

Supplementary Materials for Bayesian semiparametric univariate and multivariate density deconvolution with the R Package BayesDecon

BLAKE MOYA¹, MAINAK MANNA¹, AND ABHRA SARKAR¹

¹DEPARTMENT OF STATISTICS AND DATA SCIENCES, THE UNIVERSITY OF TEXAS AT AUSTIN, TEXAS
78705, USA

Contents

S.1 Additional Simulation Experiments	S.2
S.2 Additional NHANES Data Analyses	S.6

S.1 Additional Simulation Experiments

Here we present results of some additional numerical experiments to validate the efficiency and reliability of the proposed package. Specifically, we replicated the simulation study reported in section S.6 of [Sarkar et al. \(2021\)](#); full details can be found therein, we present a brief overview of the simulation set-up here.

For three variables $V1$, $V2$ and $V3$, samples of 500 and 1000 units were generated from distributions of varying shapes (reflected J-shaped, bimodal, and unimodal), each unit having three replicates. Each observation of $V1$, $V2$ and $V3$ were contaminated with conditionally heteroscedastic measurement error with varying shapes and a linear variance function. After drawing the error-contaminated observations, either all, or two, or none of the variables were systematically designated to be episodic using a probability model for zero-inflation. For the variables $V1$, $V2$ and $V3$, when chosen as episodic, respectively approximately 20%, 35% and 17% of their values equaled exact zero.

For each of the six possible combinations determined by the sample size (500 or 1000) and number of episodic components (three, two or, zero), 100 data sets were simulated. For each data set, we estimated the joint and marginal densities of $V1$, $V2$ and $V3$ using the functions available in the **BayesDecon** package. The median of the integrated square errors (ISEs) in estimating the corresponding true densities are then compared with the median ISEs obtained using original R codes, and are reported in Table S.2.

The ISEs obtained by the **BayesDecon** package are overall comparable to that obtained by the original R codes. This is anticipated, given that both implementations use similar models. Nonetheless some performance gains were observed in cases where the original codes underperformed, e.g., for the case with all 3 episodic components, the median $1000 \times$ ISE for component 3 reduced from 13.03 to 2.10 and from 6.28 to 0.95 for 500 and 1000 units, respectively. These improvements are likely attributable to enhancements in modeling aspects outlined in the main paper and more effective handling of numerical instabilities in the new **C++** codes.

No of ECs	No of RCs	Sample size	Median ISE $\times$ 1000	
			Original R codes	BayesDecon
3	0	500	13.04 (1.78, 7.05, 13.03)	9.14 (2.89, 6.67, 2.10)
3	0	1000	8.12 (1.63, 4.05, 6.28)	4.93 (1.85, 2.72, 0.95)
2	1	500	5.83 (1.83, 7.06, 0.46)	5.88 (2.81, 6.49, 0.76)
2	1	1000	2.79 (1.01, 3.06, 0.43)	4.33 (2.43, 2.65, 0.40)
0	3	500	4.79 (1.52, 3.31, 0.47)	5.31 (2.44, 3.02, 0.78)
0	3	1000	2.75 (0.73, 1.76, 0.40)	3.01 (1.35, 1.47, 0.40)

Table S.2: Median ISEs in estimating the joint and marginal densities (in parenthesis) obtained using the original R codes and the new **BayesDecon** package. Here EC and RC are abbreviations for episodic component and regular component, respectively.

Next, we present results from numerical experiments designed to assess the consistency of marginal density estimates produced by the package when applied to different subsets of data containing the same variable(s). For this evaluation, we focused on the combination of two episodic variables ($V1$ and $V2$) and one regular variable ($V3$), using a sample of 1000 units each with 3 replicates. We selected the dataset corresponding to the median ISE for the joint density estimate and proceeded as follows.

We conducted univariate analyses for each variable, multivariate analyses for all possible pairs of the variables, and a multivariate analysis on all three variables. We investigated whether the marginal estimates of the variables V_1 , V_2 and V_3 remain consistent across univariate and multivariate settings. Figure S.1 supports this, showing that the estimated marginal densities are very similar across all analyses.

