Multivariate Cluster Point Process Model: Parent Location Improves Inference for Complex Biofilm Image Data - Supplemetary Materials

December 8, 2023

A Neyman-Scott process

A Neyman-Scott process is a point process used for modeling parent-offspring clustering. In the simplest setting, consider the parent process C to be a homogeneous Poisson point process with intensity λ^C. For each observation location $c \in C$, the cluster of offspring Y_c is an independent Poisson process with intensity $\alpha k(\cdot - c, h)$, where $k(\cdot - c, h)$ is a probability distribution function parameterized by h that determines the spread and distribution of the offspring locations around the parent c, and $\alpha > 0$ is the expected number of offspring per cluster. The Neyman-Scott process Y is the union of all these offspring cluster processes, namely, $Y = \bigcup_{c \in C} Y_c$. Further details can be found in Illian et al. (2008) and Chiu et al. (2013), for example.

B Images of the taxa from the human dental plaque biofilm data not visualized in the image included in the main text

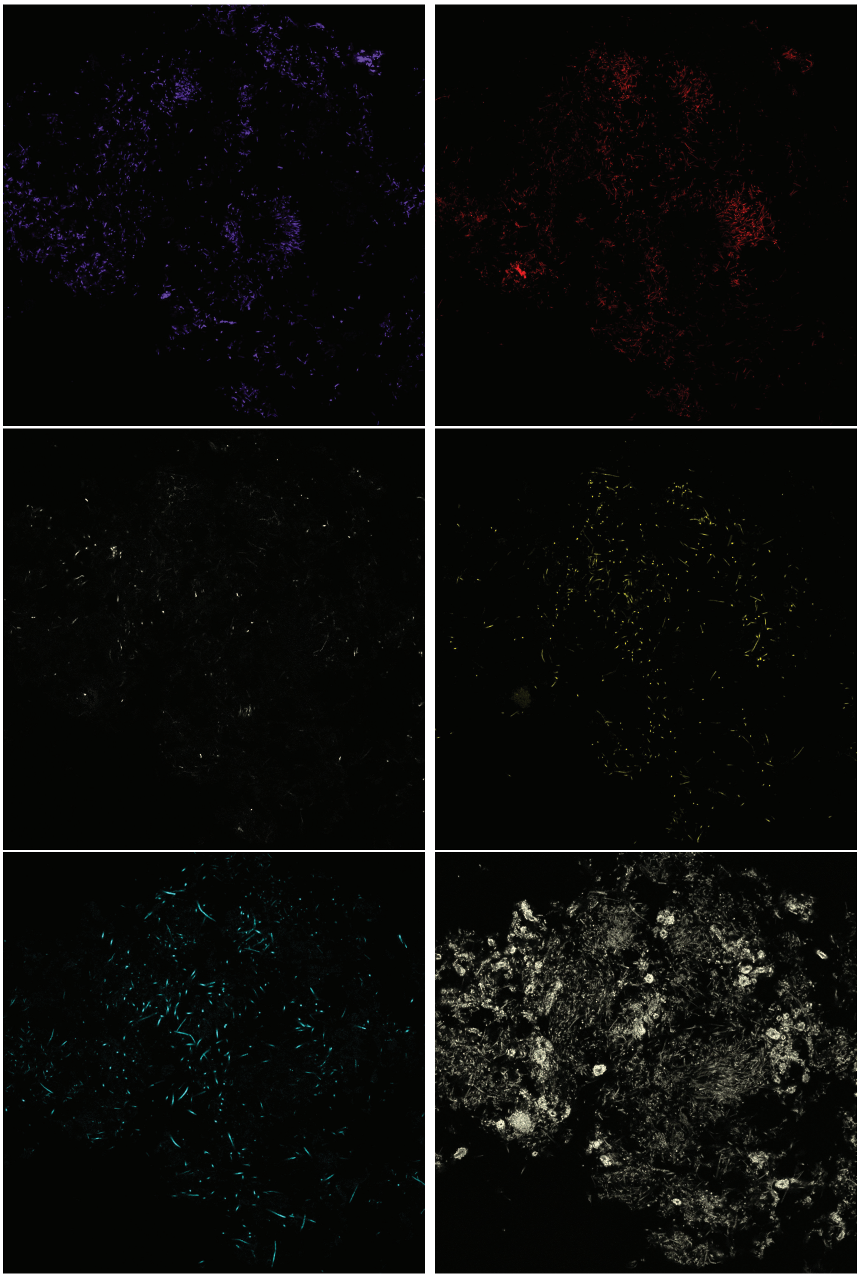

Figure S.1: RGB images of *Neisseriaceae* (top left), *Capnocytophaga* (top right), *Actinomyces* (middle left), *Fusobacterium* (middle right), *Leptotrichia* (bottom left) and *Bacterium* (bottom right) in the dental plaque biofilm sample. *Bacterium* denotes a probe for all oral bacteria. It is used for methodologic purposes to evaluate the completeness of the set of specific probes. Hence, it is omitted from analysis of community spatial structure. The genera shown here were modeled as homogeneous Poisson process in the data analysis.

C Computational details of the sampling algorithm

We use a Markov chain Monte Carlo (MCMC) method to draw samples from the joint posterior distribution of $\boldsymbol{\theta}$. In the MCMC scheme, parameters are updated either by exploiting conjugacies inherent to the proposed model or by using a Metropolis-Hastings algorithm.

