September 30, 2016

BAYESIAN MIXTURE OF PLACKETT-LUCE MODELS FOR PARTIALLY RANKED DATA

Supplementary Material

Implementation details for model assessment: posterior predictive checks $p_{B(1)}$ and $p_{B(2)}$

We remind that the posterior predictive p value represents the posterior probability that a parameter-dependent discrepancy measure $X^2_{(d)}(\underline{\pi}^{-1}_{\text{obs}};\theta)$, comparing actually observed frequencies and expected frequencies under the assumed model H, does not exceed the same discrepancy measure $X^2_{(d)}(\underline{\pi}^{-1}_{\text{rep}};\theta)$ evaluated on a replicated data set drawn from the same model, that is,

$$p_{B(d)} = \mathbf{P}\left(X^2_{(d)}(\underline{\pi}^{-1}_{\text{rep}};\theta) \geq X^2_{(d)}(\underline{\pi}^{-1}_{\text{obs}};\theta) \middle| \underline{\pi}^{-1}_{\text{obs}}, H \right).$$

The value $p_{B(d)}$ can be easily approximated by using the posterior simulations of the parameter vector θ and augmenting them with the corresponding draws of replicated data $\underline{\pi}^{-1}_{\text{rep}}$. For our first discrepancy measure $X^2_{(1)}(\underline{\pi}^{-1};\theta) = \sum_{i=1}^{K} \frac{\left(r_i(\underline{\pi}^{-1}) - r^*_i(\theta)\right)^2}{r^*_i(\theta)}$, the theoretical frequencies expected under PL mixture model with parameter $\theta = (\underline{p}, \underline{\omega})$ depend on the marginal overall support parameters $p_i = \sum_{g=1}^{G} \omega_g p_{gi}$ (for $i = 1, \ldots, K$) and are easily determined as follows:

$$r^*_i(\theta) = Np_i.$$

For the discrepancy measure $X^2_{(2)}(\underline{\pi}^{-1};\theta) = \sum_{i<i'} \frac{(\tau_{ii'}(\underline{\pi}^{-1}) - \tau^*_{ii'}(\theta))^2}{\tau^*_{ii'}(\theta)}$, one can derive the expected paired comparison frequencies under PL mixture model as follows:

$$\tau^*_{ii'}(\theta) = T_{ii'} \frac{p_i}{p_i + p_{i'}},$$

where $T_{ii'} = \tau_{ii'} + \tau_{i'i}$ indicates the total number of pairwise comparisons between item i and i'.

Implementation details for model assessment: posterior predictive checks $\tilde{p}_{B(1)}$ and $\tilde{p}_{B(2)}$

Let $m = 1, \ldots, K-1$ be the generic number of ranked items in a partial ordering of K items. We denote with $\underline{\pi}_m^{-1} = \{\pi_s^{-1} : n_s = m\}$ the subsample of N_m top-m orderings ($\sum_{m=1}^{K-1} N_m = N$). In order to assess the model adequacy regarding the homogeneity assumption on the conditional distributions given the same number m of ranked items, we define the discrepancy between each conditional distribution and the marginal distribution of the most-liked item by using the conditional frequencies $r_{i,m}$ as follows:

$$\tilde{X}^2_{(1)}(\underline{\pi}^{-1};\theta) = \sum_{m=1}^{K-1}\sum_{i=1}^{K} \frac{(r_{i,m} - r^*_{i,m}(\theta))^2}{r^*_{i,m}(\theta)},$$

where $r_{i,m} = r_i(\underline{\pi}_m^{-1})$ and $r^*_{i,m}(\theta) = N_m p_i$. Similarly, when we aim at assessing homogeneity of the conditional pairwise comparison frequencies, we define

$$\tilde{X}^2_{(2)}(\underline{\pi}^{-1};\theta) = \sum_{m=1}^{K-1}\sum_{i<i'} \frac{(\tau_{ii',m} - \tau^*_{ii',m}(\theta))^2}{\tau^*_{ii',m}(\theta)},$$

where $\tau_{ii',m} = \tau_{ii'}(\underline{\pi}_m^{-1})$ and $\tau^*_{ii',m}(\theta) = T_{ii',m}\frac{p_i}{p_i+p_{i'}}$ with $T_{ii',m} = \tau_{ii',m} + \tau_{i'i,m}$. The computation of $\tilde{p}_{B(1)}$ and $\tilde{p}_{B(2)}$ follows the general formula (4) in the main paper by replacing the desired discrepancy measure.

