Supplementary Materials to *Bayesian Semiparametric Density Deconvolution in the Presence of Conditionally Heteroscedastic Measurement Errors*

Abhra Sarkar and Bani K. Mallick
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143
USA
abhra@stat.tamu.edu and bmallick@stat.tamu.edu

John Staudenmayer
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA
01003-9305 USA
jstauden@math.umass.edu

Debdeep Pati
Department of Statistics, Florida State University, Tallahassee, FL
32306-4330 USA
debdeep@stat.fsu.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143
USA
carroll@stat.tamu.edu

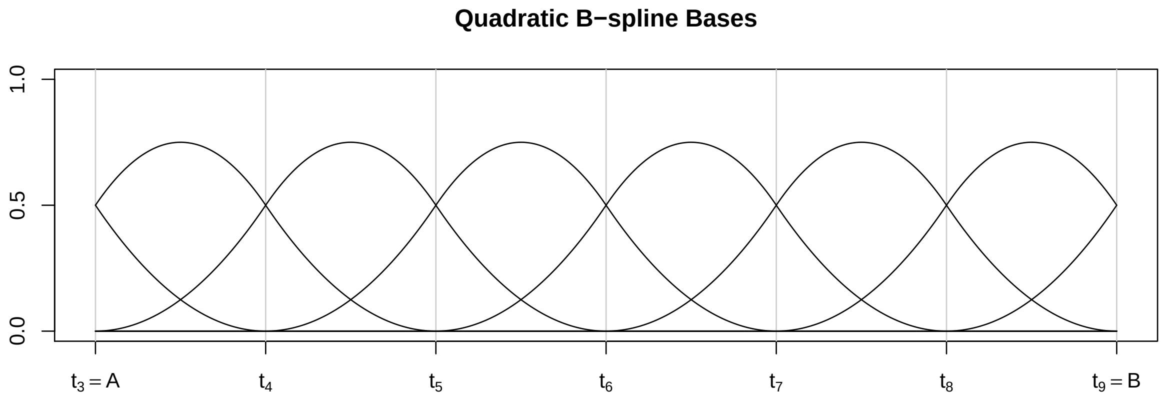


Figure S.1: Plot of 9 quadratic ($q = 2$) B-splines on $[A, B]$ defined using 11 knot points that divide $[A, B]$ into $K = 6$ equal subintervals.

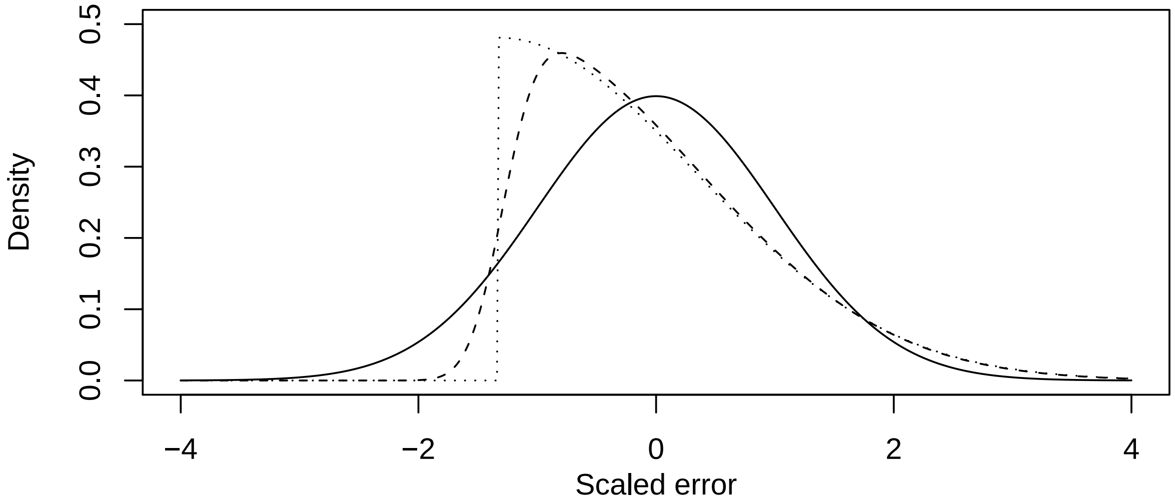


Figure S.2: Skew-normal densities with mean=0, variance=1 and varying skewness parameter λ. The solid line is the density of $\mathrm{SN}(\cdot \mid 0, 1, 0)$, the special case of standard normal distribution. The dashed line is the density of $\mathrm{SN}(\cdot \mid 0, 1, 7)$. The dotted line is the density of $\mathrm{SN}(\cdot \mid 0, 1, \infty)$ corresponding to the special case of a half-normal density.