C.1 Updating parameters associated with offspring densities

Let $\boldsymbol{\theta}^{-(\alpha)}$ denote a set of parameters $\boldsymbol{\theta}$ with α removed. The full conditional distribution for $\alpha_l\,,\ l = p+1,\ldots,p+q$ is

$$\alpha_l|\boldsymbol{\theta}^{-(\alpha_l)} \sim \ \text{Gamma}(a_Y + n_l, b_Y + \sum_{\mathbf{c}_l \in C_l} \int\limits_W k_l(\mathbf{u} - \mathbf{c}_l, h_l)\ d\mathbf{u}),$$

where n_l is the number of observations of taxon l in the window.

C.2 Updating intensity parameters in homogeneous Poisson processes

Posterior conjugacy is also achieved in the full conditional distributions of intensity parameters, $\lambda_v^C\,,\ v = 1,\ldots,p$ and $\lambda_j\,,\ j = p+q+1,\ldots,m$, which are given by

$$\lambda_v^C|\boldsymbol{\theta}^{-(\lambda_v^C)} \sim \ \text{Gamma}(a_C + n_v,\ b_C + |\mathcal{W}|)\,,\ v = 1,\ldots,p;\ \text{and}$$

$$\lambda_j|\boldsymbol{\theta}^{-(\lambda_j)} \sim \ \text{Gamma}(a + n_j,\ b + |\mathcal{W}|)\,,\ j = p+q+1,\ldots,m;$$

where n_v and n_j are the numbers of observations for taxon v and taxon j within the window, respectively.

C.3 Updating bandwidth parameters

Since the full conditionals of the bandwidth parameters do not have standard forms, we use a random work Metropolis-Hastings step to update each of $h_l\,,\ l = 1,\ldots,p$. Denote $h_j^{(t)}$ the sample for $h_j\,,\ j = p+1.\ldots,p+q$ from iteration t. For iteration $(t+1)$, we propose a candidate sample h_j^* as a random draw from $N(h_j^{(t)}, \sigma^2_{prop})$, where σ^2_{prop} is the prespecified variance of the proposal density. The corresponding acceptance ratio computes to

$$R = \frac{\exp\left(-\alpha_l \sum_{\mathbf{c}_l \in C_l} \int_W k(\mathbf{u} - \mathbf{c}_l, h_j^*)\ d\mathbf{u}\right) \prod_{\mathbf{y} \in Y_l} \left(\sum_{\mathbf{c}_l \in C_l} \int_W k(\mathbf{u} - \mathbf{c}_l, h_j^*)\right) \exp\left(-h_j^{*2}/2\sigma^2\right) \mathbb{I}(h_j^* > 0)}{\exp\left(-\alpha_l \sum_{\mathbf{c}_l \in C_l} \int_W k(\mathbf{u} - \mathbf{c}_l, h_j^{(t)})\ d\mathbf{u}\right) \prod_{\mathbf{y} \in Y_l} \left(\sum_{\mathbf{c}_l \in C_l} \int_W k(\mathbf{u} - \mathbf{c}_l, h_j^{(t)})\right) \exp\left(-h_j^{(t)2}/2\sigma^2\right)}.$$

Then, we accept the proposed candidate h_j^* as $h_j^{(t+1)}$ with probability $\min\{R, 1\}$ or keep $h_j^{(t+1)} = h_j^{(t)}$.

D Additional Tables from the simulation study

Here, we present additional details regarding the simulation scenarios (Table S.1) and results for the scenarios that included a taxon unrelated to the parent-offspring-type configuratons of interest (Table S.2). The presence of an unrelated taxon (Table S.2) did not meaningfully affect the results (Table 1). Specifically, the multivariate cluster point process (MCPP) performed better than the Neyman-Scott process (NSP) implementation in all aspects. The NSP often failed to converge, especially in scenarios where the bandwidth parameter was large.

Scenario	Unrelated taxon	Offspring density	Bandwidth
1	Absent	Sparse	Low
2	Absent	Sparse	High
3	Absent	Dense	Low
4	Absent	Dense	High
5	Absent	Mixed	Low
6	Absent	Mixed	High
7	Present	Sparse	Low
8	Present	Sparse	High
9	Present	Dense	Low
10	Present	Dense	High
11	Present	Mixed	Low
12	Present	Mixed	High

Table S.1: A summary of twelve simulation scenarios considered in Section 4. The offspring density is controlled by setting $(\alpha_2, \alpha_3) = (1.5, 1)$ for 'Sparse', $(4, 3)$ for 'Dense' and $(4, 1)$ for 'Mixed' densities. Bandwidth 'Low' sets $(h_2, h_3) = (0.01, 0.02)$ and 'High' to $(0.1, 0.01)$. The setting "Unrelated taxon" refers to whether there exists a taxon in the data spatially unrelated to the multilayered arrangement.

