The Annals of Applied Statistics

SUPPLEMENTARY FOR "PERTURBED FACTOR ANALYSIS: ACCOUNTING FOR GROUP DIFFERENCES IN EXPOSURE PROFILES"

BY ARKAPRAVA ROY*, ISAAC LAVINE†, AMY H. HERRING† AND DAVID B. DUNSON†

*University of Florida, †Duke University

1. S-1. We establish weak consistency in this section. Let $\Theta_\Lambda, \Theta_\Sigma, \Theta_E$ be the parameter spaces of Λ, Σ and E, respectively, Θ be the set of $p \times p$ positive semidefinite matrices corresponding to the parameter space of H, and $\mathcal{Q}_j$ be the parameter space of Q_j. Let $\Pi_\Lambda, \Pi_\Sigma, \Pi_E, \Pi_Q$ be the priors for Λ, Σ, E and Q_j's. We restate some of the results from Bhattacharya and Dunson (2011) for our modified factor model. With minor modification, the proofs will remain the same.

Let $g : \Theta_\Lambda \times \Theta_\Sigma \times \Theta_E \to \Theta_H$ be a continuous map such that $g(\Lambda, \Sigma, E) = \Lambda^T E \Lambda + \Sigma$.

LEMMA 1. *For any* $(\Lambda, \Sigma, E) \in \Theta_\Lambda \times \Theta_\Sigma \times \Theta_E$, *we have* $g(\Lambda, \Sigma, E) \in \Theta_H$.

The proof is similar to Lemma 1 of Bhattacharya and Dunson (2011). In our Bayesian approach, we choose independent priors for Λ, Σ and E and that induces a prior on H through the map g. We also have following proposition.

PROPOSITION 2. *If* $(\Lambda, \Sigma, E) \sim \Pi_\Lambda \otimes \Pi_\Sigma \otimes \Pi_E$, *then* $\Pi_\Lambda \otimes \Pi_\Sigma \otimes \Pi_E(\Theta_\Lambda \times \Theta_\Sigma \times \Theta_E) = 1$.

The proof is similar to Proposition 1 of Bhattacharya and Dunson (2011) with minor modifications. Now, we proceed to establish that the posterior of our multigroup model is weakly consistent under a fixed p and increasing n regime. Let us assume that the complete parameter space is $\kappa = (\Lambda, \Sigma, E, Q_2, \ldots, Q_j)$ and let κ_0 be the truth for κ.

Assumptions:

1. For some $M > 0$, the true perturbation matrices are $Q_{j0} \in \mathcal{C}_M$, with $\mathcal{C}_M$ defined in Section 2.1.1.
2. There exists some $E > 0$, $F_1 > 0$ and $F_2 > 0$ such that $\max_{ij} |\Lambda_0| < E$ and $F_1 < e_{0k} < F_2$ for all $k = 1, \ldots r$.

THEOREM 3. *Under Assumptions 1-2, the posterior for κ is weakly consistent at κ_0.*

We first show that our proposed prior has large support in the sense that the truth belongs to the Kullback-Leibler support of the prior. Thus the posterior probability of any neighbourhood around the truth converges to one in $P_{\kappa_0}^{(n)}$-probability as n goes to ∞ as a consequence of Schwartz (1965). Here $P_{\kappa}^{(n)}$ is the distribution of a sample of n observations with parameter κ. Hence, the posterior is weakly consistent.

