

Supplementary Document for Multiscale Adaptive Regression Models for Neuroimaging Data

University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

1. Simulation Studies

We conducted three sets of Monte Carlo simulations to examine the finite sample performance of $\hat{\beta}(d, h)$ and $W_\mu(d, h)$ with respect to different scales h at the levels of a single voxel and an entire region. The first two were based on simulated data on the 2D surface with the known ground truth. The third one was based on a 64×64 phantom image with four known activation regions.

1.1. Simulation Studies Part I

We simulated data at all $m = 4002$ points on the surface of a hippocampus for n subjects. At a given voxel d in \mathcal{D} , $y_i(d)$ was simulated according to $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta}(d) + \epsilon_i(d)$ for $i = 1, \dots, n$, where $\boldsymbol{\beta}(d) = (\beta_1(d), \beta_2(d), \beta_3(d))^T$, $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$, and the $\epsilon_i(d)$ were independently generated from $N(0, 1)$. We set $n = 60$ and $n = 80$. We generated x_{i2} independently from a Bernoulli distribution with the probability of success being 0.5, and generated x_{i3} independently from the uniform distribution in $[1, 2]$. The x_{i2} and x_{i3} were chosen to represent group identity and standardized age, respectively. We set $\beta_1(d) = \beta_3(d) = 0$ across all pixels d and created three different regions of interests (ROIs) by varying $\beta_2(d)$ as 0, 0.5, and 1.0, which represent different signal-to-noise ratios, to examine the performance of our method at different signal-to-noise ratios. Specifically, we created ROI1 and ROI2, which are two nested circles with radius at 3 and 5, respectively, and labeled the region outside of ROI1 and ROI2 as ROI3. We set $\beta_2(d)$ as 0 in ROI3, 0.5 in ROI2, and 1 in ROI1, respectively (Fig. 1(a)).

We fitted the linear model $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta}(d) + \epsilon_i(d)$, where $\epsilon_i(d) \sim N(0, \tau(d)^{-1})$. We used the MAET described in Example 4 to calculate adaptive parameter estimates across all voxels at 11 different scales ($h_s = 1.25^s$ and

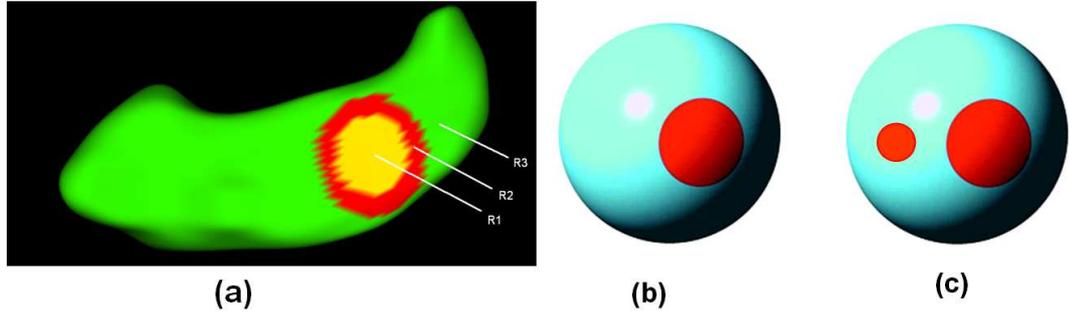


Fig. 1. Simulation study parts I and II: (a) three regions of interest ($R1$: ROI1 with yellow color; $R2$: ROI2 with red color; $R3$: ROI3 with green color) on a reference hippocampus; (b) a reference sphere with a red ROI; (c) a reference sphere with two red ROIs.

$h_0 = 0$). Since our primary interest is $\beta_2(d)$ and $\tau(d)$ was treated as nuisance parameters and fixed at $\hat{\tau}(d, h_0)$ after the first iteration. In each ROI, we calculated the average bias, the average empirical standard errors (RMS), and the average mean of the standard error estimates (SD) of $\hat{\beta}_2(d, h_s)$ based on the results from the 1,000 simulated hippocampus data sets. In ROI1 and ROI3, the biases of $\beta_2(d)$ slightly decrease from h_0 to h_{10} , whereas they fluctuate across all scales in ROI2. The biases of $\beta_2(d)$ are much smaller in ROI3 than in both ROI1 and ROI2. As s increases, the RMS and SD of $\beta_2(d)$ dramatically decrease. In addition, the RMS and its corresponding SD are relatively close to each other at all scales in each of the three ROIs (Table 1). As expected, increasing n decreases the bias, RMS and SD of parameter estimates $\hat{\beta}_2(d, h_s)$ (Table 1).

1.2. Simulation Studies Part II

In this simulation, we simulated data at all $m = 2064$ points on the surface of a reference sphere for n subjects. At a given voxel d in \mathcal{D} , a 2×1 vector $y_i(d)$ was simulated according to $y_i(d) = X_i \beta(d) + \epsilon_i(d)$, where $\beta(d) = (\beta_1(d), \dots, \beta_6(d))^T$ is a 6×1 vector, $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$, and

$$X_i = \begin{pmatrix} 1 & x_{i2} & x_{i3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_{i2} & x_{i3} \end{pmatrix}.$$

Error vectors $\epsilon_i(d) = (\epsilon_{i1}(d), \epsilon_{i2}(d))^T$ were independently generated from $N(\mathbf{0}, \mathbf{I}_2)$, in which \mathbf{I}_2 is a 2×2 identity matrix. We also simulated errors $\epsilon_{i1}(d)$ and $\epsilon_{i2}(d)$ independently from a $\chi^2(3) - 3$ distribution. Then the errors were smoothed by using heat kernel smoothing with parameters 1 and 4 iterations, which gave an effective

Table 1. Average BIAS ($\times 10^{-3}$), RMS, SD, and RE of $\hat{\beta}_2(d, h_s)$ in the three ROIs at 11 different scales ($h_0 - h_{10}$), $N(0, 1)$ distributed data, and 2 different sample sizes ($n = 60, 80$). BIAS denotes the bias of the mean of estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RE denotes the ratio of RMS over SD.

	ROI1				ROI2				ROI3			
	BIAS	RMS	SD	RE	BIAS	RMS	SD	RE	BIAS	RMS	SD	RE
$n = 60, N(0, 1)$												
h_0	-0.01	0.44	0.42	1.03	7.49	0.43	0.42	1.01	-0.49	0.43	0.42	1.02
h_1	38.80	0.19	0.18	1.06	-24.96	0.21	0.19	1.08	-0.08	0.19	0.18	1.07
h_2	28.46	0.18	0.17	1.05	-5.38	0.19	0.18	1.06	-0.54	0.18	0.17	1.07
h_3	23.22	0.18	0.17	1.05	-1.72	0.18	0.17	1.06	-0.57	0.17	0.15	1.07
h_4	19.76	0.17	0.16	1.05	0.60	0.17	0.16	1.06	-0.58	0.16	0.14	1.07
h_5	17.23	0.16	0.16	1.05	2.29	0.16	0.15	1.06	-0.58	0.14	0.13	1.07
h_6	14.70	0.16	0.15	1.05	3.54	0.15	0.14	1.06	-0.58	0.13	0.13	1.07
h_7	12.38	0.15	0.14	1.05	4.30	0.14	0.13	1.06	-0.57	0.12	0.12	1.07
h_8	10.13	0.14	0.13	1.04	4.83	0.14	0.13	1.06	-0.55	0.12	0.11	1.07
h_9	8.64	0.13	0.12	1.04	5.15	0.13	0.12	1.06	-0.55	0.11	0.10	1.07
h_{10}	7.82	0.12	0.12	1.04	5.35	0.12	0.11	1.06	-0.56	0.10	0.09	1.07
$n = 80, N(0, 1)$												
h_0	-3.02	0.37	0.37	1.01	2.90	0.37	0.37	1.01	-0.21	0.38	0.37	1.01
h_1	34.45	0.16	0.16	1.04	-23.62	0.18	0.17	1.05	0.24	0.16	0.16	1.05
h_2	20.64	0.16	0.15	1.03	-6.02	0.17	0.16	1.03	-0.09	0.15	0.14	1.04
h_3	14.37	0.15	0.15	1.03	-2.50	0.15	0.15	1.03	-0.09	0.14	0.13	1.04
h_4	10.41	0.15	0.14	1.03	-0.49	0.14	0.14	1.03	-0.09	0.13	0.13	1.04
h_5	7.68	0.14	0.14	1.03	0.68	0.14	0.13	1.03	-0.10	0.12	0.12	1.04
h_6	5.47	0.13	0.13	1.03	1.41	0.13	0.12	1.02	-0.09	0.11	0.11	1.04
h_7	3.81	0.12	0.12	1.02	1.84	0.12	0.12	1.02	-0.08	0.11	0.10	1.04
h_8	2.54	0.12	0.11	1.02	2.06	0.12	0.11	1.02	-0.08	0.10	0.09	1.04
h_9	1.78	0.11	0.11	1.02	2.10	0.11	0.11	1.02	-0.07	0.09	0.09	1.05
h_{10}	1.39	0.11	0.10	1.02	2.14	0.10	0.10	1.02	-0.08	0.08	0.08	1.05