		True	MCPP			NSP		
Scenario		Value	EST	SD	SD$_{EST}$	EST	SE	%F
7	α_2	1.50	1.53	0.10	0.10	1.46	0.33	2
	α_3	1.00	1.02	0.08	0.09	3.31	20.25	
	h_2	0.01	0.01	< 0.01	< 0.01	0.01	< 0.01	
	h_3	0.02	0.02	< 0.01	< 0.01	0.04	0.09	
	λ_1^C	150.00	161.06	12.91	12.20	171.35	34.72	
8	α_2	1.50	1.48	0.11	0.09	198.46	283.69	36
	α_3	1.00	1.02	0.08	0.09	0.98	0.28	
	h_2	0.10	0.08	0.01	0.01	10.30	28.72	
	h_3	0.01	0.01	< 0.01	< 0.01	0.01	< 0.01	
	λ_1^C	150.00	160.25	12.86	12.57	939.70	2777.40	
9	α_2	4.00	4.02	0.14	0.15	8.77	48.78	0
	α_3	3.00	3.05	0.13	0.13	2.91	0.77	
	h_2	0.01	0.01	< 0.01	< 0.01	0.02	0.08	
	h_3	0.02	0.02	< 0.01	< 0.01	0.02	< 0.01	
	λ_1^C	150.00	202.78	14.38	14.29	208.52	39.37	
10	α_2	4.00	4.00	0.17	0.17	613.34	569.20	48
	α_3	3.00	3.02	0.13	0.14	2.93	0.53	
	h_2	0.10	0.09	0.01	0.01	1.15	0.64	
	h_3	0.01	0.01	< 0.01	< 0.01	0.01	< 0.01	
	λ_1^C	150.00	200.49	14.26	14.48	13.30	28.78	
11	α_2	4.00	4.05	0.15	0.14	18.45	87.48	0
	α_3	1.00	1.00	0.07	0.08	2.05	10.04	
	h_2	0.01	0.01	< 0.01	< 0.01	0.03	0.11	
	h_3	0.02	0.02	< 0.01	< 0.01	0.47	4.41	
	λ_1^C	150.00	201.50	14.37	14.09	203.36	53.30	
12	α_2	4.00	4.02	0.17	0.17	547.61	553.61	70
	α_3	1.00	1.02	0.07	0.07	0.97	0.24	
	h_2	0.10	0.09	0.01	0.01	1.04	0.82	
	h_3	0.01	0.01	< 0.01	< 0.01	0.01	< 0.01	
	λ_1^C	150.00	199.05	14.23	15.95	26.62	41.46	

Table S.2: The true value, estimates (EST), and uncertainty measures for the offspring density (α_2, α_3), bandwidth (h_2, h_3), and parent process (λ_1^C) parameters from the MCPP and NSP analyses in the last six simulated scenarios (those with an unrelated taxon). For the MCPP model, the estimates are the posterior means averaged over different datasets, the SD is computed by averaging the posterior standard deviation over different datasets, and the SD$_{EST}$ is computed as the standard deviation of the estimates over the datasets. For the NSP model, the estimates are the outputs of the minimum contrast method, and SE is calculated similarly by using these estimates. The SD for the NSP model is not computed, as the method does not provide an uncertainty measure. The last column (%F) refers to the percentage of datasets in which the NSP model failed to converge for a given scenario.

E Sensitivity analyses regarding choice of prior for the bandwidth parameters

As part of the simulation study described in Section 4, we also evaluated the sensitivity of the MCPP method to choice of prior distribution for the bandwidth parameters. Specifically, we considered four different prior distributions, namely 1) half-normal, 2) uniform, 3) log-normal with a flat tail and high variance and 4) log-normal with a slim tail and higher peak. For the uniform prior, the lower and upper bounds were taken to be 0 and 0.2, respectively. Both the log-normal priors had $\mu = \log 0.05$; the flat-tailed prior had $\sigma = 1$, and the high-peaked prior had $\sigma = 0.1$ as the hyperparameter. The hyperparameter setting for the half-normal prior was the same as in Section 4. We compared performance of the MCPP for the different prior distributions in Scenario 5 and 6 (Table 1): both scenarios considered mixed offspring density (α_2=4 and α_3=1), one had low bandwidth (h_2=0.01 and h_3=0.02), and the other had high bandwidth (h_2=0.1 and h_3=0.01).

We report the mean absolute percentage bias for estimating the corresponding parameters in the two scenarios for the four different prior settings: i) Half-normal, ii) Uniform, iii) a flat Log-normal, and iv) a tight Log-normal. The half-normal prior-based MCPP model performed the best, and the performance was similar to that for the original model. When the true bandwidth was low, all the models—irrespective of prior choice—generally performed well and similarly to each other, with almost all biases $< 8\%$. Differences in performance emerged when the true bandwidth was high, where the analyses with tighter priors produced much less biased estimates ($< 10\%$ except in one instance) than the analyses with flatter priors (4-137%; Table S.3). However, using an informative log-normal prior backfired even for the low-bandwidth scenario when the offspring density was also low, as for the second offspring process (∼20-25%).

				Log-normal	Log-normal
		Half-normal	Uniform	(flat)	(tight)
Low bandwidth	α_2	0.03	0.03	0.03	0.03
	α_3	0.07	0.07	0.07	0.07
	h_2	0.02	0.02	0.02	0.05
	h_3	0.04	0.04	0.04	0.24
	λ_1^C	0.06	0.06	0.06	0.06
High bandwidth	α_2	0.03	0.31	0.52	0.03
	α_3	0.05	0.05	0.05	0.05
	h_2	0.09	0.99	1.37	0.10
	h_3	0.03	0.03	0.03	0.20
	λ_1^C	0.06	0.06	0.06	0.06

Table S.3: Results of MCPP based analysis of simulated data, comparing different choice of priors for the bandwidth parameters. The true values for the offspring densities were α_2=4 and α_3=1. The true values for the bandwidth parameters were h_2=0.01 and h_3=0.02 under low bandwidth and h_2=0.1 and h_3=0.01 under high bandwidth. The parent process is denoted λ^C. Results are presented as mean absolute percentage bias of the estimated parameter values based on posterior means of each of the 100 simulated datasets. There were no other taxa unrelated to these multi-layered arrangements.