		P-value combined from 4!=24 tests			
	Dietary Component	Truncation Limit $\varsigma = 0.05$		Truncation Limit $\varsigma = 0.50$	
		Kendall's τ Test	Spearman's ρ Test	Kendall's τ Test	Spearman's ρ Test
1	Calcium	1	1	0.511	0.984
2	Carbohydrate	1	1	0.824	1
3	Carotene	1	1	0.816	0.993
4	Cholesterol	1	1	0.978	1
5	Copper	1	1	0.982	1
6	Monosaturated Fat	1	1	0.777	1
7	Polysatuared Fat	1	1	1	1
8	Saturated Fat	1	1	0.987	1
9	Fiber	1	1	0.627	0.995
10	Folate	1	1	1	1
11	Iron	1	1	0.996	1
12	Magnesium	1	1	1	1
13	Niacin	1	1	0.910	0.999
14	Phosphorus	0.986	1	0.769	0.986
15	Potassium	1	1	0.989	1
16	Protein	1	1	0.969	1
17	Riboflavin	1	1	1	1
18	Sodium	1	1	0.856	0.999
19	Thiamin	1	1	1	1
20	Vitamin A	1	1	0.999	1
21	Vitamin B6	1	1	0.985	1
22	Vitamin B12	1	1	0.999	1
23	Vitamin C	0.980	1	0.507	0.970
24	Vitamin E	1	1	1	1
25	Zinc	1	1	1	1

Table S.1: Combined p-values for 4! = 24 nonparametric tests of association between W_{j_1} and $C_{j_2 j_3 j_4} = \{(W_{j_2} - W_{j_3})/(W_{j_2} - W_{j_4})\}$ for various $j_1 \neq j_2 \neq j_3 \neq j_4$ for 25 regularly consumed dietary components for which daily intakes were recorded in the EATS study. See Section 3 for additional details.

S.1 Quadratic B-Splines Used to Model Variance Functions in Section 2 of the Main Paper

Consider knot-points $t_1 = t_2 = t_3 = A < t_4 < \cdots < B = t_{K+3} = t_{K+4} = t_{K+5}$, where $t_{3:(K+3)}$ are equidistant with $\delta = (t_4 - t_3)$. For $j = 3, 4, \ldots, (K + 2)$, define

$$b_{2,j}(X) = \begin{cases} \{(X - t_j)/\delta\}^2/2 & \text{if } t_j \leq X < t_{j+1}, \\ -\{(X - t_{j+1})/\delta\}^2 + (X - t_{j+1})/\delta + 1/2 & \text{if } t_{j+1} \leq X < t_{j+2}, \\ \{1 - (X - t_{j+2})/\delta\}^2 & \text{if } t_{j+2} \leq X < t_{j+3}, \\ 0 & \text{otherwise.} \end{cases}$$

Also define

$$b_{2,1}(X) = \begin{cases} \{1 - (X - t_1)/\delta\}^2/2 & \text{if } t_3 \leq X < t_4, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{2,2}(X) = \begin{cases} -\{(X - t_3)/\delta\}^2 + (X - t_4)/\delta + 1/2 & \text{if } t_3 \leq X < t_4, \\ \{1 - (X - t_4)/\delta\}^2/2 & \text{if } t_4 \leq X < t_5, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{2,K+1}(X) = \begin{cases} \{(X - t_{K+1})/\delta\}^2/2 & \text{if } t_{K+1} \leq X < t_{K+2}, \\ -\{(X - t_{K+2})/\delta\}^2 + (X - t_{K+2})/\delta + 1/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{2,K+2}(X) = \begin{cases} \{(X - t_{K+2})/\delta\}^2/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