Table 2. Simulation Study for $W_\mu(d, h)$: estimates (ES) and standard errors (SE) of rejection rates for pixels inside and outside ROIs were reported at 3 different scales (h_0, h_5, h_{10}), $N(0, 1)$ and $\chi^2(3) - 3$ distributed data, and 2 different sample sizes ($n = 60, 80$) at $\alpha = 5\%$. For each case, 2,000 simulated datasets were used.

s	$n = 60, N(0, 1)$		$n = 80, N(0, 1)$		$n = 60, \chi^2(3) - 3$		$n = 80, \chi^2(3) - 3$		
	ROI	Inside	Outside	Inside	Outside	Inside	Outside	Inside	Outside
		ES (SE)	ES (SE)	ES (SE)	ES (SE)	ES (SE)	ES (SE)	ES (SE)	ES (SE)
h_0		0.29 (0.05)	0.06 (0.01)	0.34 (0.06)	0.06 (0.01)	0.11 (0.04)	0.06 (0.01)	0.11 (0.04)	0.05 (0.01)
h_5		0.72 (0.05)	0.07 (0.01)	0.76 (0.04)	0.07 (0.01)	0.29 (0.11)	0.07 (0.01)	0.33 (0.12)	0.07 (0.01)
h_{10}		0.78 (0.04)	0.08 (0.02)	0.79 (0.03)	0.07 (0.02)	0.45 (0.15)	0.08 (0.02)	0.52 (0.16)	0.07 (0.02)

smoothness of about 2 points (Chung *et al.*, 2005). We generated x_{i2} independently from a Bernoulli distribution with an equal probability and generated x_{i3} independently from the uniform distribution in $[0, 1]$. We set $n = 60$ and $n = 80$. To assess both Type I and II error rates at the voxel level, we selected two region-of-interest (ROI). One is with 64 points on the reference sphere and the other one with 17 points on the reference sphere. We set $\beta(d) = \mathbf{0}_6$ across the whole sphere and then change $\beta_{12}(d)$ from 0 to 0.5 for all points d in ROIs (Fig 1(b)).

We fitted the same multivariate linear model $y_i(d) = X_i\beta(d) + \epsilon_i(d)$, where $\epsilon_i(d) \sim N(\mathbf{0}, \Sigma(d))$, where $\Sigma(d)$ is a 2×2 covariance matrix. We test the hypotheses $H_0 : \beta_2(d) = 0$ and $H_1 : \beta_2(d) \neq 0$ across all points. Since our primary interest is $\beta(d)$ and the elements in $\Sigma(d)$ can be treated as nuisance parameters, we fixed $\Sigma(d)$ at $\hat{\Sigma}(d, h_0)$ after the first iteration. We applied the MAET procedure described in Example 4 to calculate $\hat{\beta}(d, h_s)$ and $W(d, h_s)$ across all voxels at 11 different scales. The 2,000 replications were used to calculate the estimates and standard errors of the rejection rates with the significance level $\alpha = 5\%$. For the test statistic $W_\mu(d, h)$, the Type I rejection rates outside the ROI were relatively accurate for all radius, while the statistical power for rejecting the null hypothesis in the ROI significantly increased with the radius h (Table 2).

1.3. Simulation Studies Part III

We simulated data at all $m = 4096$ pixels on the 64×64 phantom image for n subjects. At a given pixel d in \mathcal{D} , $y_i(d)$ was simulated according to $y_i(d) = \mathbf{x}_i^T \beta(d) + \epsilon_i(d)$ for $i = 1, \dots, n$, where $\beta(d) = (\beta_1(d), \beta_2(d), \beta_3(d))^T$ and $\mathbf{x}_i = (1, x_{i2}, x_{i3})^T$. Errors $\epsilon_i(d)$ were first independently generated from $N(0, 1)$ and $\chi^2(3) - 3$, respectively, and then they were smoothed by using heat kernel smoothing with parameters 1 and 4 iterations, which gave

an effective smoothness of about 2 pixels (Chung *et al.*, 2005). The $\chi^2(3) - 3$ distribution is a very skewed distribution. We set $n = 60$ and $n = 80$. We generated x_{i2} independently from a Bernoulli distribution with probability of success being 0.5, and generated x_{i3} independently from the uniform distribution on $[1, 2]$. The x_{i2} and x_{i3} were chosen to represent group identity and standardized age, respectively. We set $\beta_1(d) = \beta_3(d) = 0$ across all pixels d and created five different regions of interests (ROIs) with different shapes by varying $\beta_2(d)$ as 0, 0.2, 0.4, 0.6 and 0.8, which represent different signal-to-noise ratios, to examine the performance of our method at different signal-to-noise ratio and also to test whether MARM can perform well for different shapes. The true $\beta_2(d)$ was displayed for all ROIs with black, blue, red, yellow, and white colors representing $\beta_2(d)=0, 0.2, 0.4, 0.6,$ and $0.8,$ respectively (Fig. 2(k)).

We fit the linear model $y_i(d) = \mathbf{x}_i^T \boldsymbol{\beta}(d) + \epsilon_i(d)$, where $\epsilon_i(d) \sim N(0, \tau(d)^{-1})$. We applied the MAET procedure described in Example 4 to calculate adaptive parameter estimates across all pixels at 10 different scales. We calculated the bias, the empirical standard error (RMS), the mean of the standard error estimate (SD), and the ratio of RMS over SD (RE) at each pixel of all five ROIs based on the results obtained from the 1,000 simulated data sets. The biases are slightly increased from h_0 to h_{10} (Figs. 2-5 (b) and (g)), whereas RMS and SD at h_{10} are much smaller than those at h_0 (Figs. 2-5 (c), (d), (h), and (i)). In addition, the RMS and its corresponding SD are relatively close to each other at all scales for both the normal and Chi-square distributed data (Figs. 2-5 (e) and (j)). The biases, SDs, and RMSs of $\beta_2(d)$ are smaller in the normal distributed data than in the chi-square distributed data, because the SNRs in the normal distributed data are 2.45 times bigger than the SNRs in the chi-square distributed data. Moreover, SDs in these pixels of ROIs with $\beta_2(d) \neq 0$ are larger than SDs in those pixels of ROI with $\beta_2(d) = 0$ (Figs. 2-5 (i)), because the interior of ROI with $\beta_2(d) = 0$ contains more pixels. MAET performs better in ROIs with relatively large $\beta_2(d)$ values than in ROIs with small $\beta_2(d)$ values. The MAET procedure can really adapt to the shapes and boundaries of ROIs. Increasing sample size and signal-to-noise ratio decreases the bias, RMS, and SD of parameter estimates.

To compare with the conventional analyses of imaging data, we first smoothed 1,000 imaging data by using heat kernel smoothing with parameters 1 and 16 iterations, which gave an effective smoothness of about 4 pixels (Chung *et al.*, 2005), and then we fit the linear model at each voxel. We calculated the bias, the empirical standard error (RMS), the mean of the standard error estimate (SD), and the ratio of RMS over SD (RE) at each pixel of all five ROIs based on the results obtained from the 1,000 smoothed data sets. We calculated the average bias,

Table 3. Average Bias, RMS, SD, and RE of $\beta_2(d)$ parameters obtained from the conventional analyses in the five ROIs, $N(0, 1)$ and $\chi^2(3) - 3$ distributed data, and 2 different sample sizes ($n = 60, 80$). BIAS denotes the bias of the mean of estimates; RMS denotes the root-mean-square error; SD denotes the mean of the standard deviation estimates; RE denotes the ratio of RMS over SD. For each case, 1,000 simulated datasets were used.

