Supplementary material for Centered Partition Processes: Informative Priors for Clustering

Sally Paganin, Amy H. Herring, Andrew F. Olshan, David B. Dunson and The National Birth Defects Prevention Study

1 Simulation study

In Section 5.3 we presented a simulation study tailored on the motivating application to the NBDPS data, characterized by categorical variables. However our proposed CP process prior is not limited to this kind of applications. We consider again the simulation scenario in Section 5.3, with a number of dataset $N = 12$ equally partitioned in 4 groups, but instead assuming continuous type of data.

Since this type of data naturally contains more information, we consider a more challenging scenario by assuming $p = 40$ explanatory variables and a more heterogeneous number of observations across defects, as reported in Table 1. We generated most of coefficients $\beta_{i1}, \dots, \beta_{i40}$ from a random noise centered in 0 while fixing the significant shared coefficient as reported in Table 1. In this case we generated responses y_{ij} independently from a normal distribution with mean $\mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i}$ and variance 1 for $i = 1, \dots, 12$ and $j = 1, \dots, n_i$. We simplify the model in equation (17) of Section 5.2, considering a linear regression for responses y_{ij} while assuming the same prior settings as in Section 5.3. We estimated the model via MCMC using a simplified version of Algorithm 3, since in this case the Pólya-gamma data augmentation is unnecessary and one can exploit normal conjugacy. We run the MCMC for $5,000$ using a burn-in of $1,000$, with results reported in Figures 1-2.

We obtain quite similar results with respect to the scenario with categorical data in terms of clustering, while performances on coefficients estimation suffer more of the mis-

	Subgroups numerosities	Shared coefficients
Group 1	$\{n_1, n_2, n_3\} = \{5, 20, 70\}$	$\{\beta_1, \dots, \beta_5\} = 2$ $\{\beta_6, \dots, \beta_{10}\} = -2$
Group 2	$\{n_4, n_5, n_6\} = \{10, 40, 50\}$	$\{\beta_6, \dots, \beta_{10}\} = -2$ $\{\beta_{11}, \beta_{12}, \beta_{13}\} = 1.8$
Group 3	$\{n_7, n_8, n_9\} = \{10, 50, 70\}$	$\{\beta_{11}, \dots, \beta_{15}\} = -1.8$ $\{\beta_{16}, \dots, \beta_{20}\} = 1.8$
Group 4	$\{n_{10}, n_{11}, n_{12}\} = \{40, 10, 80\}$	$\{\beta_1, \dots, \beta_5\} = 2$ $\{\beta_{15}, \dots, \beta_{19}\} = 1.8$

Table 1: Setting of the simulation example.