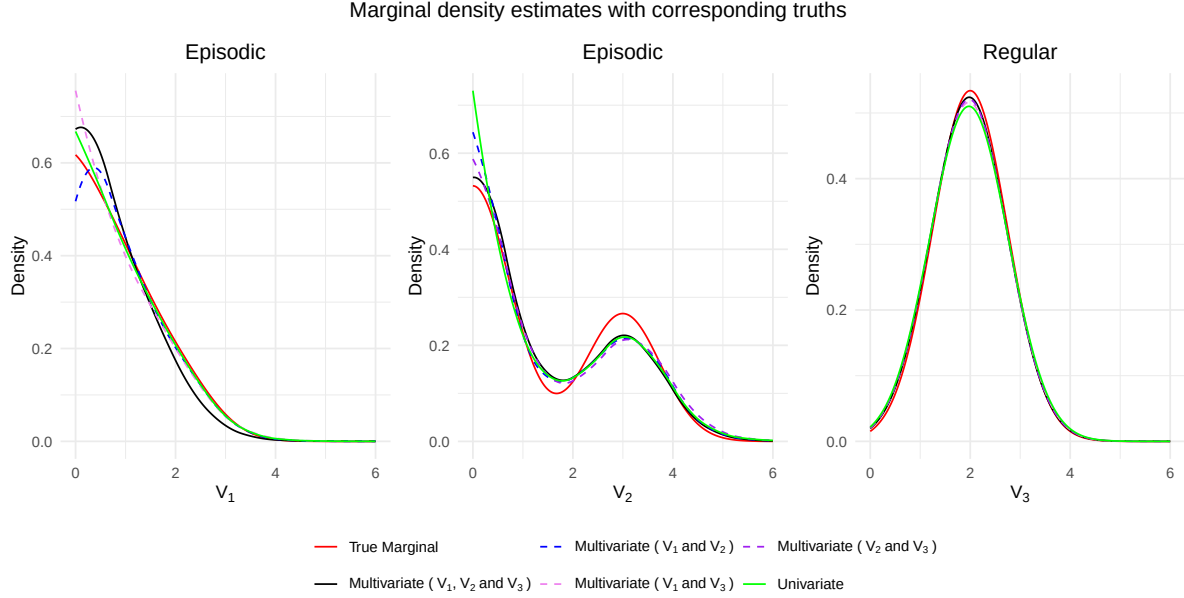
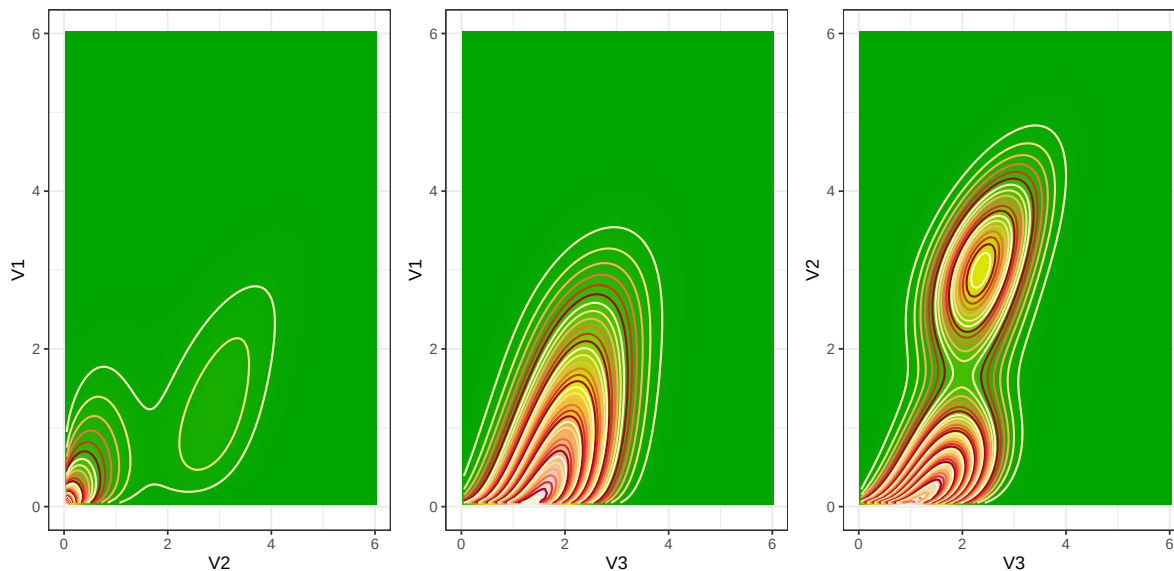