Supplemental tables and figures

TABLE SM-1.

Simulation study (Censoring setting A) Distribution (%) of the optimal number $\hat{G}$ of components identified, respectively, by the Bayesian PL mixture analysis via alternative model selection criteria and by the BNPPLM analysis via Dahl's procedure. In the simulation study, 100 data sets with 1000 partial orderings of 6 items were generated from each PL mixture scenario with alternative true number G^* of components. Best agreement rates under each simulation scenario are highlighted in bold.

$G^* = 1$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	99	98	**100**	**100**	**100**	99	**100**	**100**
2	1	2	0	0	0	1	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 2$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	2	2	2	2	6	6	3	4
2	96	93	**98**	95	93	89	97	91
3	2	3	0	3	1	3	0	4
4	0	1	0	0	0	2	0	0
5	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 3$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	0	0	0	0	0	0	0	0
2	4	4	5	6	8	8	7	6
3	91	89	**94**	92	91	88	93	88
4	5	5	1	2	1	3	0	5
5	0	1	0	0	0	1	0	1
6	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0

$G^* = 4$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	0	0	1	1	1	1	1	1
2	0	0	0	0	3	3	3	1
3	18	14	22	22	30	30	30	25
4	**81**	67	77	70	65	60	66	50
5	1	10	0	3	1	6	0	19
6	0	7	0	4	0	0	0	3
7	0	2	0	0	0	0	0	1

TABLE SM-2.

Simulation study (Censoring setting B) Distribution (%) of the optimal number $\hat{G}$ of components identified, respectively, by the Bayesian PL mixture analysis via alternative model selection criteria and by the BNPPLM analysis via Dahl's procedure. In the simulation study, 100 data sets with 1000 partial orderings of 6 items were generated from each PL mixture scenario with alternative true number G^* of components. Best agreement rates under each simulation scenario are highlighted in bold.

$G^* = 1$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	99	**100**	**100**	**100**	**100**	98	**100**	**100**
2	1	0	0	0	0	2	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 2$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	2	2	3	2	5	5	5	4
2	97	93	97	**98**	95	89	95	89
3	1	3	0	0	0	5	0	6
4	0	2	0	0	0	1	0	1
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 3$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	0	0	0	0	0	0	0	0
2	5	5	5	6	10	11	8	7
3	92	88	**94**	92	90	82	92	79
4	3	5	1	2	0	7	0	11
5	0	2	0	0	0	0	0	2
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1

$G^* = 4$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	1	1	1	1	1	1	1	1
2	0	0	0	1	6	6	5	2
3	21	21	30	30	39	39	37	19
4	**76**	61	69	65	53	48	57	50
5	2	9	0	3	1	6	0	20
6	0	7	0	0	0	0	0	5
7	0	1	0	0	0	0	0	2
8	0	0	0	0	0	0	0	1

TABLE SM-3.

Simulation study (Censoring setting C) Distribution (%) of the optimal number $\hat{G}$ of components identified, respectively, by the Bayesian PL mixture analysis via alternative model selection criteria and by the BNPPLM analysis via Dahl's procedure. In the simulation study, 100 data sets with 1000 partial orderings of 6 items were generated from each PL mixture scenario with alternative true number G^* of components. Best agreement rates under each simulation scenario are highlighted in bold.