PROOF. For $q, q^* \in$ the space of probability measure $\mathcal{P}$, let the Kullback-Leibler divergences be given by

$$KL(q^*, q) = \int q^* \log \frac{q^*}{q}.$$

Let $K(\kappa_0, \kappa)$ denote the Kullback-Leibler divergence

$$\sum_{j=1}^{J} KL(\mathrm{N}(0, Q_j^{-1}[\Lambda E \Lambda^T + \Sigma](Q_j^T)^{-1}), \mathrm{N}(0, Q_{j0}^{-1}[\Lambda_0 E_0 \Lambda_0^T + \Sigma_0](Q_{j0}^{-1})^T)).$$

For our model, we have

$$K(\kappa_0, \kappa) = \frac{1}{2n}\Bigg[\sum_{j=1}^{J} -n_j \log |Q_{j0}^{-1} H_0 (Q_{j0}^T)^{-1} \{Q_j^{-1} H (Q_j^T)^{-1}\}^{-1}| + n_j \mathrm{tr}\big(Q_{j0}^{-1} H_0 (Q_{j0}^T)^{-1} \{Q_j^{-1} H (Q_j^T)^{-1}\}^{-1} - I_p\big)\Bigg]$$

To prove Theorem 3, we rely on the following Lemma.

LEMMA 4. *For any $\epsilon > 0$, there exists $\epsilon_1 > 0, \epsilon_2 > 0$, and $\epsilon_{3j} > 0$ for $j \in \{2, \ldots, J\}$ such that $\|\Lambda - \Lambda_0\|_F^2 \leq \epsilon_1^2$, $\|E - E_0\|_F^2 \leq \epsilon_2^2$ and $\|Q_j - Q_{j0}\|_F^2 \leq \epsilon_{3j}^2$, then we have*

$$\Pi\big\{K(\kappa_0, \kappa) \leq \epsilon\big\} \geq \Pi\big\{\|\Lambda - \Lambda_0\|_F^2 \leq \epsilon_1^2, \|E - E_0\|_F^2 \leq \epsilon_2^2, \|Q_j - Q_{j0}\|_F^2 \leq \epsilon_{3j}^2, j = 2, \ldots, J\big\}$$

Due to continuity of the functions such as determinant, trace and $g(\cdot)$, the above result is immediate following the proof of Theorem 2 in Bhattacharya and Dunson (2011). For our proposed priors, the prior probability of the R.H.S. of Lemma 4 is positive. Thus the prior probability of any Kullback-Leibler neighborhood around the truth is positive. This proves Theorem 3. □

2. S-2. This section deals with some additional simulation experiments and results from the NHANES dataset.

2.1. *Case 1: Multi-group, additive perturbation.* In Section 5.2 of the manuscript, each group is multiplied by a unique perturbation matrix Q_j. In this case, we perturb the data by adding a group-specific loading matrix Ψ_j to a shared loading matrix Λ, as in model (2.1).

$$\begin{aligned} Y_i =&(\Lambda + \Psi_j)\eta_i + \epsilon_{1i} \text{ and } Y_i \in G_j, \\ \epsilon_{1i} \sim&\mathrm{N}(0, \Sigma) \quad \eta_i \sim \mathrm{N}(0, I_p), \end{aligned} \tag{2.1}$$

where the group-specific loadings (Ψ_j's) are lower in magnitude in comparison to the shared loading matrix Λ. This can also be an alternative perturbation model. The group-specific loadings are generated in Figure 1 which shows the true and estimated loading matrices from BMFSA, FBPFA, and PFA across a range of values of α. The estimated loadings from PFA are much closer to the truth than BMFSA. We find that the estimated loading matrices are some permutation of the true loading matrix with some exceptions. For the case with higher perturbation (row 2 of Figure 1), BMSFA estimate is not good and PFA estimate is also bad for lower values of α. Similarly, for the case with lower perturbations, PFA estimate with lower α is better than higher α estimated loading. Table 1 compares predictive likelihoods of the two methods in different cases, and again, PFA and FBPFA outperform BMSFA.

TABLE 1
Average predictive log-likelihood for PFA for different choices of α, FBPFA and BMSFA in Simulation Case 1.