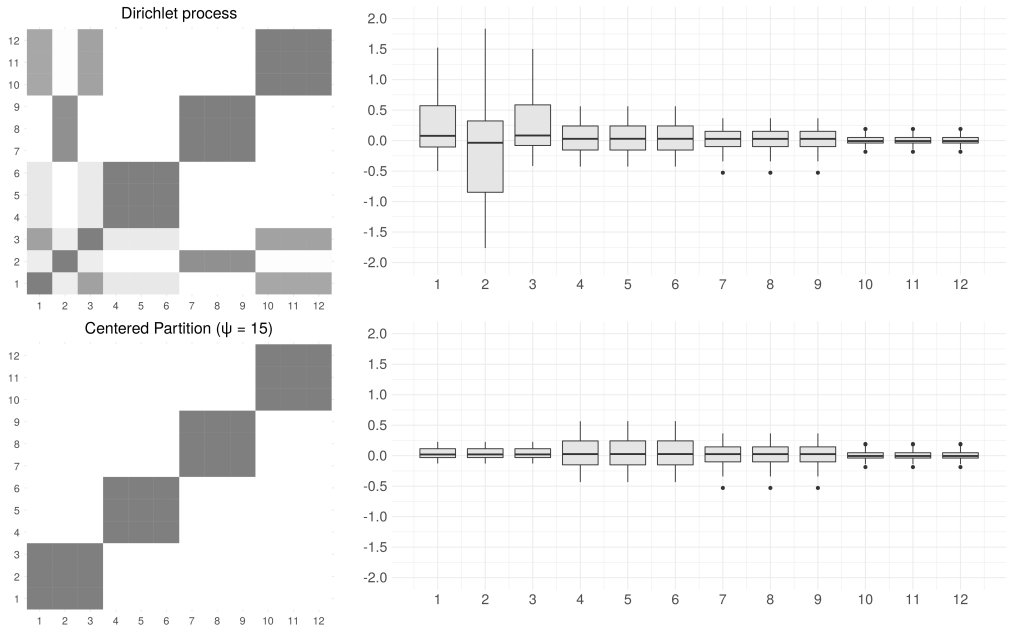

Figure 1: Results from grouped linear regressions with DP($\alpha = 1$) prior and CP process prior with DP($\alpha = 1$) base EPPF for $\psi = 15$, centered on the true partition. Heatmaps on the left side show the posterior similarity matrix. On the right side, boxplots show the distribution of deviations from the maximum likelihood baseline coefficients and posterior mean estimates for each dataset $i = 1, \ldots, 12$.

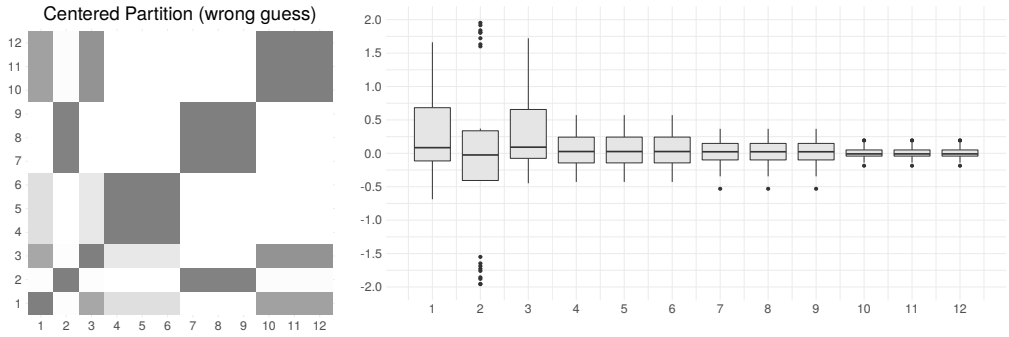

Figure 2: Results from grouped linear regression using CP process prior with DP($\alpha = 1$) base EPPF for $\psi = 15$ centered on partition $\boldsymbol{c}_0' = \{1, 5, 9\}\{2, 6, 10\}\{3, 7, 11\}\{4, 8, 12\}$ which has distance 3.16 from the true one. Heatmaps on the left side show the posterior similarity matrix. On the right side, boxplots show the distribution of deviations from the maximum likelihood baseline coefficients and posterior mean estimates for each dataset $i = 1, \ldots, 12$.

classification. In this setting the DP prior mostly struggles to identify the first group, which is the one characterized by the lowest number of observations, while it is correctly individuated under the CP process. In this case a value of $\psi = 15$ happened to be sufficient enough to recover the true clustering. We finally evaluate the CP prior performances when centered on wrong prior guess $\boldsymbol{c}_0' = \{1, 5, 9\}\{2, 6, 10\}\{3, 7, 11\}\{4, 8, 12\}$, having distance from $\boldsymbol{c}_0$ of approximately 3.16, obtaining posterior similarity matrix close to the of the DP prior.

2 Algorithms

Prior calibration

Algorithm 1 : Estimation of counts statistics related to distances neighborhoods of $\boldsymbol{c}_0$

Local search

0. Start from the base partition $\boldsymbol{c}_0$ with K_0 clusters and configuration $\boldsymbol{\lambda}_{m_0}$ and set $\delta_0 = 0$ and $\mathcal{N}_0(\boldsymbol{c}_0) = \boldsymbol{c}_0$.

for $t = 1, \dots, T$ **do**

1. Obtain $\mathcal{N}_t(\boldsymbol{c}_0)$ from partitions in $\mathcal{N}_{t-1}(\boldsymbol{c}_0)$ by exploring all directed connections, i.e. partitions obtained with one operation of split/merge on elements $\mathcal{N}_{t-1}(\boldsymbol{c}_0)$.

end for

2. Compute the distance from $\boldsymbol{c}_0$ for all partitions in $\mathcal{N}_T(\boldsymbol{c}_0)$ and take the minimum distance, δ_{L^*}; discard all partitions having distances greater than δ_{L^*}.