$G^* = 1$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	97	98	98	**100**	**100**	**100**	**100**	**100**
2	3	1	2	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 2$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	2	2	3	3	8	8	5	5
2	**97**	91	**97**	95	91	87	95	90
3	1	6	0	2	1	5	0	5
4	0	1	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

$G^* = 3$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	0	0	0	0	0	0	0	0
2	4	5	6	6	12	11	8	8
3	**95**	89	94	92	88	82	92	63
4	1	5	0	2	0	5	0	22
5	0	1	0	0	0	1	0	6
6	0	0	0	0	0	0	0	1
7	0	0	0	0	0	1	0	0

$G^* = 4$

$\hat{G}$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	BNPPLM
1	0	0	1	1	2	2	1	1
2	1	1	0	1	9	9	6	2
3	25	23	32	30	41	39	38	22
4	**72**	62	67	64	48	46	55	37
5	2	7	0	3	0	1	0	16
6	0	6	0	1	0	0	0	13
7	0	1	0	0	0	3	0	6
8	0	0	0	0	0	0	0	3

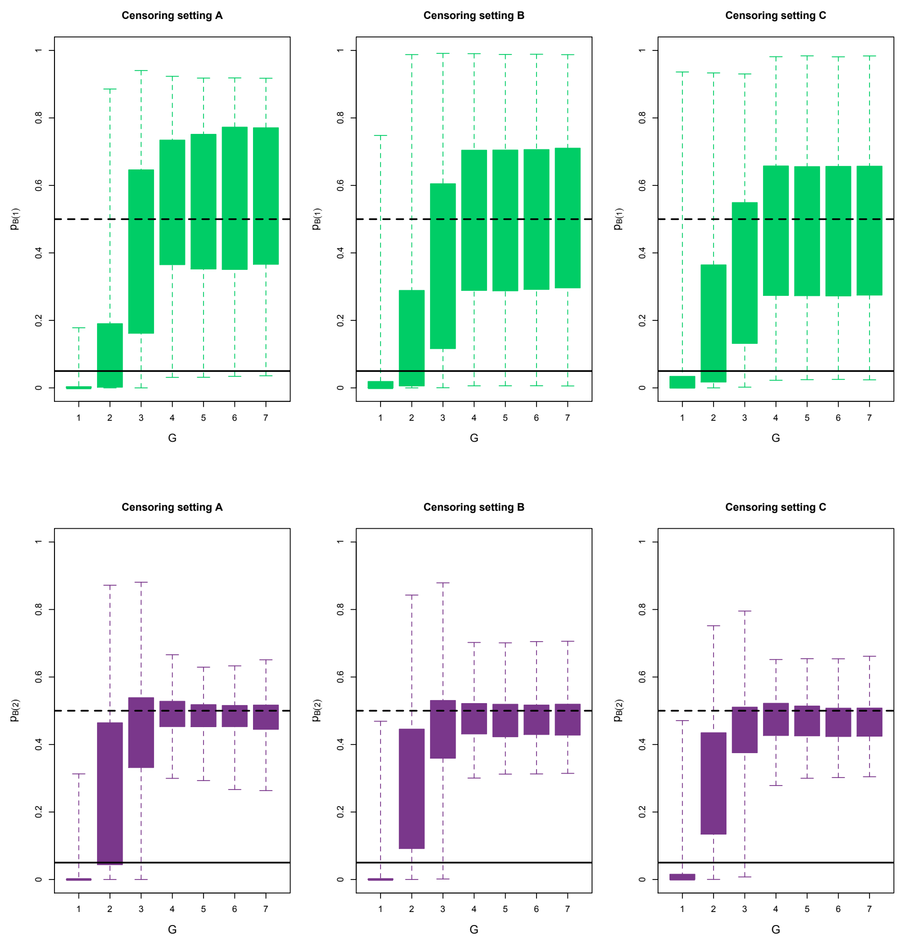

FIGURE SM-1.

Simulation study (Scenario 4) Model assessment criteria with a varying number G of fitted components for the Bayesian PL mixtures fitted to the 100 data sets simulated from the population scenario with $G^* = 4$ groups. The solid and dashed lines represent, respectively, the critical threshold 0.05 and the reference value 0.5 expected under correct model specification.

TABLE SM-4.