		$\beta = 0.2$	0.4	0.6	0.8	0.0
$\chi^2(3) - 3, n = 60$	Bias	0.02	-0.15	-0.23	-0.27	0.02
	RMS	0.19	0.17	0.17	0.17	0.19
	SD	0.18	0.17	0.17	0.17	0.18
	RE	1.02	0.98	1.01	1.01	1.02
$\chi^2(3) - 3, n = 80$	Bias	-0.08	-0.13	-0.22	-0.26	0.02
	RMS	0.14	0.15	0.15	0.15	0.16
	SD	0.15	0.15	0.15	0.15	0.16
	RE	0.97	1.01	1.01	1.00	1.01
$N(0, 1), n = 60$	Bias	-0.08	-0.14	-0.22	-0.27	0.02
	RMS	0.07	0.07	0.07	0.07	0.07
	SD	0.07	0.07	0.07	0.07	0.07
	RE	1.02	1.02	1.03	1.00	1.02
$N(0, 1), n = 80$	Bias	-0.08	-0.14	-0.22	-0.26	0.02
	RMS	0.06	0.06	0.06	0.06	0.06
	SD	0.06	0.06	0.06	0.06	0.06
	RE	0.99	0.98	1.02	1.03	1.01

RMS, SD, and RE in each of the five ROIs and presented in Table 3. For all ROIs, the biases are significantly larger in the smoothed data sets (Table 3) than those in the Table 1 of Li *et al.*, (2010). The biases of $\beta_2(d)$ are comparable in both the normal and chi-square distributed data (Table 3). Compared with the conventional analyses, MARM performs much better in preserving the edges of ROIs (Fig. 6).

To assess both Type I and II error rates at the pixel level, we tested the hypotheses $H_0 : \beta_2(d) = 0$ and $H_1 : \beta_2(d) \neq 0$ across all pixels. We calculated the estimates and standard errors of rejection rates with significance level $\alpha = 5\%$ based on the 1,000 smoothed images. For the test statistic $W_\mu(d, h)$, the Type I rejection rates in ROI with $\beta_2(d) = 0$ were much larger than the significance level 5% in the normal distributed data. Moreover,

Table 4. Simulation Study for $W_\mu(d, h)$: estimates (ES) and standard errors (SE) of rejection rates for pixels in the five ROIs were reported at $N(0, 1)$ and $\chi^2(3) - 3$ distributed data, and 2 different sample sizes ($n = 60, 80$) at $\alpha = 5\%$. For each case, 1,000 simulated datasets were used.

β	$n = 60, N(0, 1)$		$n = 80, N(0, 1)$		$n = 60, \chi^2(3) - 3$		$n = 80, \chi^2(3) - 3$	
	ES	SE	ES	SE	ES	SE	ES	SE
0.2	0.35	0.139	0.42	0.122	0.11	0.095	0.13	0.101
0.4	0.86	0.073	0.91	0.066	0.30	0.170	0.44	0.181
0.6	0.96	0.031	0.98	0.018	0.52	0.093	0.64	0.096
0.8	0.99	0.009	1.00	0.001	0.77	0.096	0.78	0.193
0.0	0.11	0.012	0.12	0.013	0.07	0.013	0.04	0.032

the initial smoothing step can really increase the statistical power for rejecting the null hypothesis in ROIs with $\beta_2(d) \neq 0$ (Table 4).

2. Assumptions and Proof

The following assumptions are needed to facilitate development of our methods, although they are not the weakest possible conditions.

(C1) $1 \geq \omega(d, d'; h) \geq 0$ and $\omega(d, d; h) = 1$ for all $d, d' \in \mathcal{D}$ and $h \geq 0$.

(C2) The data $\{\mathbf{Z}_i = (\mathbf{x}_i, \mathbf{Y}_{i, \mathcal{D}}) : i = 1, \dots, n\}$ form an independent and identical sequence.

(C3) For any $d \in \mathcal{D}$, the maximum $\boldsymbol{\theta}_*(d)$ of $E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]$ is a unique interior point of \mathcal{B} , where \mathcal{B} is a compact set in R^p and the expectation is taken with respect to the true distribution of $Y(d)$ given \mathbf{x} .

(C4) For all voxels $d \in \mathcal{D}$, $\ell(\boldsymbol{\theta}(d)) = \log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))$ is twice continuously differentiable on Θ . For all $j, k, l = 1, \dots, p$, $\ell(\boldsymbol{\theta}(d))$, $|\partial_j \ell(\boldsymbol{\theta}(d))|^2$, and $|\partial_j \partial_k \ell(\boldsymbol{\theta}(d))|^2$ are dominated by an integral function $G(Y(d), \mathbf{x})$ such that $E[\max_{d \in \mathcal{D}} |G(Y(d), \mathbf{x})|^r] < \infty$ for a $r > 1$, where $\partial_j = \partial / \partial \theta_j(d)$, in which $\theta_j(d)$ is the j -th component of $\boldsymbol{\theta}(d)$.

(C5) For a fixed $\delta > 0$,

$$\begin{aligned} \infty &> \sup_{d \in \mathcal{D}} \max_{\boldsymbol{\theta}(d) \in B(\boldsymbol{\theta}_*(d), \delta)} (\lambda_{\max}\{E[-\partial_{\boldsymbol{\theta}(d)}^2 \ell(\boldsymbol{\theta}(d))]\}) \geq \inf_{d \in \mathcal{D}} \min_{\boldsymbol{\theta}(d) \in B(\boldsymbol{\theta}_*(d), \delta)} (\lambda_{\min}\{E[-\partial_{\boldsymbol{\theta}(d)}^2 \ell(\boldsymbol{\theta}(d))]\}) > 0, \\ \infty &> \sup_{d \in \mathcal{D}} \max_{\boldsymbol{\theta}(d) \in B(\boldsymbol{\theta}_*(d), \delta)} (\lambda_{\max}\{E[\partial_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))^{\otimes 2}]\}) \geq \inf_{d \in \mathcal{D}} \min_{\boldsymbol{\theta}(d) \in B(\boldsymbol{\theta}_*(d), \delta)} (\lambda_{\min}\{E[\partial_{\boldsymbol{\theta}(d)} \ell(\boldsymbol{\theta}(d))^{\otimes 2}]\}) > 0, \end{aligned}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix, respectively.

(C6) The kernel functions $K_{st}(u)$ and $K_{loc}(u)$ are continuous decreasing functions of $u \geq 0$ such that $K_{st}(0) = K_{loc}(0) = 1$, $\lim_{u \rightarrow \infty} K_{st}(u) = \lim_{u \rightarrow \infty} K_{loc}(u) = 0$, and $\lim_{u \rightarrow \infty} u^{1/2} K_{st}(u) = 0$.

$$(C7) \lim_{n \rightarrow \infty} C_n/n = \lim_{n \rightarrow \infty} C_n^{-1} \log(N(\mathcal{D})) = \lim_{n \rightarrow \infty} C_n^{-1} = 0.$$

Remarks A1: Assumption (C2) is needed just for notational simplicity and can be easily modified to accommodate independent and non-identical distributed settings. For each fixed $d \in \mathcal{D}$, Assumptions (C3)-(C5) are generalizations of the standard conditions for ensuring first order asymptotic properties (e.g., consistency and asymptotic normality) of M-estimators (van der Vaart, 1998). Assumption (C3) is an identification condition, whereas Assumption (C4) is a uniform smoothness and integration condition. Particularly, Assumption (C4) ensures that $\ell(\boldsymbol{\theta}(d))$, $|\partial_j \ell(\boldsymbol{\theta}(d))|^2$, and $|\partial_k \partial_j \ell(\boldsymbol{\theta}(d))|^2$ are uniformly integrable for all $d \in \mathcal{D}$. Assumption (C5) is needed to ensure that the covariance matrix of $\hat{\boldsymbol{\theta}}(d, h)$ is positive definite for all $d \in \mathcal{D}$. Assumptions (C6) and (C7) on $K_{st}(\cdot)$ and $K_{loc}(\cdot)$ are needed just for ensuring desirable asymptotic properties of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ based on the stochastic weights for the AET procedure.

Remarks A2: Assumption (C7) ensures that $\lim_{n \rightarrow \infty} \log(N(\mathcal{D}))/n = 0$. In neuroimaging data, although $N(\mathcal{D})$ is much larger than the sample size n , Assumption (C7) claims that we just need a relatively large sample size compared to $\log(N(\mathcal{D}))$. For instance, in most neuroimaging data, $N(\mathcal{D}) \approx 100^3$ and $\log(10^3) = 14$. Therefore, a sample size such as 100 may be reasonable to use asymptotic normality in making statistical inferences for MARM. Assumption (C7) is needed to invoke maximal inequalities (van der Vaart and Wellner, 1996). Moreover, Assumption (C7) also requires a large value of C_n relative to $\log N(\mathcal{D})$, but it may be weakened by assuming spatial smoothness in the neuroimaging data.

Remarks A3: We first discuss the asymptotic properties of $\hat{\boldsymbol{\theta}}(d, h)$ and $W_\mu(d, h)$ in the case with fixed weights $\omega(d, d'; h)$ for a fixed scale h . Let $Y_i(d, h) = (Y_i(d') : d' \in B(d, h))$ for $i = 1, \dots, n$. Without loss of generality, we assume that the $(Y_i(d, h), \mathbf{x}_i)$ are independently and identically distributed as the true density $p(Y(d, h), \mathbf{x})$.