3. Obtain counts n_l and n_{lm} relative to distances $\delta_1, \dots, \delta_{L^*}$ for $m = 1, \dots, M$.

Monte Carlo approximation

for $r = 1, \dots, R$ **do**

4. Sample the number of clusters K from the discrete probability distribution

$$p(K = k) = e^{-1} k^N / (k! \mathcal{B}_N), \quad k \in \{1, 2, 3, \dots\}.$$

5. Conditional on K generate a partition $\boldsymbol{c}^{(r)} = \{c_1^{(r)}, \dots, c_N^{(r)}\}$ by sampling each $c_i^{(r)}$ from a discrete uniform distribution on $\{1, \dots, K\}$.

6. If $d(\boldsymbol{c}^{(r)}, \boldsymbol{c}_0) < \delta_{L^*}$ reject the partition.

end for

7. Let R^* be the number of accepted partitions, and estimate counts $\hat{n}_l$ and $\hat{n}_{lm}$ for $m = 1, \dots, M$ according to (4.6)-(4.7) in Section 4.2 conditional on the observed distance values $\hat{\delta}_{(L^*+1)}, \dots, \hat{\delta}_{L'}$.

Marginal sampling using variation of information

We describe how to compute the penalization term in the marginal sampling step described in Section 3.4 using the Variation of Information as a distance, but the same procedure applies when using other distances based on blocks sizes.

Algorithm 2 : Computation strategy for the penalization term in marginal sampling

Let K^- and K_0^- denote respectively the number of clusters in $\boldsymbol{c}^{-i}$ and $\boldsymbol{c}_0^{-i}$, i.e. partitions $\boldsymbol{c}$ and $\boldsymbol{c}_0$ after removing the i observation.
for $i = 1, \ldots, N$ **do**
1. Compute cardinalities $\{\lambda_1^{-i}, \ldots, \lambda_{K^-}^{-i}\}$ representing the number of observations in each cluster for $\boldsymbol{c}^{-i}$.
2. Compute λ_{lm}^{-i}, the number of observations in cluster l under $\boldsymbol{c}^{-i}$ and cluster m under $\boldsymbol{c}_0^{-i}$ for $l = 1, \ldots, K^-$ and $m = 1, \ldots, K_0^-$.
for $k = 1, \ldots, K^-, K^- + 1$ **do**
Let $c_{i,0}$ be the cluster of index i under partition $\boldsymbol{c}_0$.
Compute $d(\boldsymbol{c}, \boldsymbol{c}_0) \propto -H(\boldsymbol{c}) + 2H(\boldsymbol{c} \wedge \boldsymbol{c}_0)$ for $\boldsymbol{c} = \{\boldsymbol{c}^{-i} \cup k\}$ using

$$-H(\boldsymbol{c}) = \sum_{l \neq k}^{K} \left\{ \frac{\lambda_l^{-i}}{N} \log \frac{\lambda_l^{-i}}{N} \right\} + \left(\frac{\lambda_k^{-i} + 1}{N} \right) \log \left(\frac{\lambda_k^{-i} + 1}{N} \right)$$

$$H(\boldsymbol{c} \wedge \boldsymbol{c}_0) = -\left\{ \sum_{l=1}^{K} \sum_{m=1}^{K_0^-} \frac{\lambda_{lm}^{-i}}{N} \log \left(\frac{\lambda_{lm}^{-i}}{N} \right) - \frac{\lambda_{kc_{i,0}}^{-i}}{N} \log \left(\frac{\lambda_{kc_{i,0}}^{-i}}{N} \right) + \frac{\lambda_{kc_{i,0}}^{-i} + 1}{N} \log \left(\frac{\lambda_{kc_{i,0}}^{-i} + 1}{N} \right) \right\}$$

end for
end for

Gibbs sampling for shared logistic regression

In estimating the model, a Pólya-gamma data augmentation strategy is employed; for each y_{ij} we introduce a latent variable $\omega_{ij} \sim PG(1, \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i})$ for each observation j in defect-specific dataset i for $i = 1, \ldots, N$.