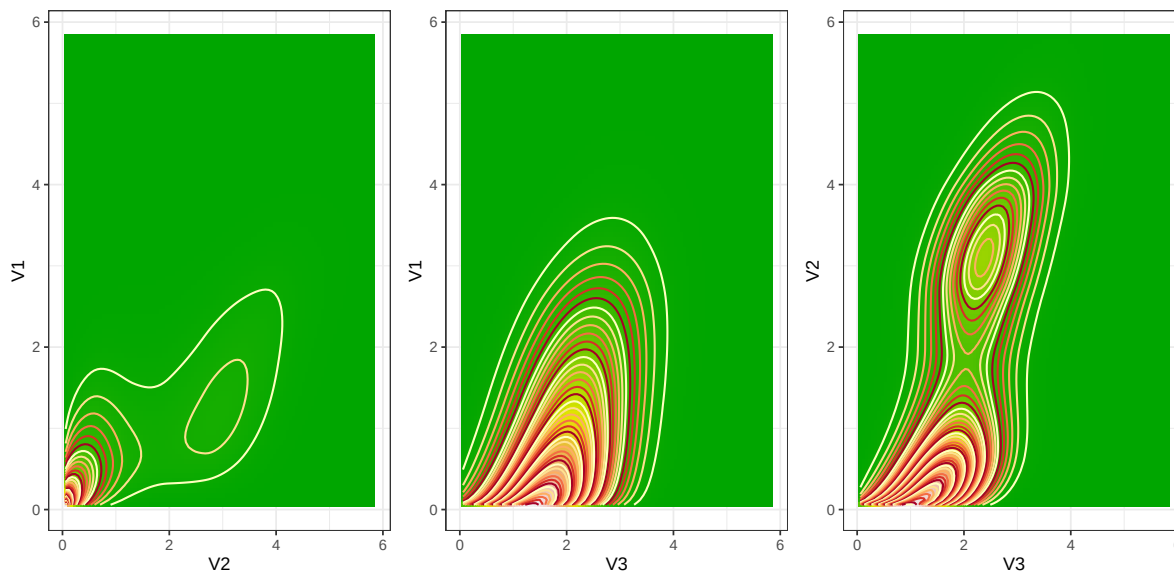
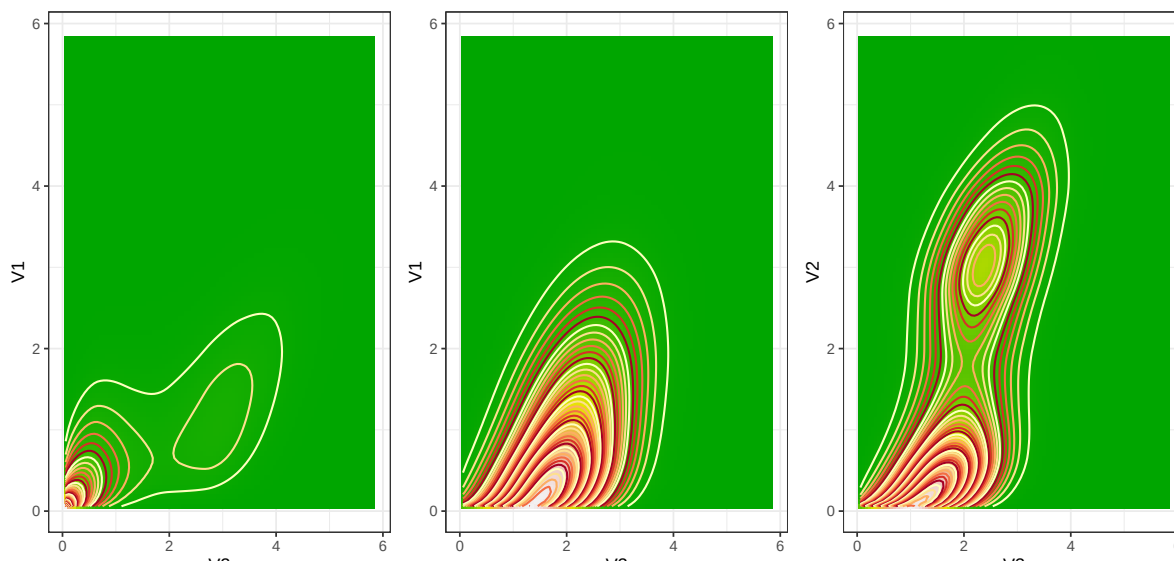