CARCONF data (435 full and partial orderings) Model selection criteria and posterior predictive p values for the Bayesian PL mixtures with a varying number G of components. For each selection criterion, the optimal choice of the number of components corresponds to the minimum value (*in bold*).

G	DIC_1	DIC_2	BPIC_1	BPIC_2	BICM_1	BICM_2	BIC	$p_{B(1)}$	$p_{B(2)}$
1	5288.34	5288.29	5293.32	5293.24	**5308.44**	**5308.39**	**5308.74**	0.000	0.247
2	**5268.73**	**5268.90**	**5280.15**	**5280.48**	5316.09	5316.25	5312.73	0.079	0.505
3	5278.45	5273.38	5301.99	5291.84	5348.62	5343.55	5334.66	0.092	0.515
4	5289.34	5272.67	5324.82	5291.47	5349.29	5332.61	5358.12	0.103	0.508
5	5295.06	5273.46	5336.93	5293.75	5356.14	5334.55	5387.49	0.107	0.516
6	5305.01	5274.43	5357.28	5296.12	5362.83	5332.25	5413.11	0.122	0.518

TABLE SM-5.

CARCONF data (435 full and partial orderings) Posterior means of the parameters and component-specific modal orderings of the optimal Bayesian 2-component PL mixture. Posterior standard deviations are shown in parentheses.

g	$\hat{\omega}_g$		$\hat{\sigma}_g^{-1}$	$\hat{p}_{g1}$		$\hat{p}_{g2}$		$\hat{p}_{g3}$		$\hat{p}_{g4}$		$\hat{p}_{g5}$		$\hat{p}_{g6}$	
1	.713	(.10)	(2,6,4,3,1,5)	.079	(.02)	.263	(.02)	.185	(.02)	.191	(.01)	.071	(.01)	.211	(.02)
2	.287	(.10)	(1,3,4,2,6,5)	.436	(.13)	.124	(.04)	.157	(.05)	.138	(.03)	.043	(.02)	.101	(.03)

TABLE SM-6.

APA election data (15449 full and partial orderings) Percentage of voters who assign position t to Candidate i (*upper panel*) and average rank vector (*lower panel*).

	Candidate				
Rank	A	B	C	D	E
1	18.8	14.8	26.0	21.0	19.4
2	27.7	17.7	16.9	16.9	20.7
3	23.6	24.1	14.0	18.6	19.7
4	17.5	24.7	18.3	20.3	19.3
5	14.8	18.4	23.1	23.4	20.3
$\bar{\pi}$	2.37	2.66	2.34	2.51	2.47

TABLE SM-7.

APA election data (15449 full and partial orderings) Percentage of voters who rank Candidate i in the first position conditionally on the number m of ranked candidates.

	Candidate				
m	A	B	C	D	E
1	17.4	17.1	23.3	22.3	19.9
2	21.6	11.8	31.9	18.8	16.0
3	20.1	16.3	20.1	21.8	21.7
4	18.4	13.5	28.0	20.4	19.7

TABLE SM-8.

APA election data (global analysis with 15449 full and partial orderings) Model selection criteria and posterior predictive p values for the Bayesian PL mixtures with a varying number G of components. For each selection criterion, the optimal choice of the number of components corresponds to the minimum value (*in bold*).

Full + Partial orderings

G	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	$BICM_1$	$BICM_2$	BIC	$p_{B(1)}$	$p_{B(2)}$
1	103204.59	103204.65	103208.57	103208.71	103235.66	103235.73	103235.19	0.000	0.000
2	100772.97	100772.87	100781.63	100781.44	100838.40	100838.30	100842.44	0.000	0.610
3	100591.84	100591.89	100603.00	100603.10	100677.58	100677.62	100704.56	0.004	0.493
4	100436.20	100445.60	100443.54	100462.34	100573.58	100582.98	100604.78	0.298	0.411
5	100396.98	100401.44	100413.46	100422.39	**100561.56**	**100566.03**	**100595.51**	0.343	0.523
6	100360.59	100369.50	100377.16	100394.99	100564.33	100573.24	100607.17	0.375	0.503
7	100336.08	100344.30	100361.35	100377.81	100600.44	100608.66	100613.47	0.233	0.536
8	100341.12	100347.07	100381.82	100393.71	100703.71	100709.65	100635.88	0.390	0.492
9	100341.55	100350.84	100390.59	100409.18	100796.84	100806.13	100667.85	0.426	0.531
10	**100317.68**	100346.31	**100358.35**	100415.63	100876.22	100904.86	100708.95	0.471	0.528
11	100321.54	**100314.49**	100376.02	**100361.92**	100677.10	100670.05	100733.43	0.382	0.531
12	100340.38	100349.80	100408.73	100427.57	100944.37	100953.79	100772.76	0.440	0.521