Algorithm 3 : Gibbs sampling for posterior computation

Conditionally on the cluster allocation vector $\mathbf{c} = (c_1, \ldots, c_n)$ and data $\{\mathbf{y}_i, \mathbf{X}_i\}$ for $i = 1, \ldots, N$, update mixture related parameters and Pólya-gamma latent variables as follows.

[1] Sample Pólya-gamma latent variables for each observation in each dataset
for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ **do**

$$(\omega_{ij}|-) \sim PG(1, \alpha_i + \mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i})$$

end for

[2] Update defect-specific intercept, exploiting Pólya-gamma conjugancy
for $i = 1, \ldots, N$ **do**

$$(\alpha_i|-) \sim \mathcal{N}(a^*, \tau^*)$$

with $\tau^* = \tau_0 + \sum_{j=1}^{n_i} \omega_{ij}$ and $a^* = [a_0\tau_0 + \sum_{j=1}^{n_i}(y_{ij} - 1/2 - \omega_{ij}\mathbf{x}_{ij}^T \boldsymbol{\beta}_{c_i})]/\tau^*$

end for

---

[3] Defining $\kappa_{ij} := y_{ij} - 1/2 - w_{ij}\alpha_i$, then the vector $(\kappa_{ij}/\omega_{ij}|c_i = k, \omega_{ij}) \sim \mathcal{N}(\mathbf{x}_{ij}^T\boldsymbol{\beta}_k, 1/\omega_{ij})$, and each cluster-specific coefficient vector $\boldsymbol{\beta}_k$ can be updated by aggregating all observations and augmented data relative to birth defects that are in the same cluster.

for $k = 1, \ldots, K$ **do**

Let $\mathbf{X}^{(k)}$, $\mathbf{y}^{(k)}$, $\boldsymbol{\kappa}^{(k)}$ be the obtained quantities relative to cluster k, and $\boldsymbol{\Omega}^{(k)}$ a diagonal matrix with the corresponding Pólya-gamma augmented variables. Then update cluster-specific coefficients vector from

$$(\boldsymbol{\beta}_k|-) \sim \mathcal{N}_p(\mathbf{b}^{(k)}, \mathbf{Q}^{(k)})$$

with $\mathbf{Q}^{(k)} = (\mathbf{X}^{(k)T}\boldsymbol{\Omega}^{(k)}\mathbf{X}^{(k)} + \mathbf{Q}^{-1})^{-1}$ and $\mathbf{b}^{(k)} = \mathbf{Q}^{(k)}(\mathbf{X}^{(k)T}\boldsymbol{\kappa}^{(k)} + \mathbf{Q}^{-1}\mathbf{b})$.

end for

---

[4] Allocate each birth defect i to one of the clusters

for $i = 1, \ldots, N$ **do**

Sample the class indicator c_i conditionally on $\boldsymbol{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$ from the discrete distribution with probabilities

$$\Pr(c_i = k|\boldsymbol{c}_{-i}, -) \propto \Pr(c_i = k|\boldsymbol{c}_{-i})\Pr(\mathbf{y}_i|\mathbf{X}_i, \alpha_i, c_i = k, \boldsymbol{\beta}_k)$$

with

$$\Pr(\mathbf{y}_i|\mathbf{X}_i, \alpha_i, c_j = k, \boldsymbol{\beta}_k) = \prod_{j=1}^{n_i} \left[\exp(\alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}_k)^{y_{ij}}\right] \left[1 + \exp(\alpha_i + \mathbf{x}_{ij}^T\boldsymbol{\beta}_k)\right]^{(-1)}$$

being the model likelihood evaluated for cluster k and $\Pr(c_i = k|\boldsymbol{c}^{-(i)})$ computed as described in Section 3.4.

end for

---

3 Results for NBDPS data application

Figure 3: **CP process with** $\psi = 0$. Posterior mean estimates of log odds-ratios, where dots indicate values significant at 95% using credibility interval. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor, while green a protective effect.