Figure S.1: Results for simulated data: Marginal density estimates from univariate and multivariate analyses for two episodic variables (V_1 and V_2) and one regular variable (V_3) – true marginal densities (red line), estimated densities from univariate analyses (V_1 , V_2 , V_3 ; green lines), estimated densities from multivariate analyses of variable pairs (dashed lines in various colors), and from the joint multivariate analysis of all three variables together (black line).

Similar experiments were also done for the joint densities. Figure S.2 shows the true joint densities for each pair of those variables and their corresponding estimates obtained from the multivariate analyses described above.

We also conducted experiments to verify consistency of the results for the same data across different random number generators. We picked the data for which the above experiments were done, and ran the analysis for seed numbers 100 and 500. The high consistency of the results across these seeds can be verified from Figure S.3.

True joint densities for different pairs of V_1 , V_2 and V_3 Estimated joint densities for (V_1, V_2) , (V_1, V_3) and (V_2, V_3) obtained from bivariate analysisEstimated joint densities for different pairs of V_1 , V_2 and V_3 obtained from multivariate analysis

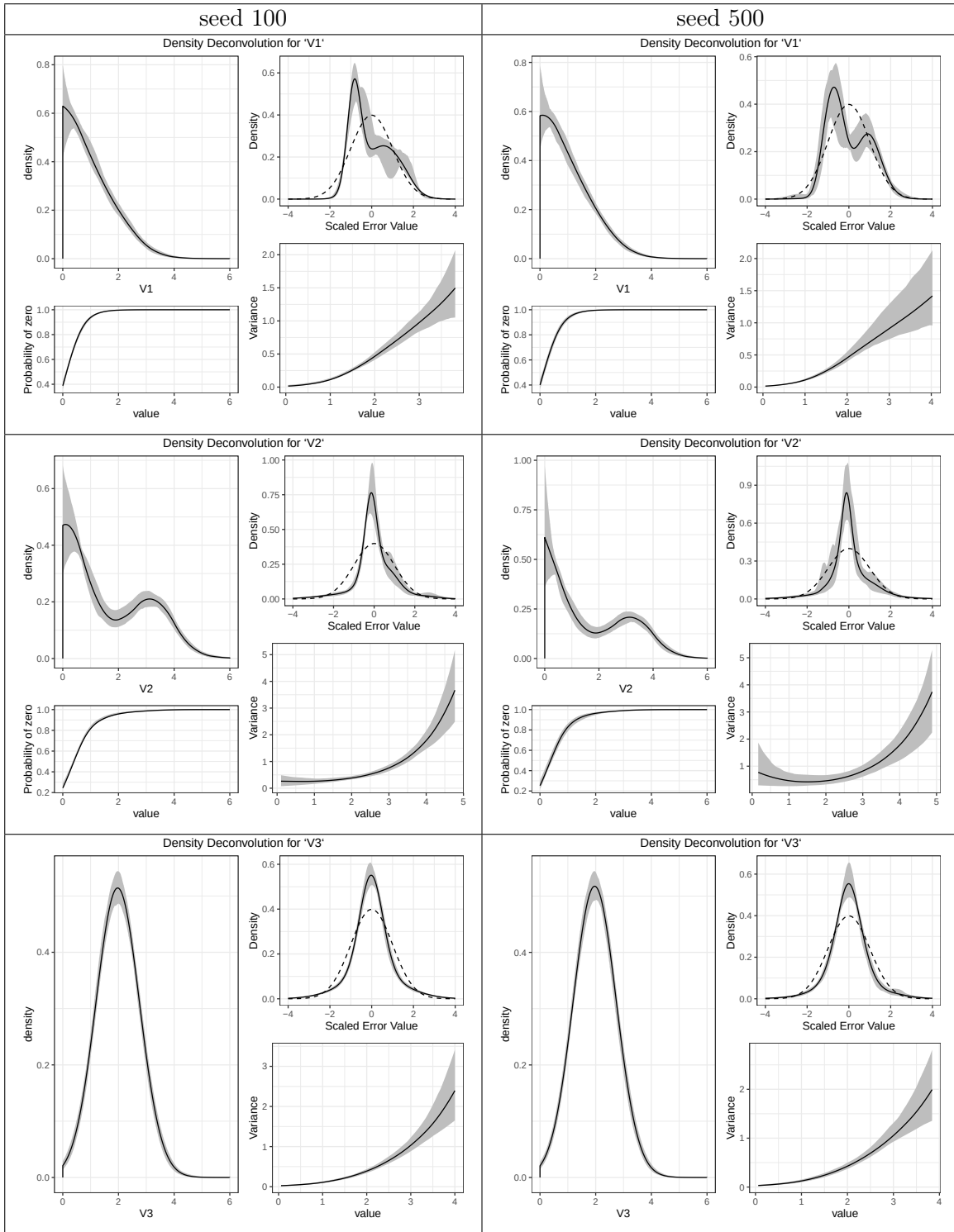


Figure S.3: Results for simulated data: Comparison of estimates obtained from the `bdeconv` function applied to simulated data across two different seeds.