Table SM-9.

APA election data (separate analysis for subsets of ballots with the same number of ranked candidates) Model selection criteria and posterior predictive p values for the Bayesian PL mixtures with a varying number G of components. For each selection criterion, the optimal choice of the number of components corresponds to the minimum value (*in bold*).

Top-2 orderings

G	DIC_1	DIC_2	BPIC_1	BPIC_2	BICM_1	BICM_2	BIC	$p_{B(1)}$	$p_{B(2)}$
1	14255.36	14255.24	14259.33	14259.09	14277.59	14277.47	14278.66	0.000	0.049
2	13427.79	13427.99	**13436.89**	**13437.29**	**13481.98**	**13482.18**	**13479.88**	0.295	0.502
3	13427.70	13426.86	13443.00	13441.33	13510.92	13510.08	13506.41	0.379	0.530
4	13434.10	13427.37	13458.77	13445.32	13531.62	13524.90	13533.12	0.410	0.530
5	13433.24	13425.80	13458.63	13443.74	13530.04	13522.60	13569.87	0.443	0.531
6	13431.30	13426.03	13455.70	13445.16	13537.16	13531.89	13608.96	0.455	0.527
7	13429.54	13425.14	13453.09	13444.29	13536.39	13531.99	13647.93	0.463	0.522
8	13429.62	13425.23	13453.22	13444.45	13536.84	13532.45	13686.96	0.472	0.526
9	13428.02	**13423.42**	13450.81	13441.60	13529.06	13524.45	13726.03	0.478	0.525
10	13427.65	13424.41	13450.28	13443.79	13537.01	13533.76	13765.02	0.479	0.522
11	13427.58	13423.66	13450.16	13442.32	13532.07	13528.15	13804.08	0.476	0.525
12	**13426.88**	13423.49	13449.11	13442.32	13532.91	13529.51	13843.13	0.478	0.513

Top-3 orderings

G	DIC_1	DIC_2	BPIC_1	BPIC_2	BICM_1	BICM_2	BIC	$p_{B(1)}$	$p_{B(2)}$
1	17147.88	17147.84	17151.92	17151.83	17170.41	17170.37	17170.43	0.000	0.180
2	16732.01	16733.01	16741.62	16743.63	**16793.05**	**16794.05**	**16781.65**	0.262	0.468
3	**16717.06**	16717.76	**16731.79**	**16733.19**	16805.00	16805.69	16794.74	0.278	0.529
4	16717.27	16719.43	16742.81	16747.13	16876.03	16878.19	16811.62	0.440	0.530
5	16718.74	16720.82	16752.55	16756.70	16923.69	16925.77	16834.81	0.479	0.523
6	16723.28	16713.24	16762.76	16742.70	16879.76	16869.73	16866.26	0.483	0.511
7	16725.94	16716.31	16769.12	16749.85	16905.96	16896.33	16899.80	0.477	0.516
8	16726.43	16715.82	16771.67	16750.46	16911.66	16901.06	16934.43	0.468	0.509
9	16729.16	16715.99	16776.54	16750.20	16909.40	16896.23	16971.16	0.497	0.509
10	16735.15	16716.82	16788.59	16751.92	16915.26	16896.92	17003.31	0.491	0.511
11	16736.19	16718.88	16790.71	16756.09	16929.25	16911.94	17040.45	0.478	0.507
12	16735.28	**16710.82**	16790.19	16741.29	16883.04	16858.58	17077.01	0.497	0.505