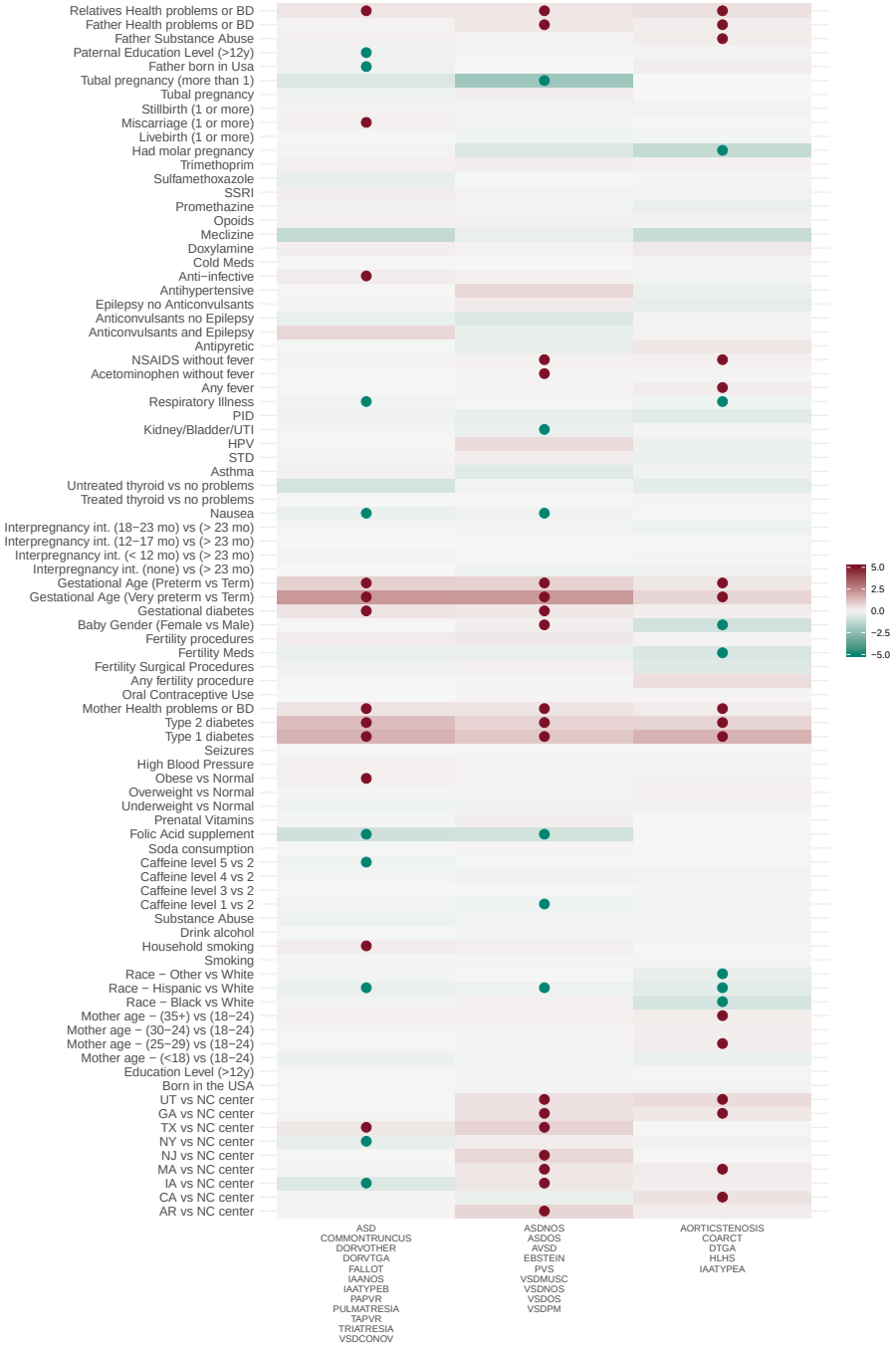

Figure 4: **CP process with** $\psi = 40$. Posterior mean estimates of log odds-ratios, where dots indicate values significant at 95% using credibility interval. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor, while green a protective effect.