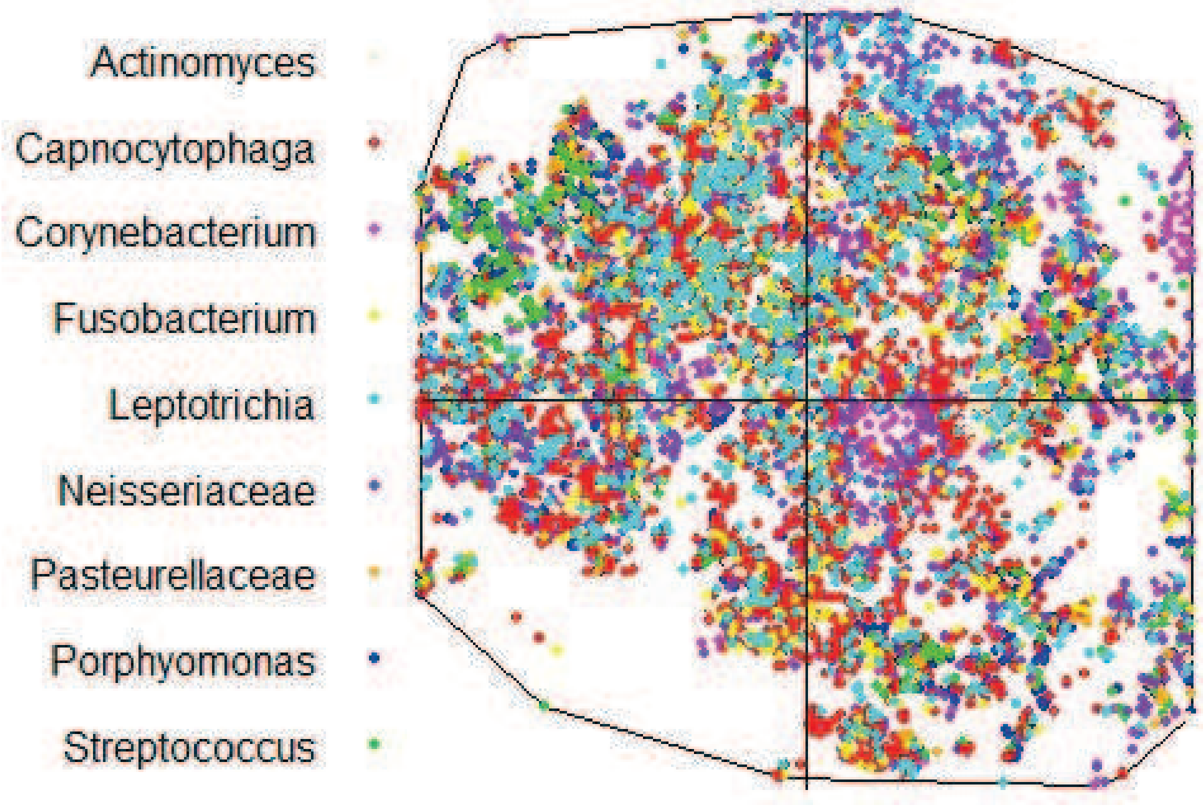

Figure S.2: Division of the dental plaque sample image into the first (bottom left), second (top left), third (bottom right) and fourth (top right) quadrants. Black space has been removed.

F Additional Figures and Tables for the Analysis of Human Microbiome Biofilm Image Data

Here, we present the subsetted image of the human microbiome biofilm data (Figure S.2), with abundances that differ in the four quadrants (Table S.4). The estimates for the intensity functions for the five taxa (*Neisseriaceae*, *Capnocytophaga*, *Actinomyces*, *Fusobacterium*, *Leptotrichia*) that have no visible spatial relationship with the parent-offspring-type configurations also varied across the quadrants (Table S.5). K-functions for the whole and subsetted analyses (Figures S.3 and S.4 through S.7) also varied noticably by quadrant. We also present the DIC estimates for the different models explored in Section 5.2. (Table S.6) which shows that the models with the *Fusobacterium* and *Leptotrichia* as an additional parent-offspring pair is a better fit to the data than the original model, while models fitting *Streptococcus* around *Fusobacterium* do not fit the data well.

References

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*, volume 70. John Wiley & Sons.