The MWQL estimator $\hat{\boldsymbol{\theta}}(d, h)$ maximizes the function $n^{-1}\ell_n(\boldsymbol{\theta}(d); h, \tilde{\omega})$, which converges to

$$M(\boldsymbol{\theta}(d); h, \tilde{\omega}) = \sum_{d' \in B(d, h)} \tilde{\omega}(d, d'; h) E[\log p(Y(d')|\mathbf{x}, \boldsymbol{\theta}(d))] \quad (1)$$

in probability (or almost surely) under some mild conditions as $n \rightarrow \infty$, where the expectation is taken with respect to $p(Y(d, h), \mathbf{x})$. Under some identifiability conditions, $\hat{\boldsymbol{\theta}}(d; h)$ converges to $\boldsymbol{\theta}_*(d; h)$, which maximizes $M(\boldsymbol{\theta}(d); h, \tilde{\omega})$ (van der Vaart, 1998). When $h = 0$, $\boldsymbol{\theta}_*(d; 0) = \boldsymbol{\theta}_*(d)$ is the ‘pseudo’ true value in voxel d . When $h > 0$, $\boldsymbol{\theta}_*(d; h)$ can only be regarded as a weighted combination of all $\boldsymbol{\theta}_*(d')$ for $d' \in B(d, h)$. In a homogeneous region, that is $\boldsymbol{\theta}_*(d') = \boldsymbol{\theta}_*(d)$, $\boldsymbol{\theta}_*(d; h) = \boldsymbol{\theta}_*(d)$ even for $h > 0$. However, in a nonhomogeneous region, an arbitrary set of weights $\omega(d, d'; h)$ can lead to undesirable consequences, such as smoothing out the boundary of activated regions and reducing statistical power in detecting activated regions.

Proof of Theorem 1. The proof of Theorem 1 consists of three steps. In Step 1, we will show that $\hat{\boldsymbol{\theta}}(h_0) = (\hat{\boldsymbol{\theta}}(d, h_0) : d \in \mathcal{D})$ converges $\boldsymbol{\theta}_* = (\boldsymbol{\theta}_*(d) : d \in \mathcal{D})$ in probability. We need to introduce some notation. Let \mathbf{T} be a bounded brain region in R^g containing all voxels $d \in \mathcal{D}$, where $g = 2$ for the 2D surface and $g = 3$ for the 3D volume. Let $\Theta = \prod_{d \in \mathcal{D}} \mathcal{B}$ be the parameter space for $\boldsymbol{\theta}$ and $\ell^\infty(\mathbf{T})^p$ is the product of p $\ell^\infty(\mathbf{T}) = \{z : \mathbf{T} \rightarrow R, \sup_{\mathbf{t} \in \mathbf{T}} |z(\mathbf{t})| < \infty\}$. Let $\Psi_n : \Theta \rightarrow \ell^\infty(\mathbf{T})^p$ and $\Psi : \Theta \rightarrow \ell^\infty(\mathbf{T})^p$ be random maps and a deterministic map, respectively, such that

$$\Psi_n(\boldsymbol{\theta})(\mathbf{t}) = n^{-1} \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d_{\mathbf{t}})} \log p(Y_i(d_{\mathbf{t}})|\mathbf{x}_i, \boldsymbol{\theta}(d_{\mathbf{t}})) \quad \text{and} \quad \Psi(\boldsymbol{\theta})(\mathbf{t}) = E[\partial_{\boldsymbol{\theta}(d_{\mathbf{t}})} \log p(Y(d_{\mathbf{t}})|\mathbf{x}, \boldsymbol{\theta}(d_{\mathbf{t}}))],$$

in which $d_{\mathbf{t}}$ denotes the voxel covering \mathbf{t} .

To prove the consistency of $\hat{\boldsymbol{\theta}}(h_0)$, we will show that

$$\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{t} \in \mathbf{T}} \|\Psi_n(\boldsymbol{\theta})(\mathbf{t}) - \Psi(\boldsymbol{\theta})(\mathbf{t})\|_2 \rightarrow 0 \quad \text{and} \quad \inf_{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\| \geq \epsilon} \sup_{\mathbf{t} \in \mathbf{T}} \|\Psi(\boldsymbol{\theta})(\mathbf{t})\|_2 > \sup_{\mathbf{t} \in \mathbf{T}} \|\Psi(\boldsymbol{\theta}_*)(\mathbf{t})\|_2. \quad (2)$$

It follows from Assumptions (C3) and (C4) that the second term in equation (2) is true. To prove the first term in equation (2), we note that

$$\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{t} \in \mathbf{T}} \|\Psi_n(\boldsymbol{\theta})(\mathbf{t}) - \Psi(\boldsymbol{\theta})(\mathbf{t})\|_2 = \max_{d \in \mathcal{D}} A_n(d), \quad (3)$$

where $A_n(d) = \sup_{\boldsymbol{\theta}(d) \in \mathcal{B}} |n^{-1} \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d)) - E[\partial_{\boldsymbol{\theta}(d)} \log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]|$. Then, we consider $\mathcal{F} = \{\partial_{\boldsymbol{\theta}(d)} \log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d)) : d \in \mathcal{D}, \boldsymbol{\theta}(d) \in \mathcal{B}\}$ with an envelope $\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x})$. Following the arguments in Theorem 2.4.3 of van der Vaart and Wellner (1996), we can show that $E[\max_{d \in \mathcal{D}} A_n(d)]$ is bounded above by

$$\sqrt{[1 + p \log(C_1(\epsilon)K) + \log(N(\mathcal{D}))]/n} C_2 K + 2E[\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x}) \mathbf{1}\{\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x}) > K\}] + \epsilon \rightarrow 0,$$

where C_2 is a constant independent of ϵ , K can be chosen such that the second term of the above equation is arbitrarily small, and $C_1(\epsilon)$ is a constant depending on ϵ . Finally, following the arguments in Theorems 5.7 and 5.9 of van der Vaart (1998), we can prove consistency of $\hat{\boldsymbol{\theta}}(h_0)$.

In Step 2, we will prove the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}}(h_0) - \boldsymbol{\theta}_*)$. For each $d \in \mathcal{D}$, a Taylor's series expansion gives

$$\mathbf{0} = \Psi_n(\hat{\boldsymbol{\theta}}(h_0))(d) = \Psi_n(\boldsymbol{\theta}_*)(d) + \partial_{\boldsymbol{\theta}(d)}\Psi_n(\tilde{\boldsymbol{\theta}})(d)[\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)], \quad (4)$$

where $\tilde{\boldsymbol{\theta}} \in \Theta$ and $\tilde{\boldsymbol{\theta}}(d)$ is on the line connecting $\boldsymbol{\theta}(d)$ and $\boldsymbol{\theta}_*(d)$. Similar to the proof of (3), we can show that

$$\sup_{\boldsymbol{\theta} \in \Theta: \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \leq \epsilon} \sup_{\mathbf{t} \in \mathbf{T}} \|\partial_{\boldsymbol{\theta}(d_t)}\Psi_n(\boldsymbol{\theta})(\mathbf{t}) - \partial_{\boldsymbol{\theta}(d_t)}\Psi(\boldsymbol{\theta})(\mathbf{t})\|_2 \rightarrow 0 \quad (5)$$

in probability, when $\log(N(\mathcal{D}))/n$ is sufficiently small. Therefore, we can show that

$$\sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)] = [-\partial_{\boldsymbol{\theta}(d)}\Psi(\boldsymbol{\theta}_*)(d) + o_{p, \mathcal{D}}(1)]^{-1} \sqrt{n}\Psi_n(\boldsymbol{\theta}_*)(d), \quad (6)$$

for all $d \in \mathcal{D}$, where $o_{p, \mathcal{D}}(1)$ denotes uniform convergence to zero for all $d \in \mathcal{D}$. It is easy to prove the asymptotic normality of $\sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)]$ for each $d \in \mathcal{D}$. Furthermore, by using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that $\sup_{d \in \mathcal{D}} \|\Psi_n(\boldsymbol{\theta}_*)(d)\|_2 = O_p(\sqrt{\log(N(\mathcal{D}))/n})$, which yields

$$\max_{d \in \mathcal{D}} \|\hat{\boldsymbol{\theta}}(d, h_0) - \boldsymbol{\theta}_*(d)\|_2 = O_p(\sqrt{\log N(\mathcal{D})/n}). \quad (7)$$

In Step 3, we will derive the rate of convergence of $D_{\boldsymbol{\theta}}(d, d'; h_0)$. Since $D_{\boldsymbol{\theta}}(d, d'; h_0)$ can be rewritten as

$$n[\hat{\Delta}(d, 0) - \hat{\Delta}(d', 0) + \Delta_*(d, d')]^T \Sigma_*(d, h)^{-1} [\hat{\Delta}(d, 0) - \hat{\Delta}(d', 0) + \Delta_*(d, d')] [1 + o_p(1)],$$

it follows from (7) that if $\Delta_*(d, d') = \mathbf{0}$, then $\max_{d, d' \in \mathcal{D}} |D_{\boldsymbol{\theta}}(d, d'; h_0)| = O_p(\log(N(\mathcal{D})))$ and $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_0)/C_n) = K_{st}(O_p(\log(N(\mathcal{D}))/C_n)) = 1 + o_p(1)$. However, if $\Delta_*(d, d') \neq \mathbf{0}$, then we have

$$D_{\boldsymbol{\theta}}(d, d'; h_0) = n \|\Sigma_*(d, h)^{-1/2} [\Delta_*(d, d') + O_p(\sqrt{\log N(\mathcal{D})/n})]\|_2^2,$$

which yields the proof of Theorem 1.