Generative Distribution of Ψ_j's	True Loading	PFA for $\alpha = 10^{-2}$	PFA for $\alpha = 10^{-4}$	FBPFA	BMSFA
$\mathrm{N}(-0.2, 0.2)$	Loading 1	-56.27	-60.71	-56.63	-429.28
	Loading 2	-1041.28	-361.82	-421.28	-5853.65
$\mathrm{N}(-0.5, 0.8)$	Loading 1	-30.68	-50.09	-54.62	-514.67
	Loading 2	-320.04	-246.06	-479.59	-6798.37

2.2. *Case 2: Observation-level perturbations.* In this case, we generate data from the model (2.2). First, a non-perturbed dataset is generated exactly the same way as in Section 5.1 and 5.2 in the manuscript. Then, we generate separate perturbation matrices $Q_{i0} \sim \mathrm{MN}(I_p, \alpha_0 I_p, \alpha_0 I_p)$ for each data vector Y_i. We repeat this simulation for different choices of α_0.

Figures 2 shows the estimated loading when they are estimated using $\alpha = 10^{-4}$. The method performs poorly when the perturbation parameter

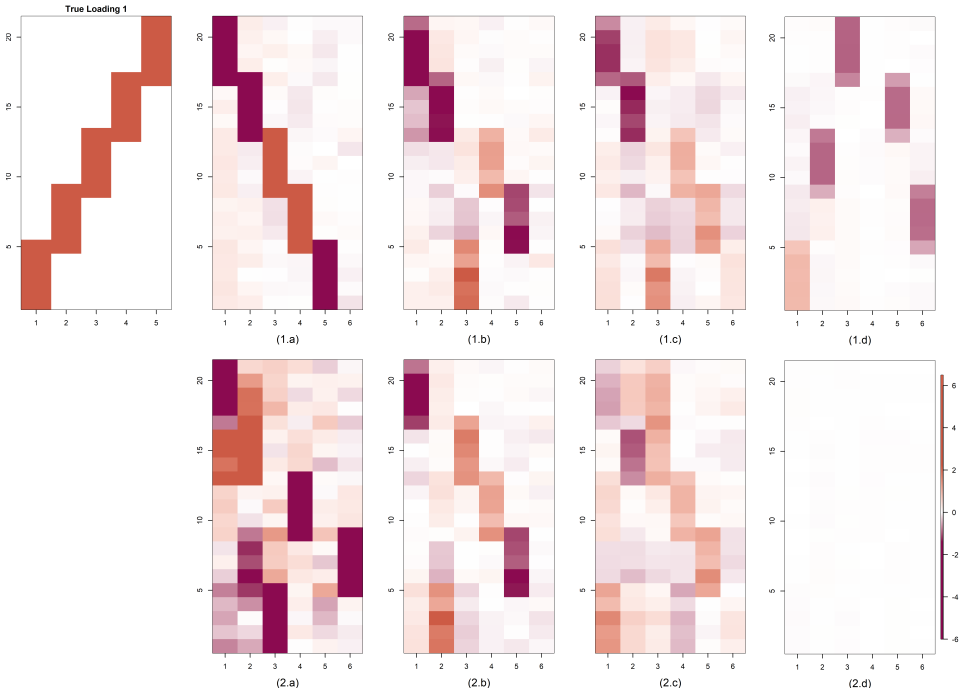


FIG 1. *Comparison of estimated loading matrices in Simulation case 2.1 when the true data generating process follows the model in (2.1). For the first two rows, true* Ψ_j*'s are generated from* $N(-0.2, 0.2)$ *and for the last two rows* Ψ_j*'s are generated from* $N(-0.5, 0.8)$ *and for each row (a) PFA with* $\alpha = 10^{-4}$*, (b) PFA with* $\alpha = 10^{-2}$*, (c) FBPFA, (d) BMSFA*