Top-4 (full) orderings

G	DIC_1	DIC_2	BPIC_1	BPIC_2	BICM_1	BICM_2	BIC	$p_{B(1)}$	$p_{B(2)}$
1	54812.60	54812.60	54816.60	54816.61	54839.29	54839.30	54839.21	0.000	0.000
2	53696.16	53695.65	53705.49	53704.48	53754.42	53753.91	53755.38	0.000	0.639
3	53576.87	53574.99	53591.48	53587.73	53659.78	53657.91	53668.81	0.043	0.655
4	53477.70	53477.33	53494.38	53493.63	53585.83	53585.46	**53608.79**	0.179	0.478
5	53458.70	53454.30	53479.35	53470.54	**53562.39**	**53557.98**	53625.13	0.183	0.470
6	53433.51	53439.25	53460.29	53471.77	53655.67	53661.42	53630.95	0.462	0.479
7	**53412.08**	53437.63	**53445.60**	53496.70	53830.73	53856.28	53639.31	0.503	0.512
8	53420.25	**53416.43**	53464.62	**53456.97**	53686.23	53682.40	53669.06	0.440	0.436
9	53449.15	53499.96	53512.83	53614.45	54261.90	54312.71	53702.59	0.413	0.489
10	53415.29	53443.76	53466.78	53523.72	53975.90	54004.37	53736.39	0.481	0.553
11	53437.74	53446.27	53503.70	53520.77	53942.07	53950.61	53773.17	0.403	0.462
12	53424.85	53438.45	53488.02	53515.22	53949.36	53962.97	53809.15	0.455	0.519