Proof of Theorem 2. We prove Theorem 2 (a) and (b) by induction. The proof primarily consists of three steps:

(i) $s = 0$; (ii) $s = 1$; (iii) $s \geq 1$. In Step 1, we have already proved the case $s = 0$ in Theorem 1.

We prove Step 2 as follows. It follows from the definition of $\tilde{\omega}(d, d'; h_1)$ that

$$\sup_{\boldsymbol{\theta}(d) \in \mathcal{B}} |n^{-1} \ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\omega}) - M(\boldsymbol{\theta}(d); h_1, \tilde{\omega})| \leq \sum_{d' \in B(d, h_1)} \tilde{\omega}(d, d'; h_1) \delta_n(d') \leq \max_{d' \in B(d, h_1)} \delta_n(d'),$$

where $\delta_n(d) = \sup_{\boldsymbol{\theta}(d) \in \mathcal{B}} |n^{-1} \sum_{i=1}^n \log p(Y_i(d)|\mathbf{x}_i, \boldsymbol{\theta}(d)) - E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]|$. Then, following arguments in Theorems 2.7.11 and 2.4.3 of van der Vaart and Wellner (1996) and assumptions (C2)-(C4), we can show that

$$E[\max_{d \in \mathcal{D}} \delta_n(d)] \leq \sqrt{[1 + p \log(C_1(\epsilon)K) + \log(N(\mathcal{D}))]/n} C_2 K + 2E[\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x}) \mathbf{1}\{\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x}) > K\}] + \epsilon \rightarrow 0.$$

Since the above arguments are independent of $\tilde{\omega}(d, d'; h_1)$, we can conclude that

$$\max_{d \in \mathcal{D}} \sup_{\boldsymbol{\theta}(d) \in \mathcal{B}} |n^{-1} \ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\omega}) - M(\boldsymbol{\theta}(d); h_1, \tilde{\omega})| \rightarrow 0 \quad (8)$$

in probability, and it holds for any adaptive weights $\tilde{\omega}(d, d'; h)$.

Let $\mathcal{D}_*(d)^c = \{d' : \Delta_*(d, d') \neq \mathbf{0}\}$ and $\mathcal{D}_*(d) = \{d' : \Delta_*(d, d') = \mathbf{0}\}$. According to Theorem 1 (c), for all $d' \in B(d, h_1) \cap \mathcal{D}_*(d)^c$ and any $d \in \mathcal{D}$, we have

$$\begin{aligned} C_n^{-1} D_{\boldsymbol{\theta}}(d, d'; h_0) &= n C_n^{-1} \lambda_{\max}(\Sigma_*(d, h_0))^{-1} \times \\ &\quad \inf_{d' \in \mathcal{D}_*(d)^c} \|\Delta_*(d, d') + O_p(n^{-1/2})\|_2^2 = \tilde{\delta}_n(d) \rightarrow \infty. \end{aligned} \quad (9)$$

It follows from (9) and (7) that

$$\begin{aligned} &\max_{d \in \mathcal{D}} \sup_{\boldsymbol{\theta}(d)} \left| M(\boldsymbol{\theta}(d); h_1, \tilde{\omega}) - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))] \right| \\ &\leq \max_{d \in \mathcal{D}} K_{st}(\tilde{\delta}_n(d)) E[\max_{d \in \mathcal{D}} G(Y(d), \mathbf{x})] \rightarrow 0. \end{aligned} \quad (10)$$

Since $\boldsymbol{\theta}_*(d) = \operatorname{argmax}_{\boldsymbol{\theta}(d)} \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h) E[\log p(Y(d)|\mathbf{x}, \boldsymbol{\theta}(d))]$, it follows from Theorem 5.7 of van der Vaart (1998) and the arguments in the proof of Theorem 1 (a) that $\hat{\boldsymbol{\theta}}(h_1) = (\hat{\boldsymbol{\theta}}(d, h_1) : d \in \mathcal{D})$ converges to $\boldsymbol{\theta}_*$ in probability.

To prove asymptotic normality of $\hat{\boldsymbol{\theta}}(d, h_1)$, we can use a Taylor's series expansion to show that

$$\mathbf{0} = \partial_{\boldsymbol{\theta}(d)} \ell_n(\hat{\boldsymbol{\theta}}(d, h_1); h_1, \tilde{\omega}) = \partial_{\boldsymbol{\theta}(d)} \ell_n(\boldsymbol{\theta}_*(d); h_1, \tilde{\omega}) + \partial_{\boldsymbol{\theta}(d)}^2 \ell_n(\tilde{\boldsymbol{\theta}}(d, h_1); h_1, \tilde{\omega}) [\hat{\boldsymbol{\theta}}(d, h_1) - \boldsymbol{\theta}_*(d)],$$

where $\tilde{\boldsymbol{\theta}}(d, h_1)$ is on the segment joining $\hat{\boldsymbol{\theta}}(d, h_1)$ and $\boldsymbol{\theta}_*(d)$. Similar to the arguments in the proof of Theorem 1 (b) and (10), we can show that

$$\begin{aligned} &\max_{d \in \mathcal{D}} \sup_{\|\boldsymbol{\theta}_*(d) - \boldsymbol{\theta}(d)\|_2 \leq \epsilon} |n^{-1} \partial_{\boldsymbol{\theta}(d)}^2 \ell_n(\boldsymbol{\theta}(d); h_1, \tilde{\omega}) - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) E[\partial_{\boldsymbol{\theta}(d)}^2 \log p(Y(d')|\mathbf{x}, \boldsymbol{\theta}(d))]| \rightarrow 0, \\ &\max_{d \in \mathcal{D}} n^{-1/2} |\partial_{\boldsymbol{\theta}(d)} \ell_n(\boldsymbol{\theta}_*(d); h_1, \tilde{\omega}) - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d')|\mathbf{x}, \boldsymbol{\theta}_*(d))| \\ &\leq n^{1/2} K_{st}(O_p(n C_n^{-1})) E[\sup_{d \in \mathcal{D}} G(Y(d), \mathbf{x})] O(1) \rightarrow 0. \end{aligned}$$

Finally, we obtain

$$\begin{aligned} \sqrt{n}[\hat{\boldsymbol{\theta}}(d, h_1) - \boldsymbol{\theta}_*(d)] &= \left\{ - \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) E[\partial_{\boldsymbol{\theta}(d)}^2 \log p(Y(d') | \mathbf{x}, \boldsymbol{\theta}_*(d))] + o_{p, \mathcal{D}}(1) \right\}^{-1} \times \\ &\quad n^{-1/2} \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}_*(d)). \end{aligned} \quad (11)$$

By using Theorem 2.14.1 of van der Vaart and Wellner (1996), we can show that

$$\max_{d \in \mathcal{D}} \|n^{-1/2} \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d) | \mathbf{x}, \boldsymbol{\theta}_*(d))\|_2 = O_p(\sqrt{\log N(\mathcal{D})}),$$

which yields that $\max_{d \in \mathcal{D}} \|\hat{\boldsymbol{\theta}}(d, h) - \boldsymbol{\theta}_*(d)\|_2 = O_p(\sqrt{\log N(\mathcal{D})/n})$. Based on these results for $\hat{\boldsymbol{\theta}}(d, h_1)$, we can prove the same results as Theorem 1 (c) and (d) for $D_{\boldsymbol{\theta}}(d, d'; h_1)$ and $K_{st}(D_{\boldsymbol{\theta}}(d, d'; h_1)C_n^{-1})$.