S.2 Additional NHANES Data Analyses

We conducted similar reproducibility experiments on the NHANES data set, where 4 nested subsets of variables were selected from the original set of 13. The groups were

- **Group 1:** TSFAT
- **Group 2:** PFALLPROTLEG, TSFAT
- **Group 3:** PFALLPROTLEG, TFACIDS, TSFAT
- **Group 4:** PFALLPROTLEG, TFACIDS, TSODI, TSFAT

Figure S.4 displays the estimated marginal densities for the TSFAT variable across all experimental settings, along with the corresponding estimated marginal error densities and marginal variance functions. Figure S.5 displays the estimated bivariate joint densities for the variables taken from Group 2, Group 3, and Group 4, respectively. Here, we took the pairs (TSFAT, PFALLPROTLEG) from Group 2, 3 and 4, and (PFALLPROTLEG, TFACIDS) from Group 3 and 4, respectively. These figures suggest that the new **BayesDecon** package is able to produce consistent density estimates for a specific group of variables obtained from marginalizations over different supersets.

Similar to the simulation setup, we conducted experiments for the NHANES data to verify the consistency of our results across different random number generators. We picked three variables, namely, T_D_TOTAL, TSODI and TSFAT, of which the first one is episodic with approximately 10% of the recalls equaling exact zeros. We show the analysis for them across the seeds 100 and 500. As discussed in Section 4.5 of main paper, the estimated densities of main interest were highly consistent across all seeds. However, unlike in simulation settings, we observed some variability in the estimated shapes of the scaled error distribution for some of the cases with the default choice of the hyper-parameters, especially when the variables have significant mass near zero and additionally have many exact zero proxies in episodic cases. See, e.g., the episodic case in Figure S.6. This behavior stems from the lack of separate identifiability between the scaled errors ϵ and the variance functions $s^2(X)$. When their product $U = s(X)\epsilon$ is small, these components are numerically difficult to distinguish, particularly given the limited information in datasets with only two replicates per subject. Consequently, some mixture components in the error model exhibit highly dispersed variances under our default hyper-parameter choices $(a_epsilon, b_epsilon) = (3, 2)$ for the inverse-Gamma priors on their variances, which represent the smallest integer values that yield a prior mean of one and a finite prior variance. The estimated f_X 's of main interest still remain consistent as their updates depend on the product $U = s(X)\epsilon$, and when these U 's are small, at least in regular cases, the corresponding proxies are actually highly informative about the underlying X 's, making the estimates of f_X 's robust against uncertainty in the estimates of scaled error distributions or the variance functions, even when the number of replicates is limited. In such cases, we can use stronger prior regularization for the scaled error mixture components' variances, e.g., by setting the hyper-parameters $(a_epsilon, b_epsilon) = (6, 5)$, which leads to more consistent and realistic results for the scaled error distributions as well. These results are illustrated in Figure S.7.

References

- Sarkar A, Pati D, Mallick BK, Carroll RJ (2021). Bayesian copula density deconvolution for zero-inflated data in nutritional epidemiology. *Journal of the American Statistical Association*, 116: 1075–1087.