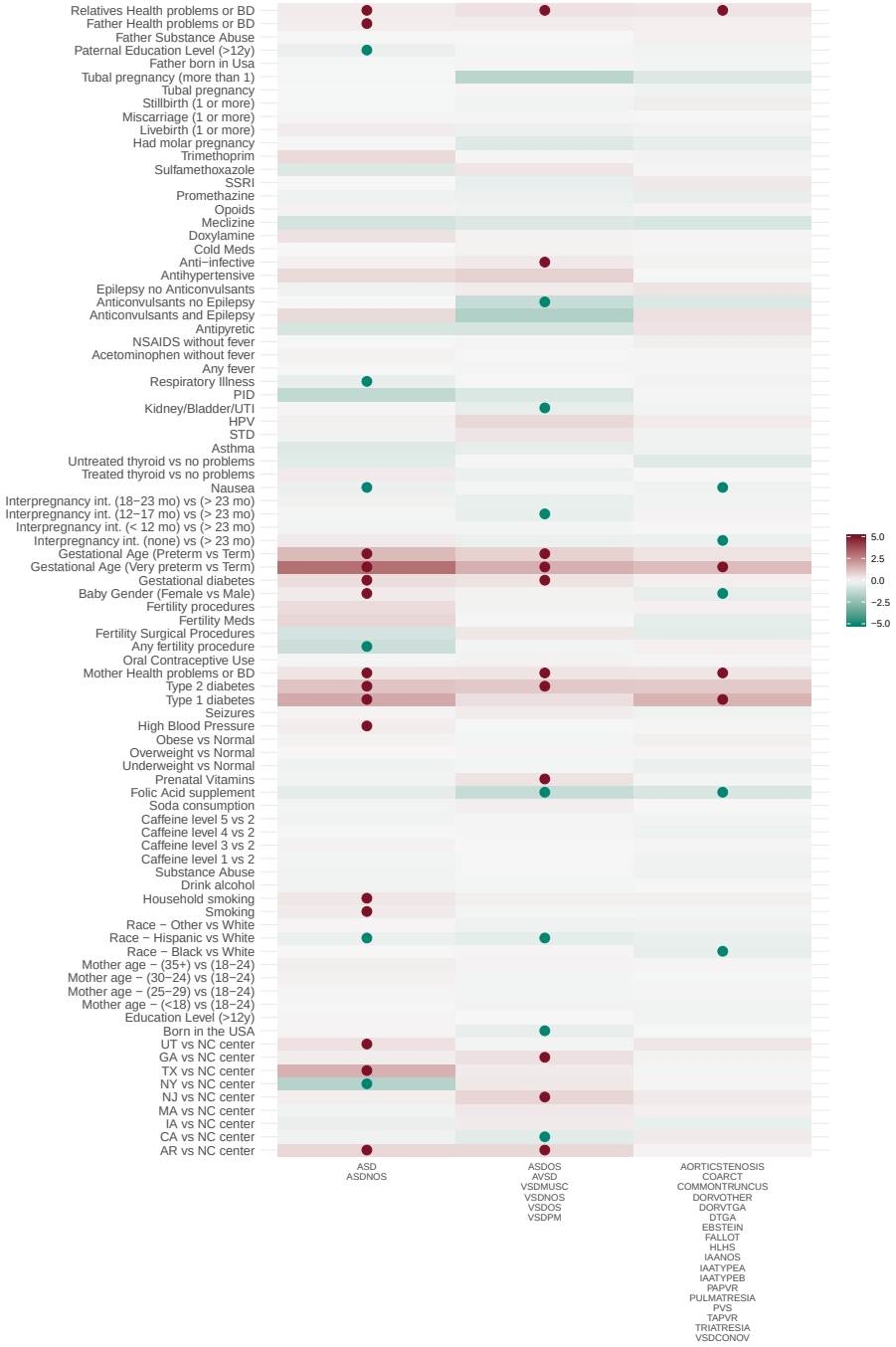

Figure 5: **CP process with** $\psi = 80$. Posterior mean estimates of log odds-ratios, where dots indicate values significant at 95% using credibility interval. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor, while green a protective effect.