In Step 3, by using induction and the above arguments in Step 2, we can prove Theorem 2 (a) and (b) for any fixed $s > 1$.

Given the results in Theorem 2 (a) and (b), we can apply standard arguments in the literature to prove Theorem 2 (c). We omit the details for simplicity.

Proof of Corollary 1. Because we can prove Corollary 1 (a) using the same arguments in proving Theorem 2 (a), we omit the details.

The proof of Corollary 1 (b) consists of two steps. In Step 1, following the same arguments in Theorem 2 (a), we can prove (11). In Step 2, we examine the asymptotic distribution of

$$A(d; h_1) = \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \omega(d, d'; h) n^{-1/2} \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}(d)_*).$$

For any $d' \in B(d, h_1) \cap \mathcal{D}_*(d)$, $D_{\boldsymbol{\theta}}(d, d'; h_0)$ converges to a random variable, denoted by $Z(d, d'; h_0)$, in distribution, and thus $\omega(d, d'; h)$ converges to $K_{st}(Z(d, d'; h_0))$ in distribution. In addition, for any $d' \in B(d, h_1) \cap \mathcal{D}_*(d)$, $n^{-1/2} \sum_{i=1}^n \partial_{\boldsymbol{\theta}(d)} \log p(Y_i(d') | \mathbf{x}, \boldsymbol{\theta}(d)_*)$ converges to a normal random vector, denoted by $Z(d')$, in distribution. Note that $Z(d')$ and $Z(d, d'; h_0)$ are correlated with each other. Finally, using the continuous mapping theorem, we can claim that $A(d; h_1)$ converges to

$$\sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} K_{loc}(\|d - d'\|_2/h_1) K_{st}(Z(d, d'; h_0)) Z(d),$$

which is not a normal random variable when there is a $d' \in B(d, h_1) \cap \mathcal{D}_*(d)$. Thus, $W_{\mu}(d, h_1)$ is not asymptotically χ^2 distributed.

Proof of Theorem 3. We prove Theorem 3 (a) using induction. The proof primarily consists of two steps: (i) $\sqrt{n}[\hat{\beta}(d, h_0) - \beta_*(d)] = A_1(d; h_0) + o_p(1)$; (ii) $\sqrt{n}[\hat{\beta}(d, h_1) - \beta_*(d)] = A_1(d; h_1) + o_p(1)$ for each voxel d . Moreover, for notational simplicity, we assume that $\tau(d)$ is known through the proof.

In Step 1, since $\hat{\beta}(d, h_0) = (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i Y_i(d) = \beta_*(d) + A_1(d; h_0)/\sqrt{n} = \beta_*(d) + (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i \epsilon_i(d)$ holds, it is easy to show that $A_1(d; h_0)$ converges to $E[\mathbf{x}^{\otimes 2}]^{-1/2} Z(d)$ in distribution for any voxel d . Following the arguments in Theorem 2.4.3, we can show that $\max_{d \in \mathcal{D}} \|n^{-1} \sum_{i=1}^n \mathbf{x}_i \epsilon_i(d)\|_2 = O_p(\sqrt{\log(N(\mathcal{D}))/n})$.

In Step 2, since $D_{\beta}(d, d'; h_0)$ can be rewritten as

$$n\tau(d) \|E[\mathbf{x}^{\otimes 2}]^{-1/2} \{\Delta_*(d, d') + (\sum_{i=1}^n \mathbf{x}_i^{\otimes 2})^{-1} \sum_{i=1}^n \mathbf{x}_i [\epsilon_i(d') - \epsilon_i(d)]\}\|_2^2,$$

where $\Delta_*(d, d') = \beta_*(d) - \beta_*(d')$, we can check that $D_{\beta}(d, d'; h_0)$ and $K_{st}(D_{\beta}(d, d'; h_0)/C_n)$ have the asymptotic expansions as described in Lemma 1. We can show that $\tilde{\omega}(d, d'; h_1)$ are smaller than $K_{st}(O_p(nC_n^{-1}))$ for all $d' \in B(d, h_1) \cap \mathcal{D}_*(d)^c$ and $\hat{\omega}(d, d'; h_1)$ converges to $C(d, d'; h_1)$ for all $d' \in B(d, h_1) \cap \mathcal{D}_*(d)$. Therefore, we have

$$\begin{aligned} \sqrt{n}[\hat{\beta}(d, h_1) - \beta_*(d)] &= \sum_{d' \in B(d, h_1) \cap \mathcal{D}_*(d)} \tilde{\omega}(d, d'; h_1) E[\mathbf{x}^{\otimes 2}]^{-1/2} n^{-1/2} \sum_{i=1}^n \mathbf{x}_i \epsilon_i(d') + o_p(1) \\ &= A_1(d, h) + o_p(1). \end{aligned}$$

Applying the continuous mapping theorem yields weak convergence of $A_1(d, h_1)$ and $\sqrt{n}[\hat{\beta}(d, h_1) - \beta_*(d)]$. We can use the same arguments in Corollary 1 (b) to prove Theorem 3 (b). Note that for the PS approach, $\hat{\omega}(d, d'; h_1)$ converges in distribution to $C(d, d'; h_1) K_{st}(\tau(d) \|Z(d) - Z(d')\|_2^2)$ for all $d' \in B(d, h_1) \cap \mathcal{D}_*(d)$.

References

- Chung, M.K., Robbins, S., Dalton, K.M., Davidson, R.J., Alexander, A.L., and Evans, A.C. (2005) Cortical thickness analysis in autism via heat kernel smoothing. *NeuroImage*, **25**, 1256-1265.
- Li, Y. M., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2010) Multiscale adaptive regression models for neuroimaging data.
- van der Vaart (1998) *Asymptotic Statistics*. Cambridge University Press.
- van der Vaart and Wellner, J. (1996) *Weak Convergence and Empirical Processes*. New York: Springer.