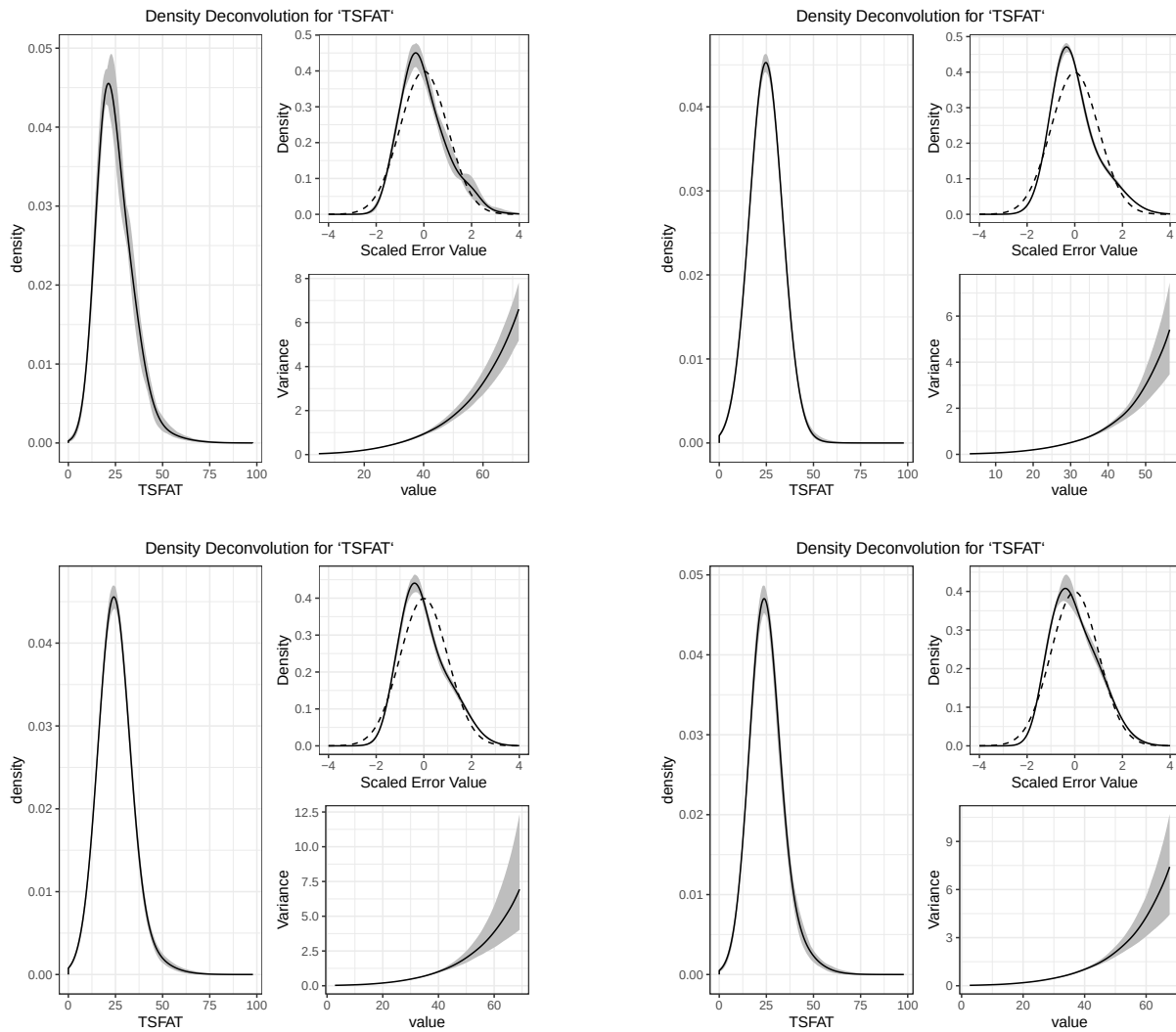


Figure S.4: Results for NHANES data: Comparison of estimates for TSFAT across different variable subsets. The four panels correspond to the experimental groups: top-left – Group 1, top-right – Group 2, bottom-left – Group 3, and bottom-right – Group 4.