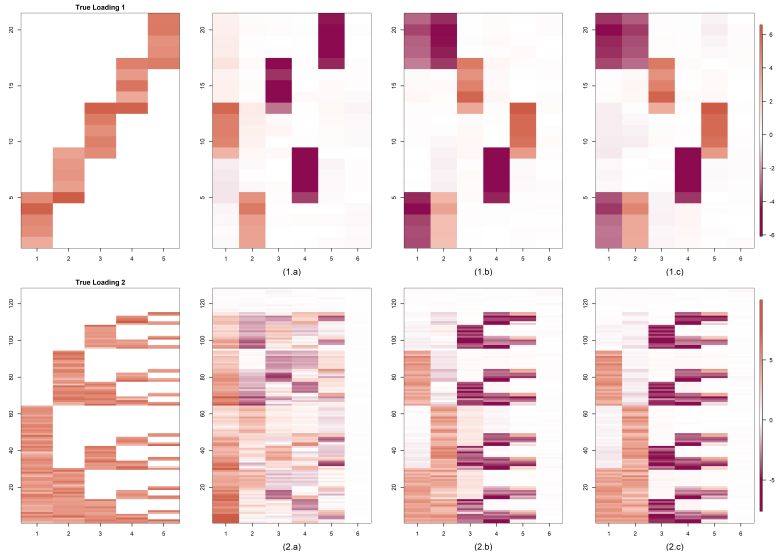


FIG 2. *Estimated loadings using PFA for different choices of α in Simulation case 2.2 for $\alpha_0 = 10^{-4}$ and two loadings. (a): Estimated loading for $\alpha = 10^{-2}$, (b): For $\alpha = 10^{-4}$, (c): For $\alpha = 10^{-5}$.*

α is lower than the true parameter α_0 as in the previous cases. Figure 3 illustrates the performance of FBPFA in this case. When the perturbation is higher, the performance deteriorates as it is more difficult to capture the loading structure accurately. Overall, our method is able to accurately estimate the true loading structure under some permutation of the columns.

2.3. *More exploratory analysis on NHANES data.* For each chemical level, we fit an one-way ANOVA model to analyse group-specific effects on each phthalate level separately. The results are provided from Table 2 to 9.

TABLE 2
Estimated group effects from a one-way ANOVA analysis for the phthalate MnBP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.85	0.11	36.27	0.00
N-H Black	0.36	0.15	2.33	0.02
N-H White	-0.29	0.13	-2.26	0.02
OH	0.46	0.18	2.55	0.01
Other/Multi	-0.09	0.22	-0.41	0.68

3. S-3. This section has some more additional figures. We have some additional figures in this section as discussed in the Manuscript. Figure 4