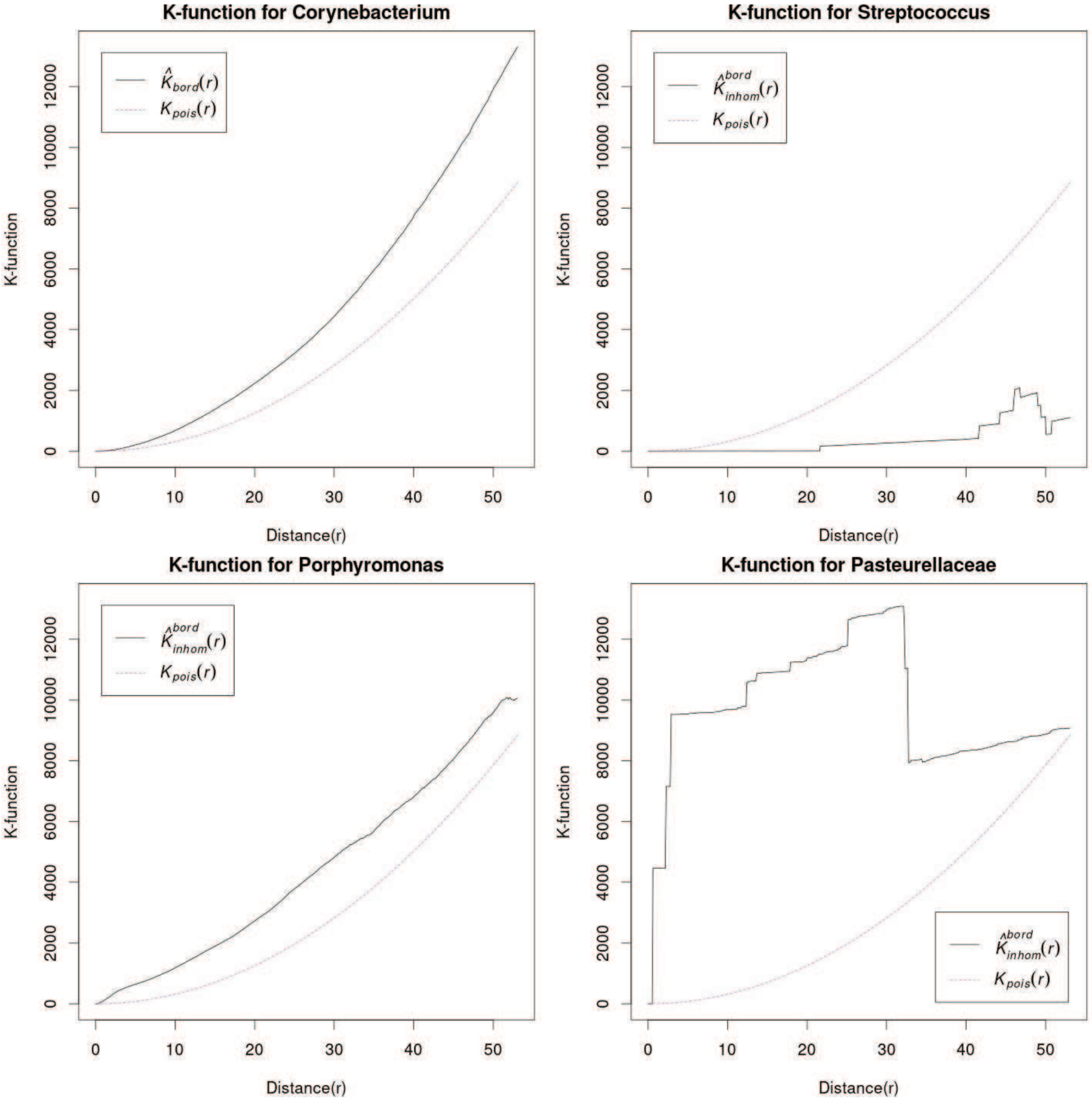

Figure S.3: Border corrected K-functions $\left(\hat{K}_{\text{bord}}(r) \text{ or } \hat{K}^{\text{bord}}_{\text{inhom}}(r)\right)$ for the processes corresponding to *Corynebacterium* (top left), *Streptococcus* (top right), *Porphyromonas* (bottom left) and *Pasteurellaceae* (bottom right) in comparison to that of a homogeneous Poisson process $\left(\hat{K}_{\text{pois}}(r)\right)$.