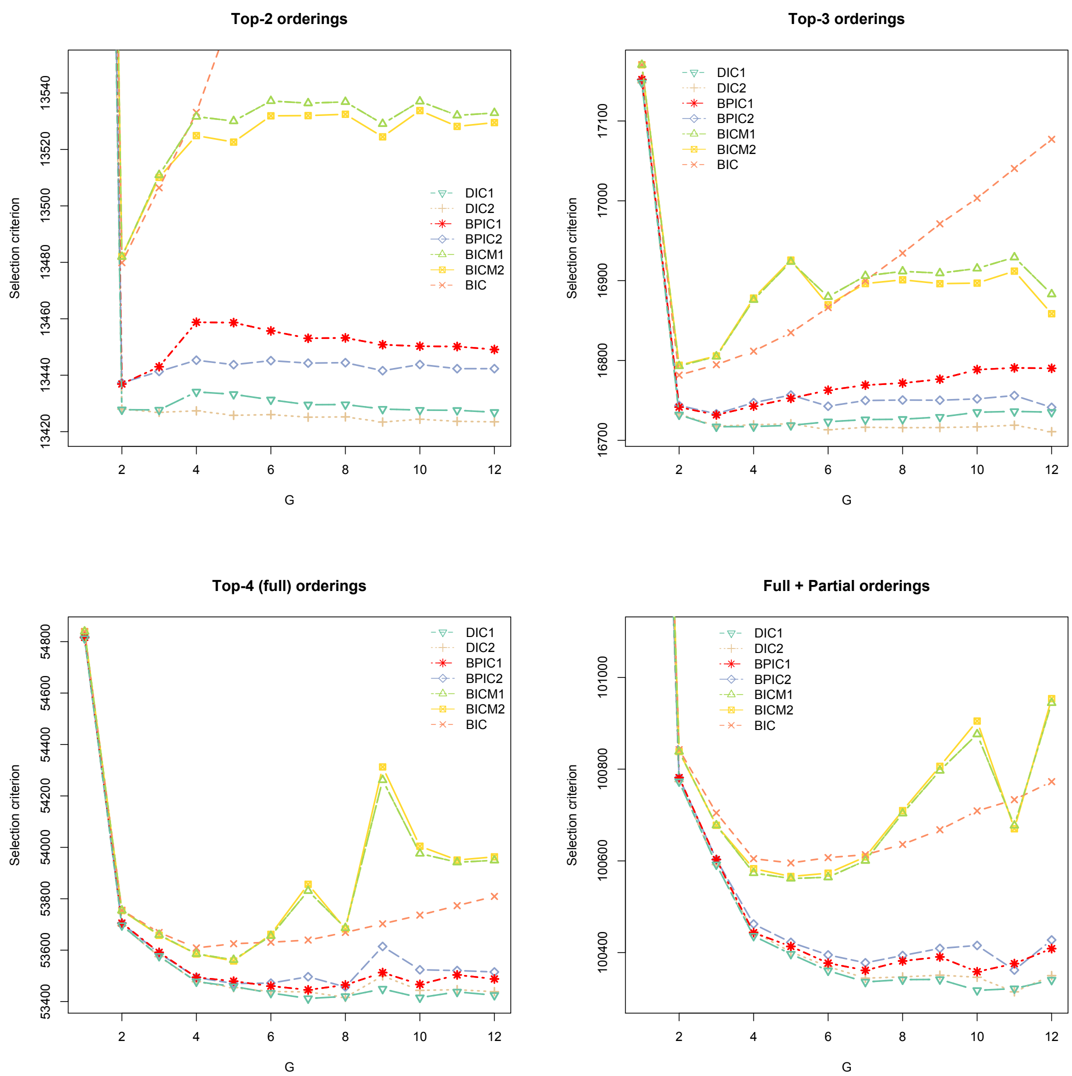

Figure SM-2.

APA election data (global and separate analysis for subsets of ballots with the same number of ranked candidates) Model selection criteria for the Bayesian PL mixtures with a varying number G of components. For each selection criterion, the optimal choice of the number of components corresponds to the minimum value.

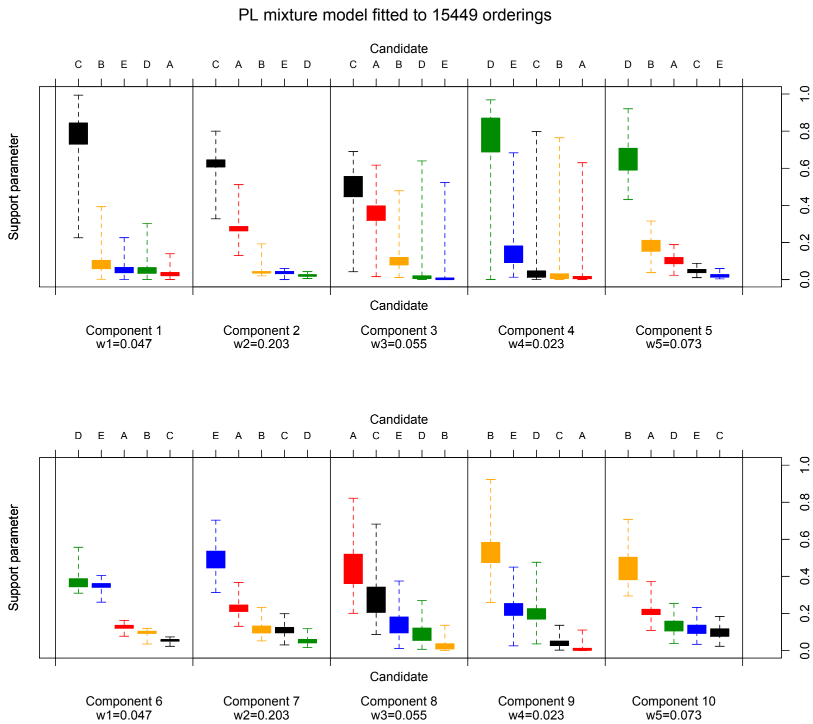

Figure SM-3.

APA election data (global analysis with 15449 full and partial orderings) Optimal Bayesian 10-component PL mixture model fitted to the entire data set.