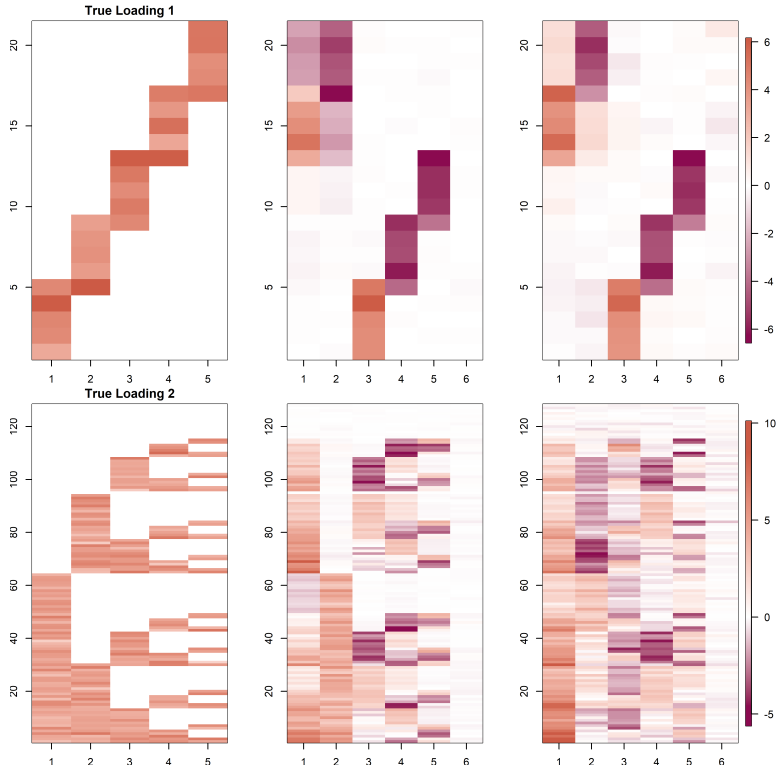


FIG 3. *Estimated loadings using FBPFA for different choices of* α_0 *in Simulation case 2.2 for two cases and two loadings. (a): Estimated loading for* $\alpha_0 = 10^{-4}$*, (b): For* $\alpha_0 = 10^{-2}$.

TABLE 3
Estimated group effects from a one-way ANOVA analysis for the phthalate MiBP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.88	0.06	49.23	0.00
N-H Black	0.59	0.08	6.97	0.00
N-H White	-0.35	0.07	-4.98	0.00
OH	0.33	0.10	3.28	0.00
Other/Multi	-0.08	0.12	-0.68	0.50

TABLE 4
Estimated group effects from a one-way ANOVA analysis for the phthalate MEP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	8.32	0.29	28.80	0.00
N-H Black	2.84	0.42	6.79	0.00
N-H White	-1.41	0.35	-4.02	0.00
OH	1.18	0.49	2.40	0.02
Other/Multi	-0.99	0.60	-1.64	0.10

shows the MCMC mixing of the perturbation matrices and loading matrices using FBPFA for NHANES application. Figure 5 illustrates estimated loadings for the second choice of loading matrix in Section 5.2 of the manuscript. From the same section, estimated loadings for the partially shared factors case are in Figure 6. Figure 7 depicts estimated loading for the same choice of loading in Section 5.3 of the manuscript.

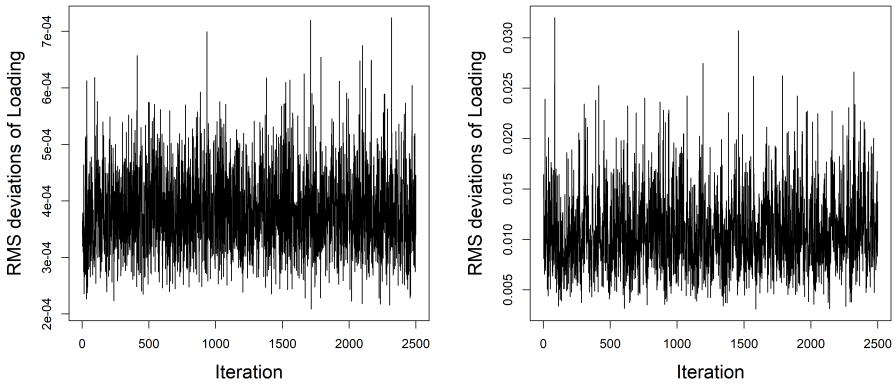
References.

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98(2)** 291–306.

SCHWARTZ, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4** 10–26.

ARKAPRAVA ROY
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FL
E-MAIL: ark007@phhp.ufl.edu

ISAAC LAVINE
AMY HERRING
DAVID B. DUNSON
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NC
E-MAIL: isaac.lavine@duke.edu
amy.herring@duke.edu
dunson@duke.edu



(a) Trace plot for Q (b) Trace plot for Λ

FIG 4. *Trace plots of root mean square (RMS) deviations across the MCMC chain for the perturbation matrices* $\frac{1}{4}\sum_{j=2}^{5}\overline{(Q_j^t - Q_j^{t+1})^2}$ *and the shared loading matrix* Λ *which is* $\overline{(\Lambda^t - \Lambda^{t+1})^2}$. *The matrices* Q_j^t *and* Λ^t *are the t-th post burn samples of* Q_j *and* Λ *respectively for NHANES application.*