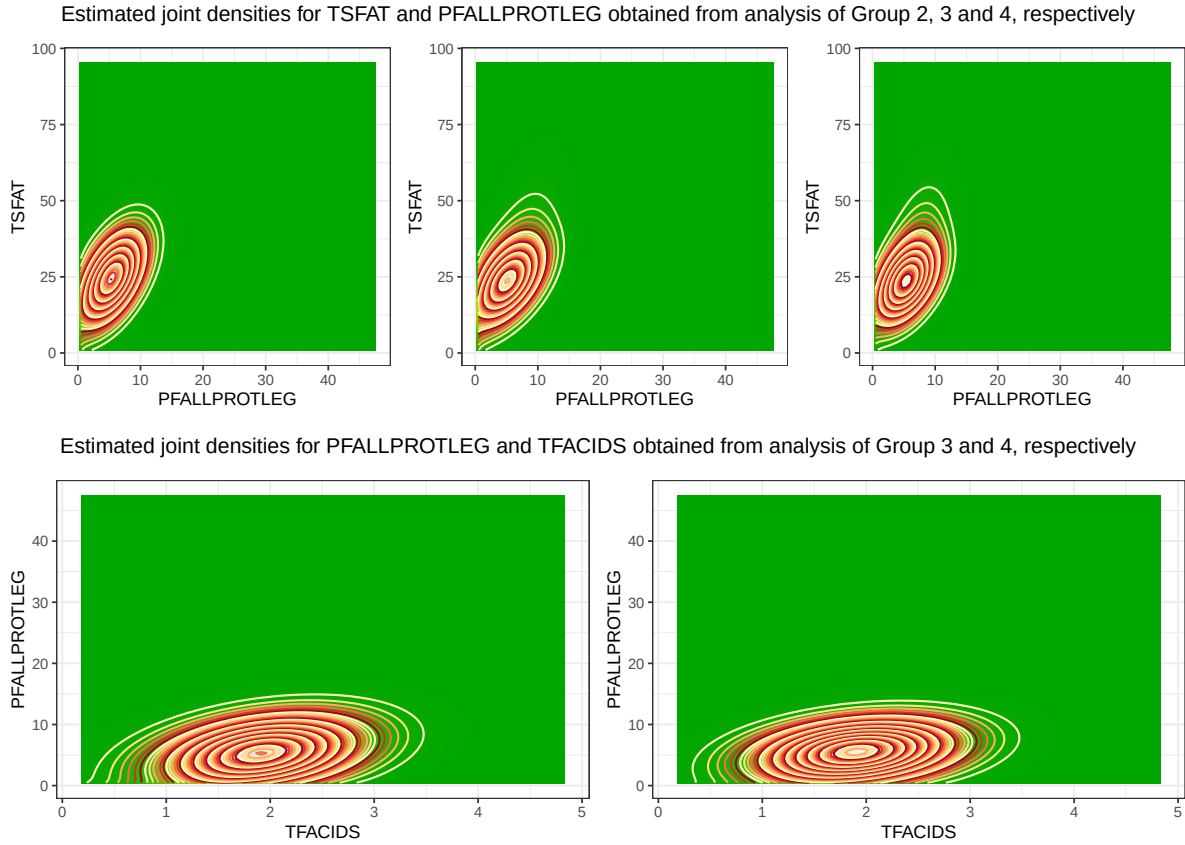


Figure S.5: Results for NHANES data: Comparison of bivariate joint density estimates across different variable subsets. The top row corresponds to the bivariate joint density estimates for TSFAT and PFALLPROTLEG: left - Group 2, middle - Group 3, right - Group 4. The bottom row corresponds to the bivariate joint density estimates for PFALLPROTLEG and TFACIDS: left - Group 3, right - Group 4.

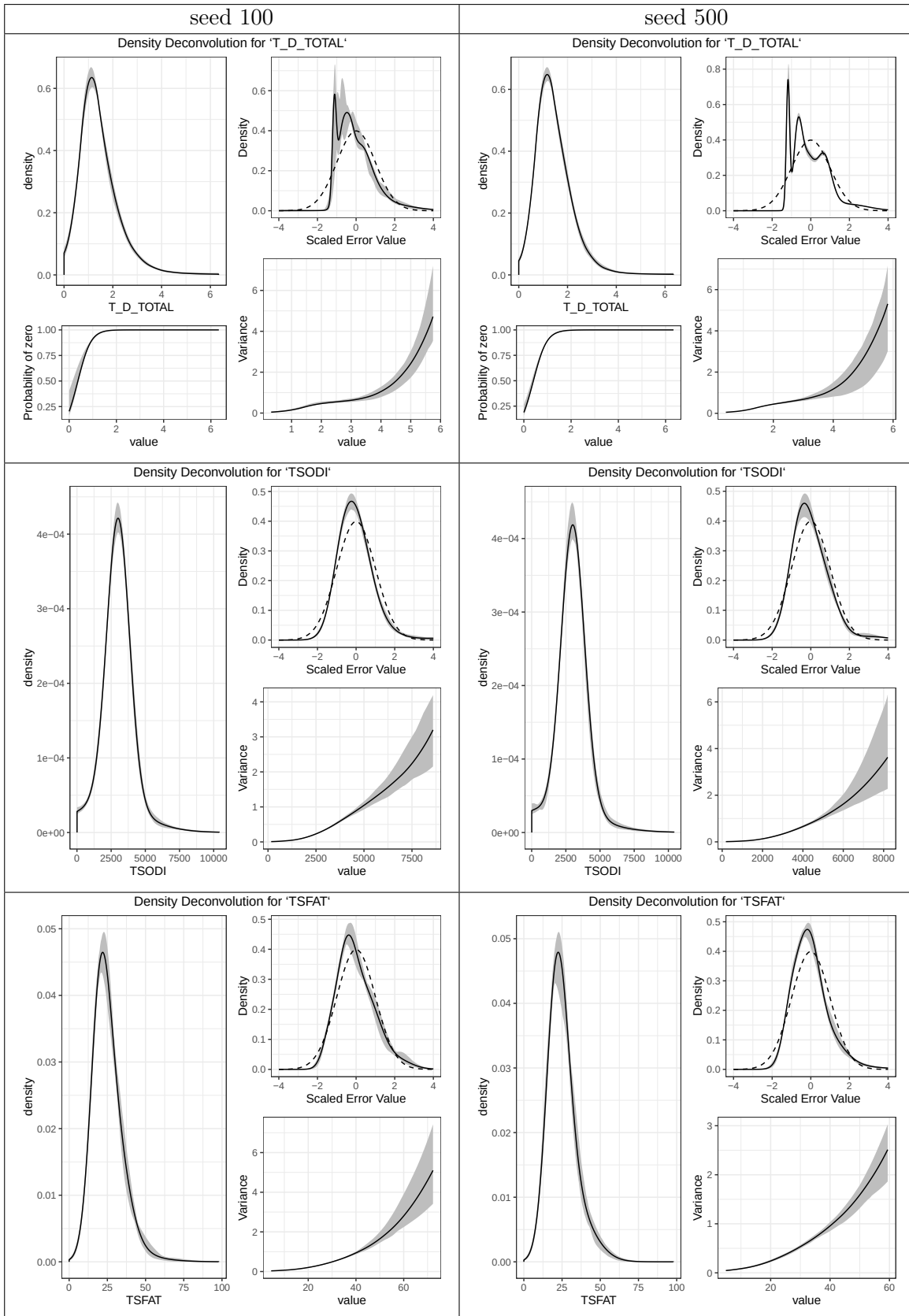


Figure S.6: Results for NHANES data: Comparison of estimates obtained from the `bdeconv` function using default hyper-parameters applied to NHANES data across two different seeds.