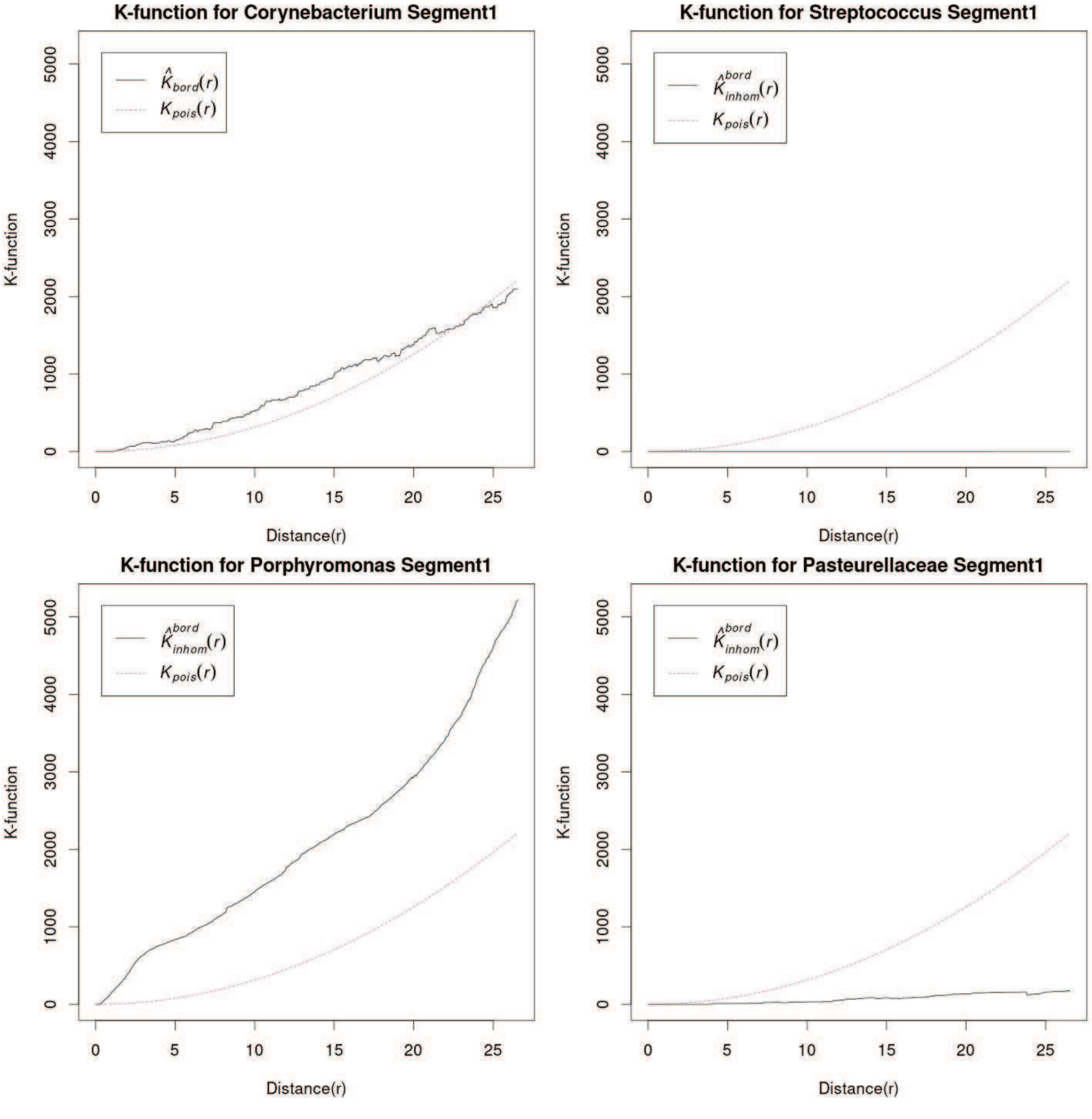

Figure S.4: Border corrected K-functions $\left(\hat{K}_{\text{bord}}(r) \text{ or } \hat{K}^{\text{bord}}_{\text{inhom}}(r)\right)$ for the processes corresponding to *Corynebacterium* (top left), *Streptococcus* (top right), *Porphyromonas* (bottom left) and *Pasteurellaceae* (bottom right) in comparison to that of a homogeneous Poisson process $\left(\hat{K}_{\text{pois}}(r)\right)$ in Segment 1

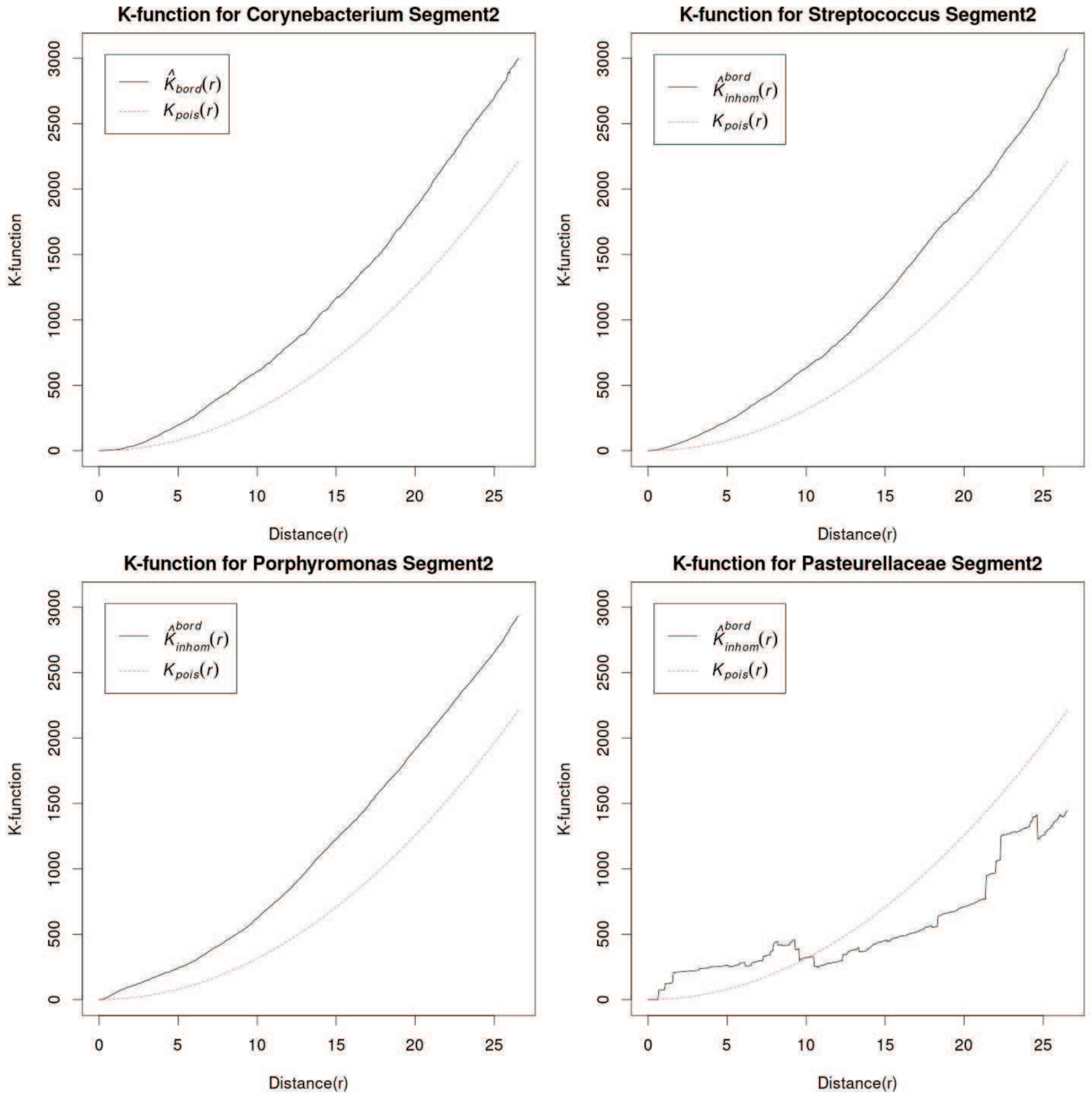

Figure S.5: Border corrected K-functions $\left(\hat{K}_{\text{bord}}(r) \text{ or } \hat{K}^{\text{bord}}_{\text{inhom}}(r)\right)$ for the processes corresponding to *Corynebacterium* (top left), *Streptococcus* (top right), *Porphyromonas* (bottom left) and *Pasteurellaceae* (bottom right) in comparison to that of a homogeneous Poisson process $\left(\hat{K}_{\text{pois}}(r)\right)$ in Segment 2