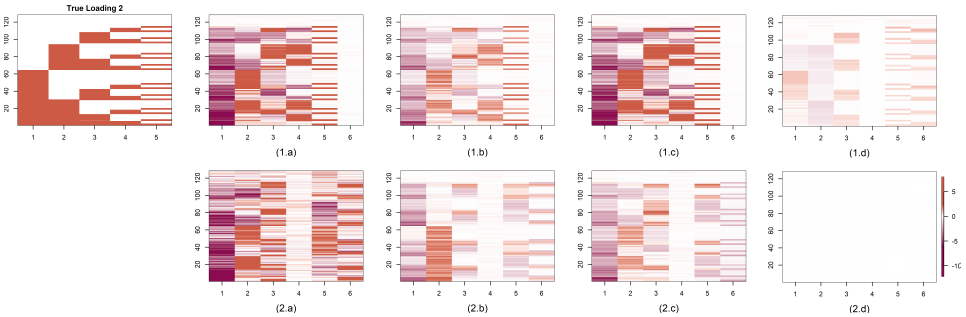


FIG 5. *Comparison of estimated loading for loading matrix 2 in simulation case 2 of the manuscript with different choices of* α_0 *and* α *where* $Q_{j0} \sim MN(I_p, \alpha_0 I_p, \alpha_0 I_p)$ *and* $U = \alpha I_p = V$. *(a)* $\alpha = 1\times10^{-4}, \alpha_0 = 1\times10^{-4}$, *(b)* $\alpha = 1\times10^{-2}, \alpha_0 = 1\times10^{-4}$, *(c) FBPFA with* $\alpha_0 = 1\times10^{-4}$, *(d) BMSFA with* $\alpha_0 = 1\times10^{-4}$, *(e)* $\alpha = 1\times10^{-4}, \alpha_0 = 1\times10^{-2}$, *(f)* $\alpha = 1\times10^{-2}, \alpha_0 = 1\times10^{-2}$, *(g) FBPFA with* $\alpha_0 = 1\times10^{-2}$, *(h) BMSFA with* $\alpha_0 = 1\times10^{-2}$. *True loading matrices are plotted twice in columns 1 for easier comparison with other images. The color gradient added in the last image holds for all the images.*

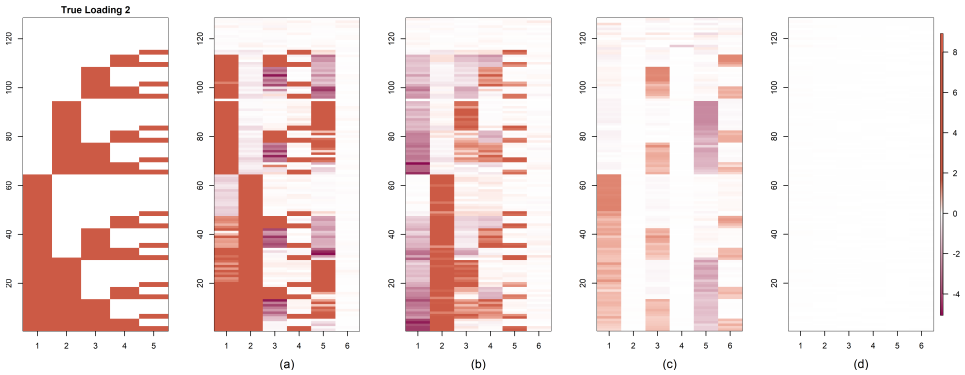


FIG 6. *Comparison of estimated loading in the partially shared modification of simulation case 2 of the manuscript. Row 1 corresponds to true loading structure 1 and row 2 to true loading structure 2. (a) FBPFA with* $\alpha_0 = 1 \times 10^{-4}$*, (b) FBPFA with* $\alpha_0 = 1 \times 10^{-2}$*, (c) BMSFA with* $\alpha_0 = 1 \times 10^{-4}$*, (d) BMSFA with* $\alpha_0 = 1 \times 10^{-2}$*, (e) FBPFA with* $\alpha_0 = 1 \times 10^{-4}$*, (f) FBPFA with* $\alpha_0 = 1 \times 10^{-2}$*, (g) BMSFA with* $\alpha_0 = 1 \times 10^{-4}$*, (h) BMSFA with* $\alpha_0 = 1 \times 10^{-2}$*. The color gradient added in the last image holds for all the images.*

