160256	141702	1792281
71	666	13171	11034	169209	149625	1917588
72	685	13706	11491	178531	157891	2049806
73	703	14267	11959	188219	166471	2189195
74	722	14839	12441	198341	175413	2336068
75	741	15434	12935	208875	184701	2490790
76	761	16045	13442	219772	194370	2653595
77	780	16660	13962	231113	204389	2824802
78	800	17290	14496	242854	214808	3004715
79	820	17945	15042	255093	225610	3193697
80	841	18611	15603	267785	236833	3392022
81	861	19307	16177	280916	248442	3600097
82	882	20008	16765	294534	260493	3818240
83	903	20713	17367	308599	272965	4046845
84	925	21454	17984	323218	285900	4286352
85	946	22202	18614	338373	299260	4537044
86	968	22982	19260	354009	313104	4799346

87	990	23774	19920	370176	327410	5073587
88	1013	24571	20595	386859	342223	5360240
89	1035	25391	21285	404172	357501	5659681
90	1058	26233	21991	422074	373309	5972353
91	1081	27094	22711	440500	389620	6298680
92	1105	27983	23448	459547	406484	6639100
93	1128	28877	24200	479155	423856	6994059
94	1152	29780	24968	499457	441804	7364046
95	1176	30722	25752	520404	460300	7749502
96	1201	31669	26553	541966	479397	8150987
97	1225	32659	27369	564199	499045	8569019
98	1250	33656	28203	587040	519319	9004036
99	1275	34658	29053	610656	540186	9456496
100	1301	35693	29920	635009	561704	9926962
101	1326	36746	30804	660026	583820	10415946
102	1352	37827	31706	685766	606612	10924007
103	1378	38932	32624	712211	630045	11451680
104	1405	40043	33561	739484	654181	11999519
105	1431	41171	34515	767575	678962	12568160
106	1458	42335	35487	796381	704473	13158290
107	1485	43511	36477	826014	730674	13770375
108	1513	44730	37486	856405	757632	14405000
109	1540	45955	38512	887694	785285	15062801
110	1568	47186	39558	919870	813722	15744388
111	1596	48461	40622	952866	842901	16450391
112	1625	49748	41705	986747	872893	17181481
113	1653	51074	42807	1021440	903631	17938314
114	1682	52419	43929	1057121	935211	18721555
115	1711	53770	45069	1093798	967585	19531923
116	1741	55148	46230	1131352	1000830	20370041
117	1770	56558	47410	1169850	1034875	21236691
118	1800	57989	48610	1209272	1069820	22132755
119	1830	59459	49830	1249738	1105615	23058898
120	1861	60936	51071	1291297	1142341	24015730
121	1891	62421	52331	1333791	1179921	25004058
122	1922	63959	53613	1377348	1218463	26024628
123	1953	65504	54915	1421889	1257911	27078204
124	1985	67100	56238	1467550	1298352	28165598
125	2016	68708	57582	1514386	1339705	29287641
126	2048	70323	58948	1562277	1382082	30445101
127	2080	71976	60334	1611296	1425424	31638919
128	2113	73655	61743	1661361	1469823	32870171
129	2145	75366	63173	1712643	1515192	34139421
130	2178	77111	64625	1765228	1561651	35447522
131	2211	78863	66099	1818932	1609135	36795378
132	2245	80633	67596	1873831	1657742	38183913
133	2278	82452	69114	1929902	1707380	39613984
134	2312	84282	70656	1987249	1758174	41086588
135	2346	86169	72220	2046012	1810056	42602574
136	2381	88064	73807	2105960	1863129	44162971
137	2415	89966	75417	2167236	1917295	45768701
138	2450	91919	77051	2229752	1972687	47420788
139	2485	93891	78707	2293624	2029230	49120185
140	2521	95907	80388	2359009	2087034	50868375
141	2556	97950	82092	2425713	2145996	52665971
142	2592	100001	83820	2493818	2206254	54513965
143	2628	102081	85572	2563232	2267730	56413420
144	2665	104204	87349	2634106	2330539	58365425
145	2701	106349	89149	2706640	2394571	60371037
146	2738	108546	90975	2780565	2459973	62431379
147	2775	110751	92825	2855965	2526660	64547499
148	2813	112964	94700	2932815	2594754	66720601
149	2850	115242	96600	3011200	2664140	68951927
150	2888	117530	98526	3091346	2734970	71242921
151	2926	119876	100476	3172971	2807155	73594370
152	2965	122241	102453	3256219	2880823	76007435
153	3003	124614	104455	3340992	2955852	78483347
154	3042	127029	106483	3427378	3032403	81023258
155	3081	129480	108537	3515649	3110380	83628483
156	3121	131965	110618	3605542	3189918	86300210
157	3160	134495	112724	3697138	3270889	89039747
158	3200	137034	114858	3790336	3353460	91848366
159	3240	139590	117018	3885301	3437531	94727319
160	3281	142209	119205	3982229	3523243	97677964
161	3321	144837	121419	4080903	3610461	100701803
162	3362	147538	123661	4181362	3699361	103800552
163	3403	150248	125929	4283578	3789835	106975060
164	3445	152967	128226	4387645	3882032	110226672
165	3486	155742	130550	4493778	3975811	113556854
166	3528	158544	132902	4601761	4071354	116966891
167	3570	161394	135282	4711691	4168549	120458340

Table 3: Values of $b_{max}(z_l, s_{max})$ and $b^*_{max}(z_l, s_{max})$