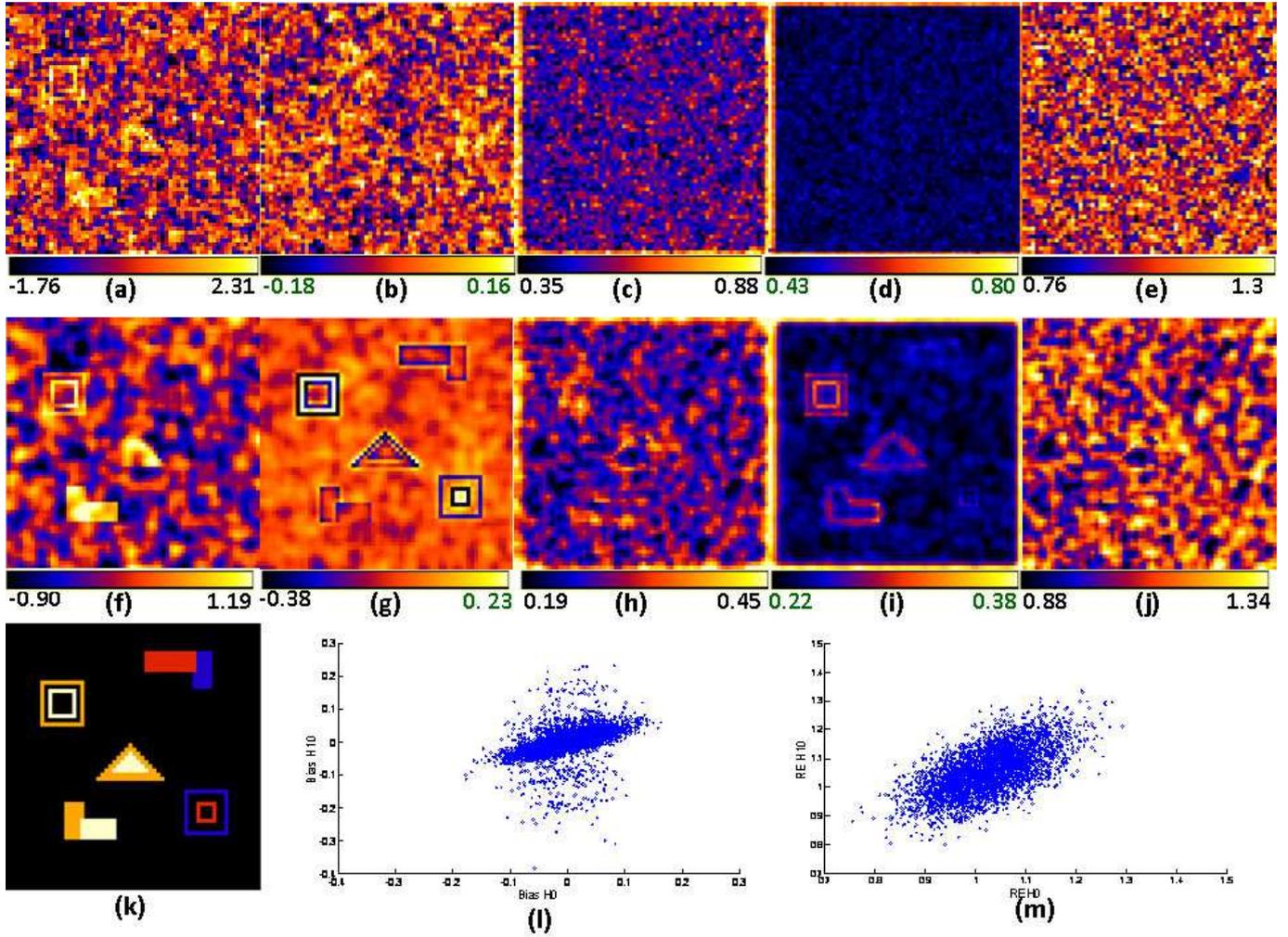


Fig. 2. Results from a simulation study of comparing voxel-wise method and MARM based on 1,000 $\chi^2(3) - 3$ distributed data with $n = 60$. Panel (k) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)=0, 0.2, 0.4, 0.6,$ and 0.8 , respectively. The first row contains the results from voxel-wise method: (a) a selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS image of $\hat{\beta}_2(d, h_0)$; (d) SD image of $\hat{\beta}_2(d, h_0)$; and (e) RE image of $\hat{\beta}_2(d, h_0)$. The second row contains the results obtained from MAET as $S = 10$ and $c_h = 1.1$: (f) a selected image of $\hat{\beta}_2(d, h_{10})$ obtained from a simulated data set; (g) bias image of $\hat{\beta}_2(d, h_{10})$; (h) RMS image of $\hat{\beta}_2(d, h_{10})$; (i) SD image of $\hat{\beta}_2(d, h_{10})$; and (j) RE image of $\hat{\beta}_2(d, h_{10})$. Panels (l) and (m) are the scatter plots of biases and REs of $\hat{\beta}_2(d, h_0)$ versus $\hat{\beta}_2(d, h_{10})$, respectively.

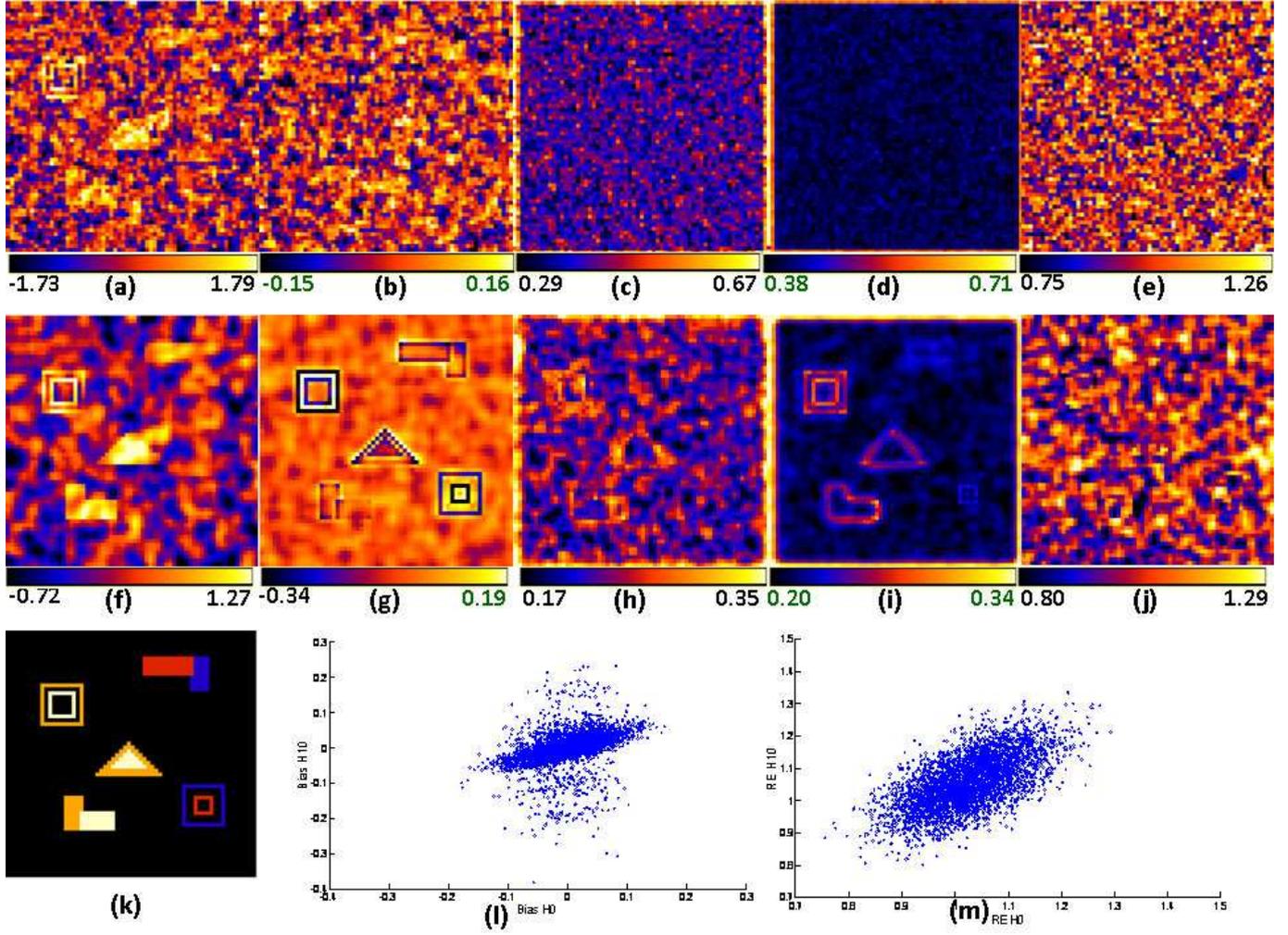


Fig. 3. Results from a simulation study of comparing voxel-wise method and MARM based on 1,000 $\chi^2(3) - 3$ distributed data with $n = 80$. Panel (k) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)=0, 0.2, 0.4, 0.6,$ and 0.8 , respectively. The first row contains the results from voxel-wise method: (a) a selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS image of $\hat{\beta}_2(d, h_0)$; (d) SD image of $\hat{\beta}_2(d, h_0)$; and (e) RE image of $\hat{\beta}_2(d, h_0)$. The second row contains the results obtained from MAET as $S = 10$ and $c_h = 1.1$: (f) a selected image of $\hat{\beta}_2(d, h_{10})$ obtained from a simulated data set; (g) bias image of $\hat{\beta}_2(d, h_{10})$; (h) RMS image of $\hat{\beta}_2(d, h_{10})$; (i) SD image of $\hat{\beta}_2(d, h_{10})$; and (j) RE image of $\hat{\beta}_2(d, h_{10})$. Panels (l) and (m) are the scatter plots of biases and REs of $\hat{\beta}_2(d, h_0)$ versus $\hat{\beta}_2(d, h_{10})$, respectively.