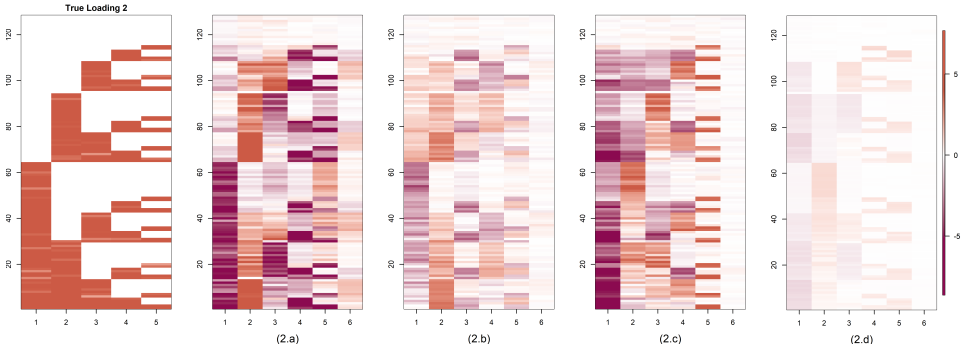


FIG 7. *Comparison of estimated loading matrices in simulation case 3 of the manuscript when the true data generating process follows the model in* (**??**)*. For the first two rows, true* Ψ_j*'s are generated from* $N(-0.2, 0.2)$ *and for the last two rows* Ψ_j*'s are generated from* $N(-0.5, 0.8)$ *and for each row (a) PFA with* $\alpha = 1 \times 10^{-2}$*, (b)PFA with* $\alpha = 1 \times 10^{-4}$*, (c) FBPFA, (d) BMSFA. The color gradient added in the last image holds for all the images.*

TABLE 5
Estimated group effects from a one-way ANOVA analysis for the phthalate MBeP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.79	0.07	39.90	0.00
N-H Black	0.27	0.10	2.66	0.01
N-H White	-0.09	0.08	-1.05	0.29
OH	-0.01	0.12	-0.06	0.95
Other/Multi	-0.23	0.15	-1.57	0.12

TABLE 6
Estimated group effects from a one-way ANOVA analysis for the phthalate MECPP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	4.80	0.12	41.52	0.00
N-H Black	-0.44	0.17	-2.63	0.01
N-H White	-0.64	0.14	-4.56	0.00
OH	-0.25	0.20	-1.28	0.20
Other/Multi	-0.46	0.24	-1.89	0.06

TABLE 7
Estimated group effects from a one-way ANOVA analysis for the phthalate MEHHP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.83	0.10	36.85	0.00
N-H Black	-0.04	0.15	-0.26	0.79
N-H White	-0.39	0.13	-3.06	0.00
OH	-0.10	0.18	-0.56	0.58
Other/Multi	-0.26	0.22	-1.19	0.23

TABLE 8
Estimated group effects from a one-way ANOVA analysis for the phthalate MEOHP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.11	0.08	39.04	0.00
N-H Black	-0.06	0.12	-0.54	0.59
N-H White	-0.33	0.10	-3.39	0.00
OH	-0.11	0.14	-0.83	0.41
Other/Multi	-0.26	0.17	-1.54	0.12

TABLE 9
Estimated group effects from a one-way ANOVA analysis for the phthalate MEHP.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	1.58	0.04	35.44	0.00
N-H Black	0.03	0.06	0.44	0.66
N-H White	-0.25	0.05	-4.57	0.00
OH	0.01	0.08	0.10	0.92
Other/Multi	0.00	0.09	0.02	0.99