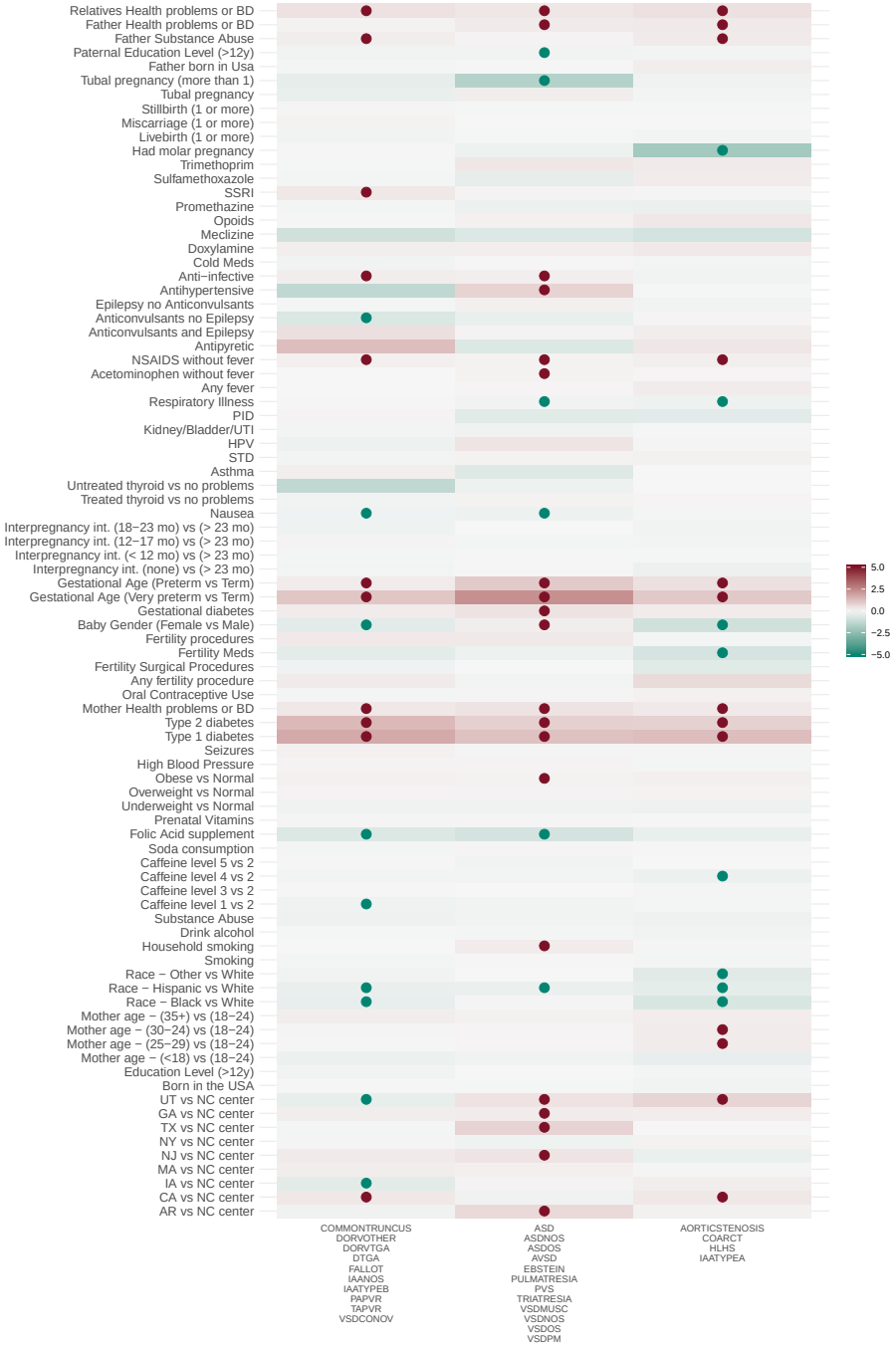

Figure 6: **CP process with** $\psi = 120$. Posterior mean estimates of log odds-ratios, where dots indicate values significant at 95% using credibility interval. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor, while green a protective effect.

Figure 7: **CP process with** $\psi = \infty$. Posterior mean estimates of log odds-ratios, where dots indicate values significant at 95% using credibility interval. Labels on the x-axis list the defects in each cluster. Red color indicates a risk factor, while green a protective effect.