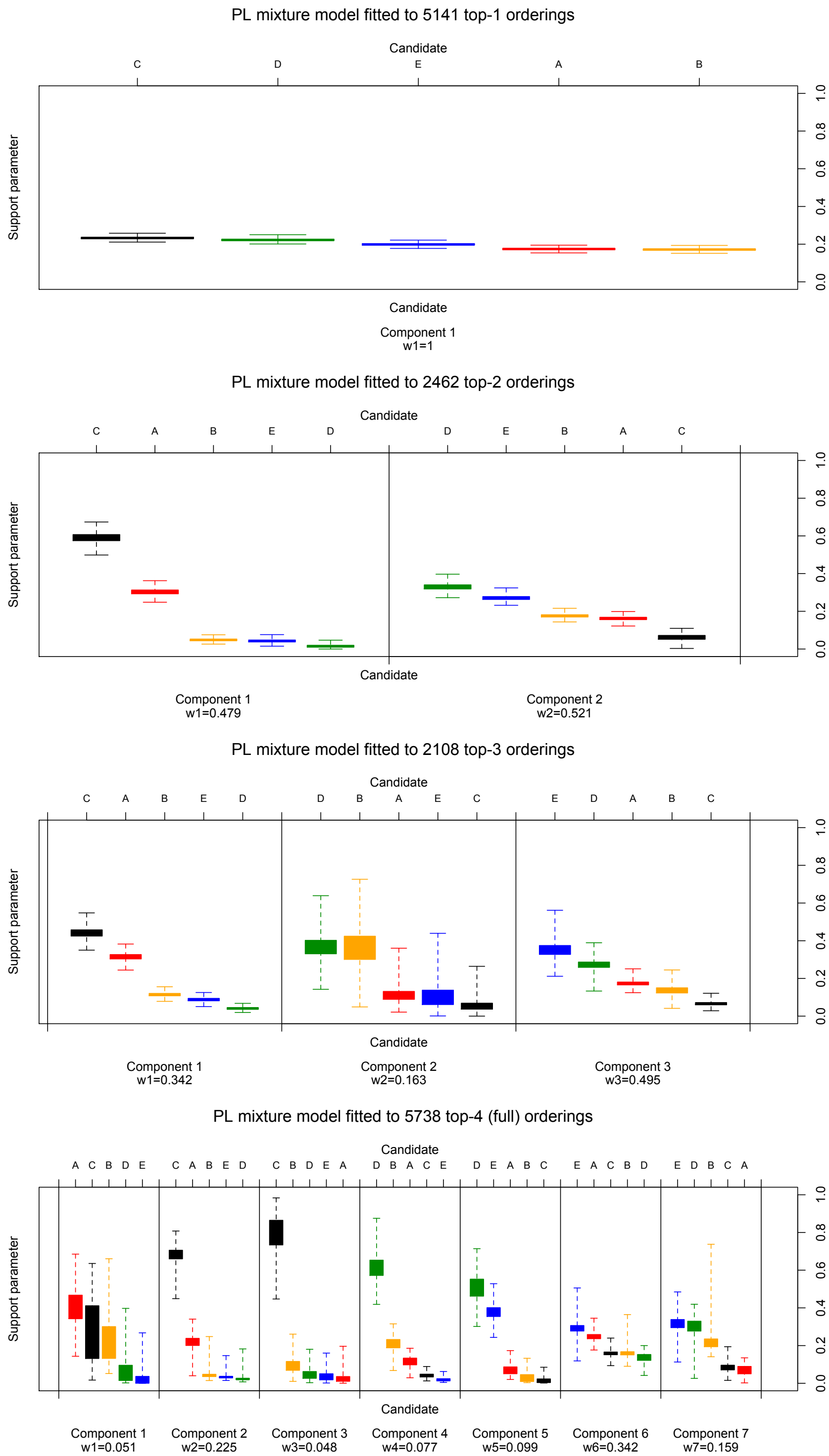

Figure SM-4.

APA election data (separate analysis for subsets of ballots with the same number of ranked candidates) Optimal Bayesian PL mixtures fitted to subsets of partial orderings with the same number m of ranked items.