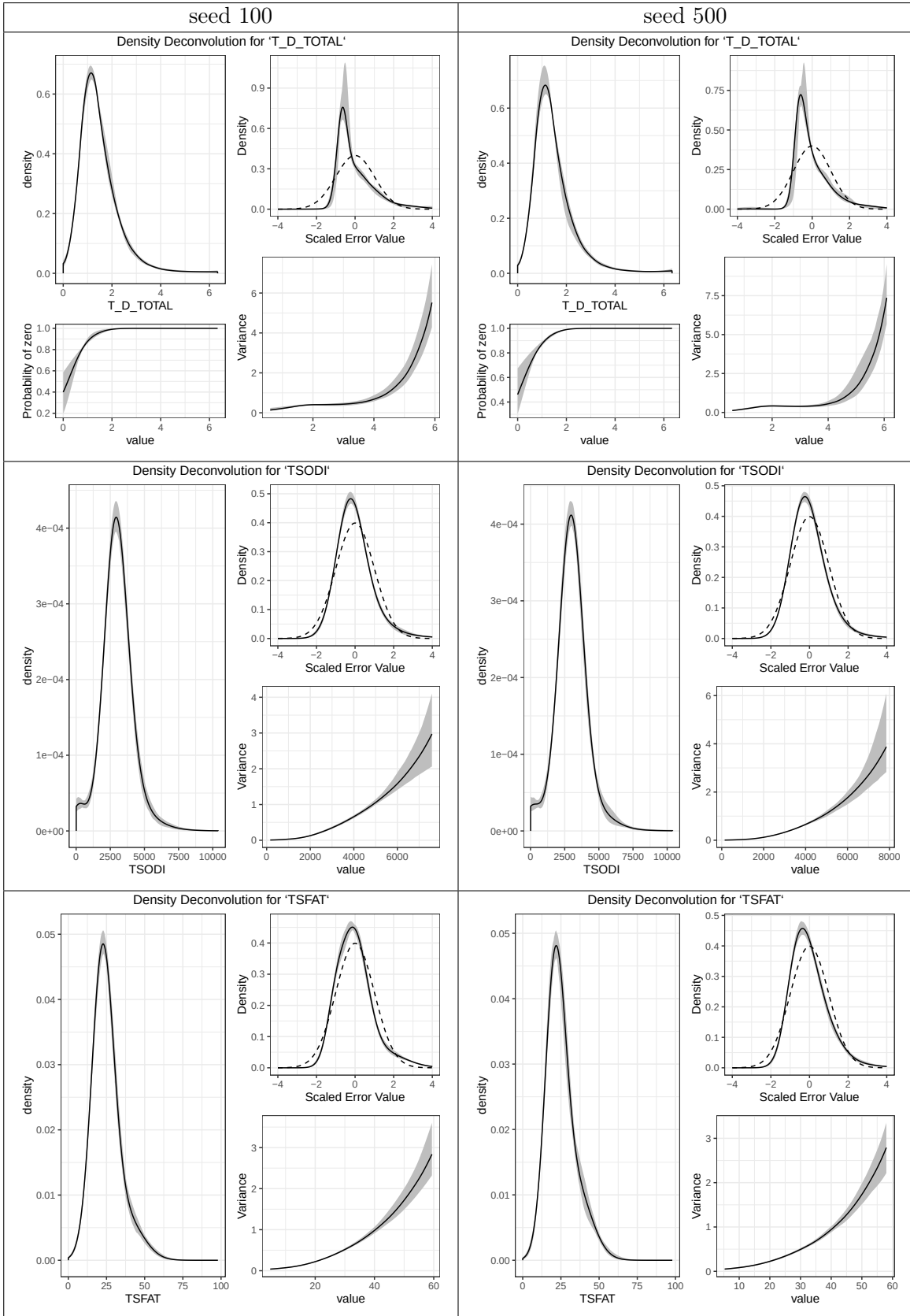


Figure S.7: Results for NHANES data: Comparison of estimates obtained from the `bdeconv` function using hyper-parameters inducing stronger prior regularization applied to NHANES data across two different seeds.