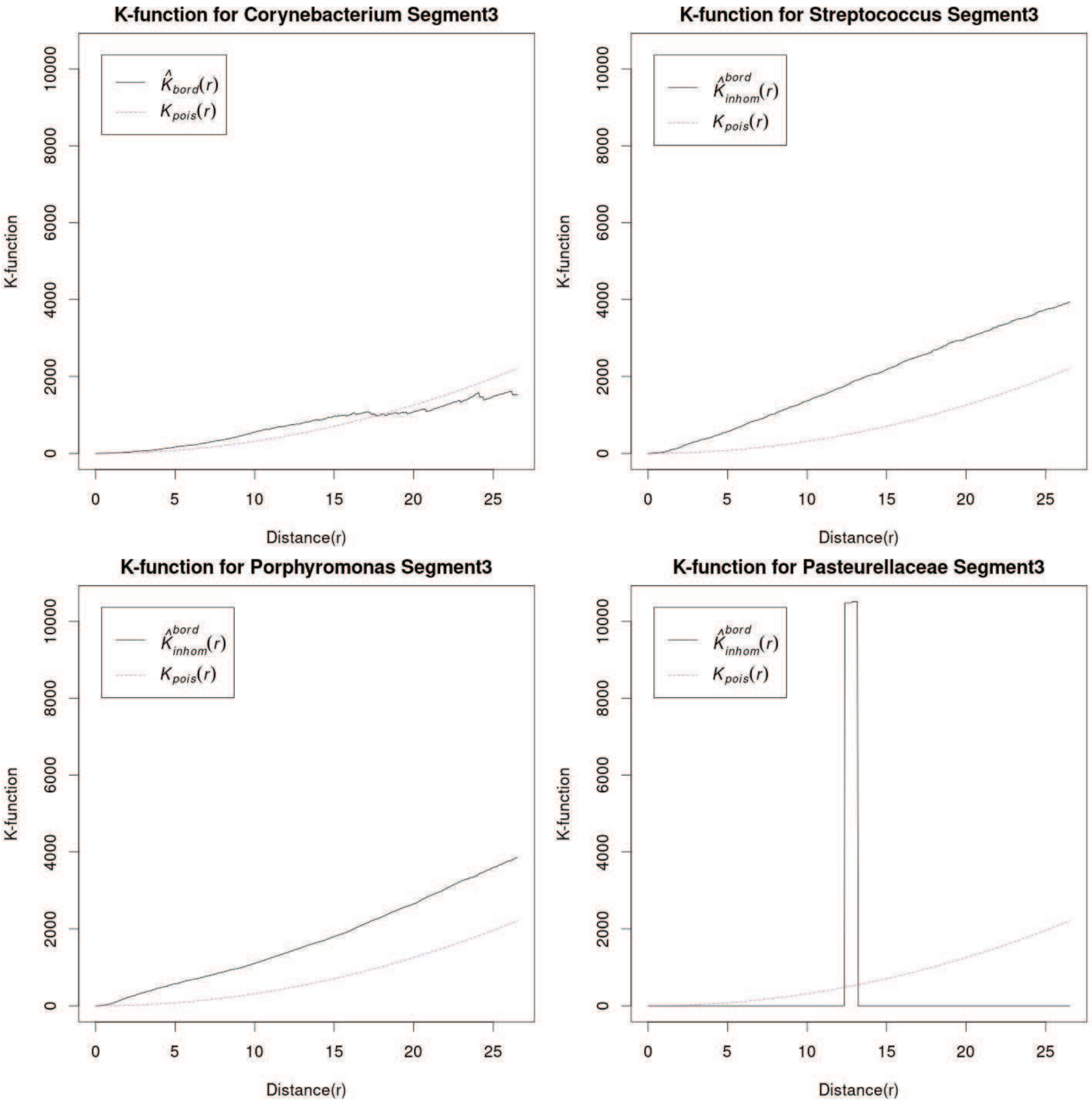

Figure S.6: Border corrected K-functions $\left(\hat{K}_{\text{bord}}(r) \text{ or } \hat{K}^{\text{bord}}_{\text{inhom}}(r)\right)$ for the processes corresponding to *Corynebacterium* (top left), *Streptococcus* (top right), *Porphyromonas* (bottom left) and *Pasteurellaceae* (bottom right) in comparison to that of a homogeneous Poisson process $\left(\hat{K}_{\text{pois}}(r)\right)$ in Segment 3