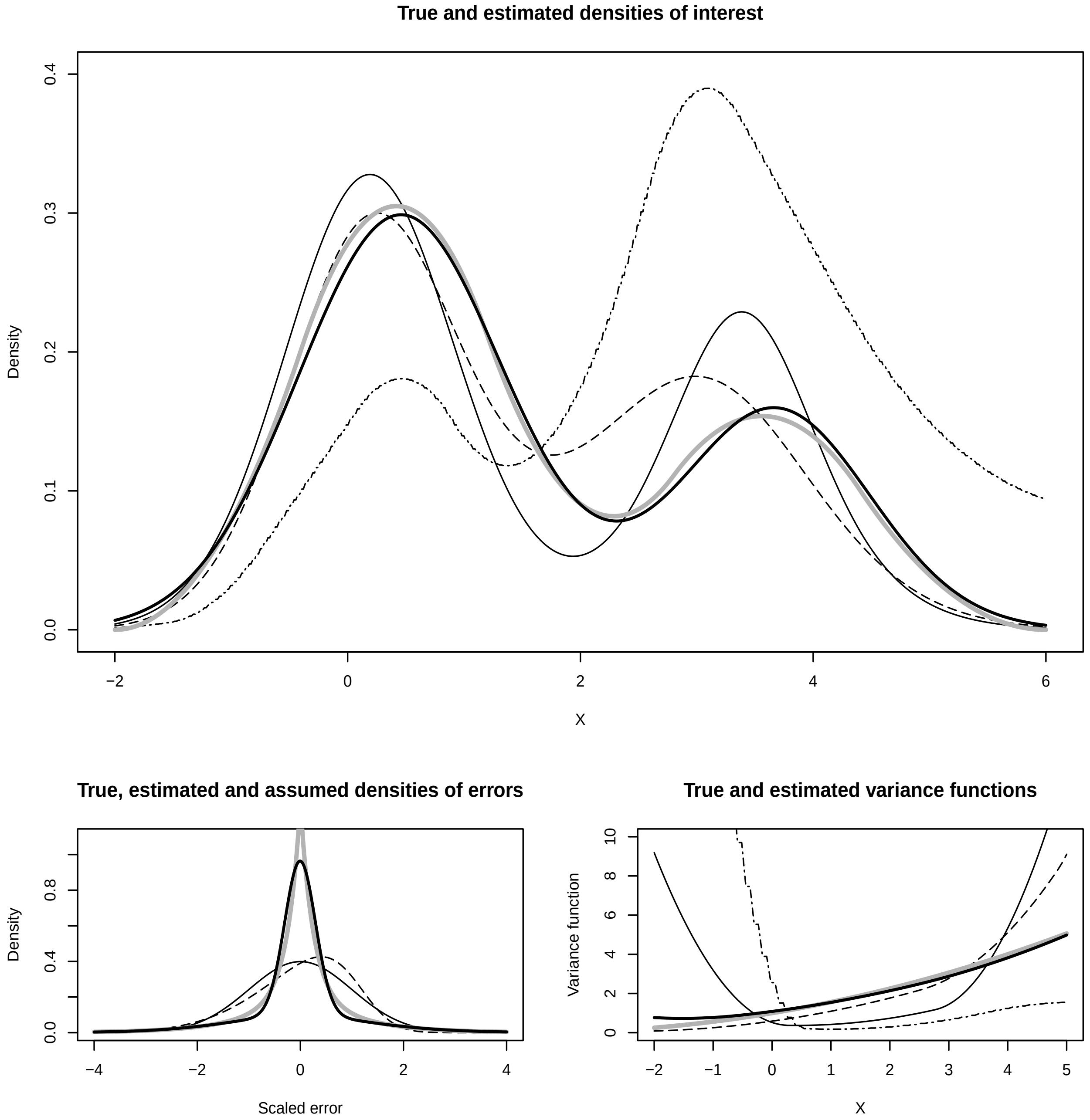
S.2 Additional Simulation Experiments

In this section, we present the results of additional simulation experiments when the true density of interest is a normalized mixture of B-splines: $f_X^3(X) \propto \sum_{k=1}^{7} b_{2,k}(X) c_k$ with $\mathbf{c} = (c_1, \ldots, c_7)^{\mathrm{T}} = (0, 0, 2, 0.1, 1, 0, 0)^{\mathrm{T}}$ and equidistant knots on $[-2, 6]$. The normalizing constant was estimated by numerical integration on a grid of 500 equidistant points in $[-2, 6]$. The true values of X were generated from f_X^3 using the inverse cumulative distribution function method. We recall that the SRB approach of Staudenmayer, et al. (2008) models f_X by normalized mixture of B-splines and assumes normality of the scaled errors. The SRB approach and the three methods we proposed in the main paper are compared over a factorial combination of three sample sizes ($n = 250, 500, 1000$), nine different types of distributions for the scaled errors (Table 1 and Figure 1), and one variance function $v(X) = (1 + X/4)^2$. For each subject, $m_i = 3$ replicates were simulated. The estimated MISEs are presented in Table S.2. Results for error distribution (i) are summarized in Figure S.3.

The results show that the deconvolution approaches proposed in Section 2 of the main paper outperform the SRB model in all 27 (3×9) cases, even in scenarios when the measurement errors were normally distributed and hence the truth actually conformed to the SRB model. This may be attributed to the fact that Models I, II and III estimate f_X by a flexible infinite mixture model, where the number of mixture components that are 'active' in the data is inferred semiautomatically from the data making it an adaptive data dependent approach. On the other hand, the SRB model estimates the density of interest by a mixture of normalized B-Splines with a fixed number of components. Model III, we recall, also relaxes parametric assumptions on the measurement errors, accommodating skewness, multimodality and heavy tails and resulting in huge reductions in MISE over other models when the measurement errors are heavy-tailed.

True Error Distribution	Sample Size	MISE ×1000			
		SRB	Model1	Model2	Model3
(a)	250	8.66	**4.58**	4.74	4.68