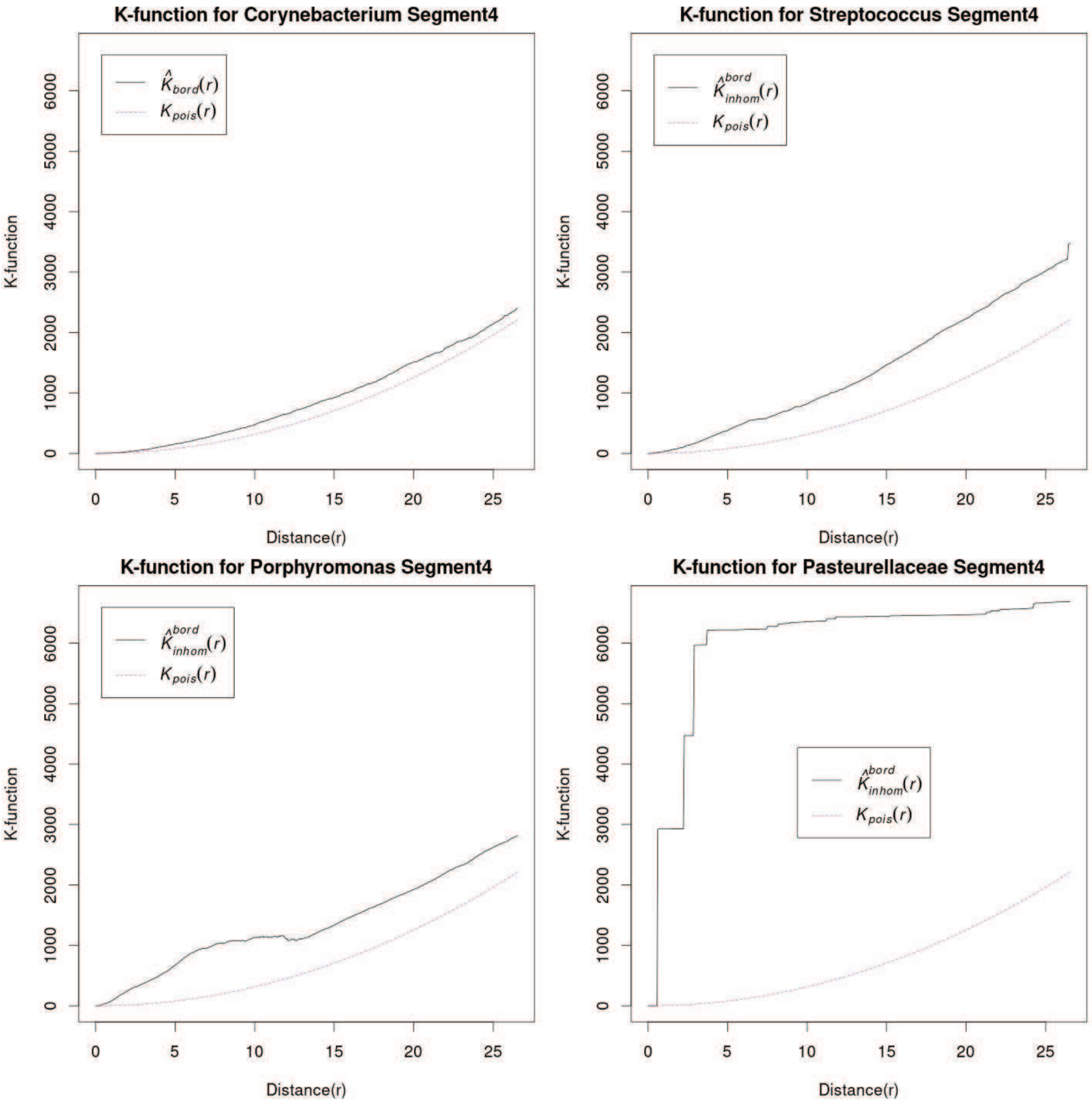

Figure S.7: Border corrected K-functions $\big(\hat{K}_{\text{bord}}(r)$ or $\hat{K}^{\text{bord}}_{\text{inhom}}(r)\big)$ for the processes corresponding to *Corynebacterium* (top left), *Streptococcus* (top right), *Porphyromonas* (bottom left) and *Pasteurellaceae* (bottom right) in comparison to that of a homogeneous Poisson process $\big(\hat{K}_{\text{pois}}(r)\big)$ in Segment 4

Taxon	Quadrant				Total
	I	II	III	IV	
Actinomyces	119	280	154	223	776
Capnocytophaga	512	755	574	573	2414
Corynebacterium	58	219	186	245	708
Fusobacterium	92	250	141	173	656
Leptotrichia	191	411	234	339	1175
Neisseriaceae	339	479	402	491	1711
Pasteurellaceae	53	130	76	106	365
Porphyromonas	227	525	269	420	1441
Streptococcus	98	379	163	249	889

Table S.4: The abundance (counts) of bacterial taxa of interest in the human dental plaque sample image data and its four subdivided quadrants.

	λ_5	λ_6	λ_7	λ_8	λ_9
Full Image	0.04	0.06	0.02	0.02	0.03
Segment 1	0.04	0.06	0.01	0.01	0.02
Segment 2	0.05	0.07	0.03	0.02	0.04
Segment 3	0.04	0.05	0.01	0.01	0.02
Segment 4	0.05	0.06	0.02	0.02	0.03

Table S.5: The posterior means of parameters associated with *Neisseriaceae* (λ_5), *Capnocytophaga* (λ_6), *Actinomyces* (λ_7), *Fusobacterium* (λ_8) and *Leptotrichia* (λ_9) obtained by applying the proposed MCPP method on the entire image and on each of the four quadrants of the dental plaque sample image. All results are rounded to two decimal places. The posterior standard deviations were all smaller than 0.01 and are not reported separately.

Identifier	Parent-Offspring Realtions Present	DIC
1	$C \rightarrow SPo \vdots S \rightarrow Pa$	124985.4
2	$C \rightarrow Po \vdots S \rightarrow Pa \vdots F \rightarrow S$	125531.1
3	$C \rightarrow SPo \vdots F \rightarrow S \vdots S \rightarrow Pa$	134007.1
4	$C \rightarrow SPo \vdots S \rightarrow Pa \vdots F \rightarrow L$	124317.0
5	$C \rightarrow SPo \vdots S \rightarrow Pa \vdots L \rightarrow F$	124303.5
6	$C \rightarrow SPo \vdots S \rightarrow Pa \vdots F \rightarrow LS$	133357.2
7	$C \rightarrow SPo \vdots S \rightarrow Pa \vdots L \rightarrow F \vdots F \rightarrow S$	133345.2

Table S.6: Different parent-offspring relationships explored in different models and their DIC. The monikers C, S, Po, Pa, F and L are used for *Corynebacterium*, *Streptococcus*, *Porphyromonas*, *Pasteurellaceae*, *Fusobacterium* and *Leptotrichia*