	500	4.80	**3.63**	3.74	3.87
	1000	4.03	**2.57**	2.75	2.68
(b)	250	9.13	5.77	**4.38**	4.48
	500	5.12	3.76	**3.53**	3.56
	1000	4.68	2.83	**2.50**	2.72
(c)	250	6.35	4.74	4.35	**4.16**
	500	6.08	3.15	3.85	**3.07**
	1000	3.93	2.54	2.96	**1.93**
(d)	250	6.31	5.17	5.95	**3.61**
	500	3.70	3.91	6.36	**2.70**
	1000	2.92	2.75	7.08	**2.03**
(e)	250	8.73	5.74	5.31	**4.06**
	500	7.42	5.63	3.70	**3.01**
	1000	7.99	3.37	2.35	**1.90**
(f)	250	8.86	5.32	5.39	**5.19**
	500	4.64	3.87	3.83	**3.12**
	1000	3.31	2.47	3.00	**2.35**
(g)	250	22.77	12.51	12.61	**3.45**
	500	19.66	17.66	17.09	**2.25**
	1000	40.55	22.66	16.36	**1.50**
(h)	250	11.15	6.61	6.38	**3.96**
	500	8.34	9.38	7.18	**3.22**
	1000	13.69	9.91	7.98	**2.03**
(i)	250	17.49	12.25	13.55	**3.28**
	500	32.99	20.40	15.19	**2.42**
	1000	40.67	19.47	12.18	**1.17**

Table S.2: Mean integrated squared error (MISE) performance of density deconvolution models described in Section 2 of this article (Models I, II and III) compared with the model of Staudenmayer, et al. (2008) (Model SRB) for different scaled error distributions when the true density of interest is a mixture of splines. The true variance function was $v(X) = (1 + X/4)^2$. See Section S.2 for additional details. The minimum value in each row is highlighted.


True and estimated densities of interest

True, estimated and assumed densities of errors

True and estimated variance functions

Figure S.3: Results for heavy-tailed error distribution (i) with sample size n=1000 corresponding to 25^{th} percentile MISE. The true density f_X is a normalized mixture of B-splines. See Section S.2 for additional details. The top panel shows the estimated densities under different models. The bottom left panel shows estimated densities of scaled errors under Model-II (dashed line) and Model-III (solid bold line) superimposed over a standard Normal density (solid line). The bottom right panel shows estimated variance functions under different models. For the top panel and the bottom right panel, the solid thin line is for Model-I; the dashed line is for Model-II; the solid bold line is for Model-III; and the dot-dashed line is for the Model of Staudenmayer, et al. (2008). In all three panels the bold gray lines represent the truth.