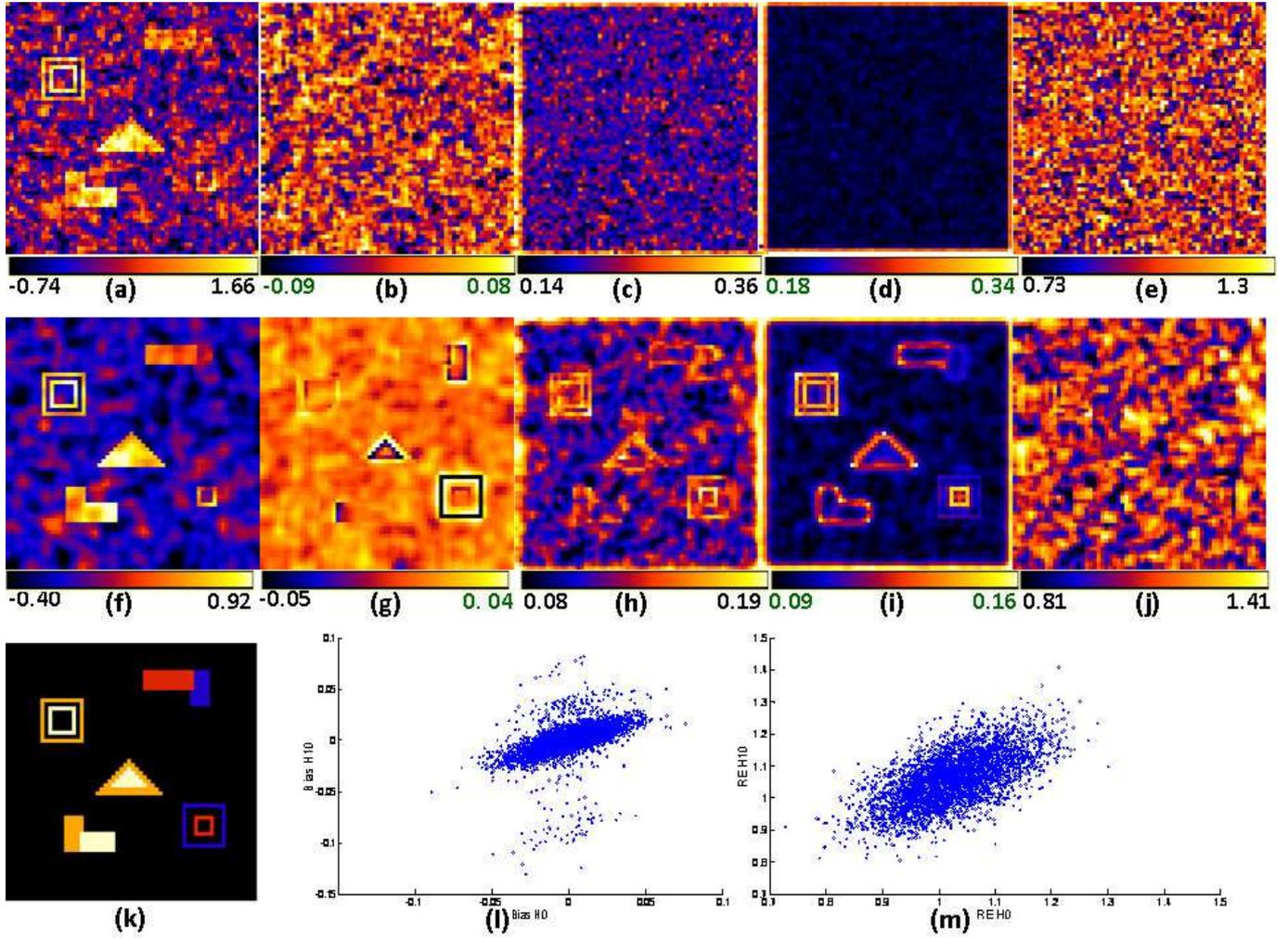


Fig. 4. Results from a simulation study of comparing voxel-wise method and MARM based on 1,000 $N(0, 1)$ distributed data with $n = 60$. Panel (k) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)=0, 0.2, 0.4, 0.6,$ and 0.8 , respectively. The first row contains the results from voxel-wise method: (a) a selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS image of $\hat{\beta}_2(d, h_0)$; (d) SD image of $\hat{\beta}_2(d, h_0)$; and (e) RE image of $\hat{\beta}_2(d, h_0)$. The second row contains the results obtained from MAET as $S = 10$ and $c_h = 1.1$: (f) a selected image of $\hat{\beta}_2(d, h_{10})$ obtained from a simulated data set; (g) bias image of $\hat{\beta}_2(d, h_{10})$; (h) RMS image of $\hat{\beta}_2(d, h_{10})$; (i) SD image of $\hat{\beta}_2(d, h_{10})$; and (j) RE image of $\hat{\beta}_2(d, h_{10})$. Panels (l) and (m) are the scatter plots of biases and REs of $\hat{\beta}_2(d, h_0)$ versus $\hat{\beta}_2(d, h_{10})$, respectively.

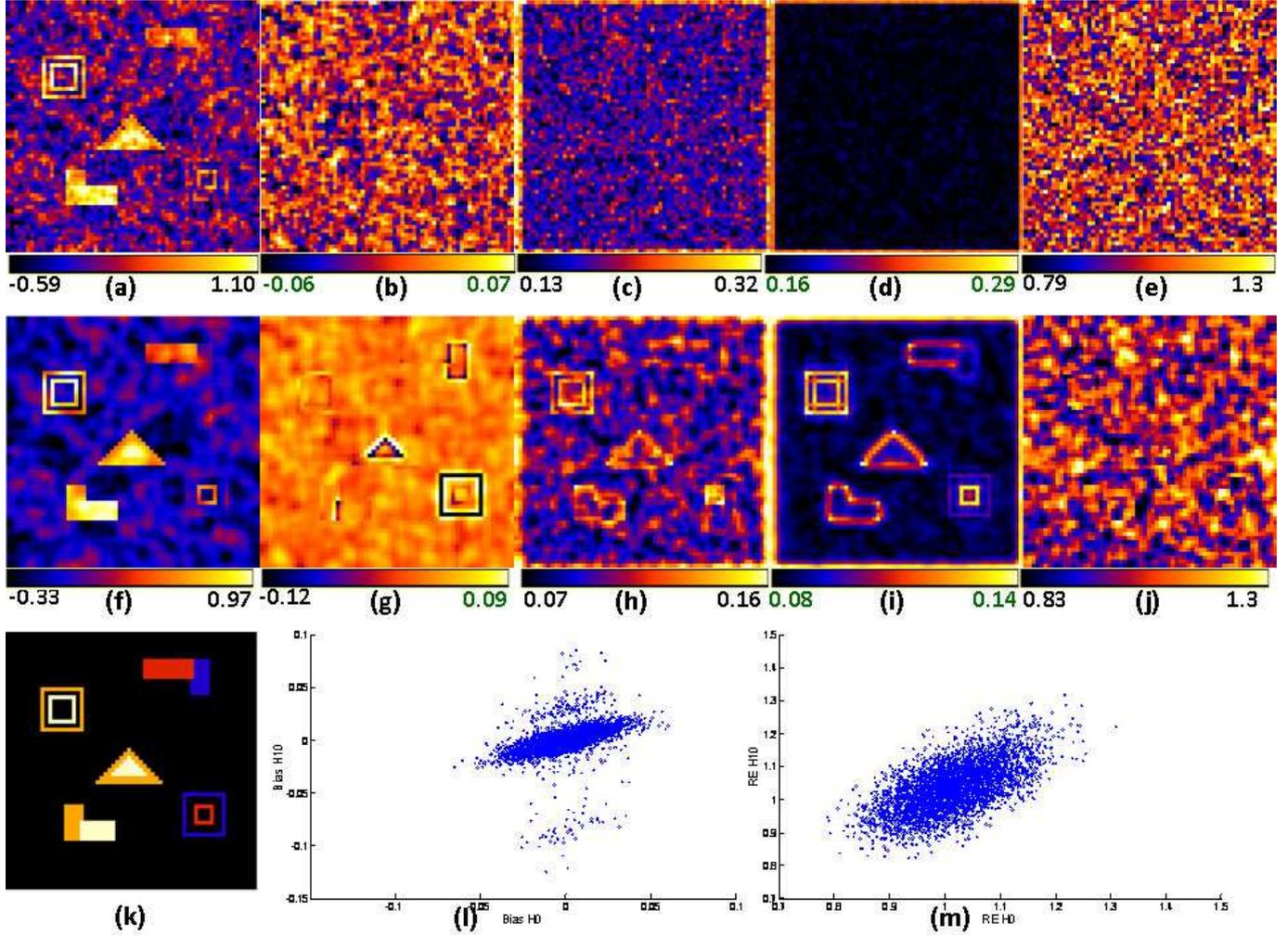


Fig. 5. Results from a simulation study of comparing voxel-wise method and MARM based on 1,000 $N(0, 1)$ distributed data with $n = 80$. Panel (k) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)=0, 0.2, 0.4, 0.6, \text{ and } 0.8$, respectively. The first row contains the results from voxel-wise method: (a) a selected image of $\hat{\beta}_2(d, h_0)$ obtained from a simulated data set; (b) bias image of $\hat{\beta}_2(d, h_0)$; (c) RMS image of $\hat{\beta}_2(d, h_0)$; (d) SD image of $\hat{\beta}_2(d, h_0)$; and (e) RE image of $\hat{\beta}_2(d, h_0)$. The second row contains the results obtained from MAET as $S = 10$ and $c_h = 1.1$: (f) a selected image of $\hat{\beta}_2(d, h_{10})$ obtained from a simulated data set; (g) bias image of $\hat{\beta}_2(d, h_{10})$; (h) RMS image of $\hat{\beta}_2(d, h_{10})$; (i) SD image of $\hat{\beta}_2(d, h_{10})$; and (j) RE image of $\hat{\beta}_2(d, h_{10})$. Panels (l) and (m) are the scatter plots of biases and REs of $\hat{\beta}_2(d, h_0)$ versus $\hat{\beta}_2(d, h_{10})$, respectively.

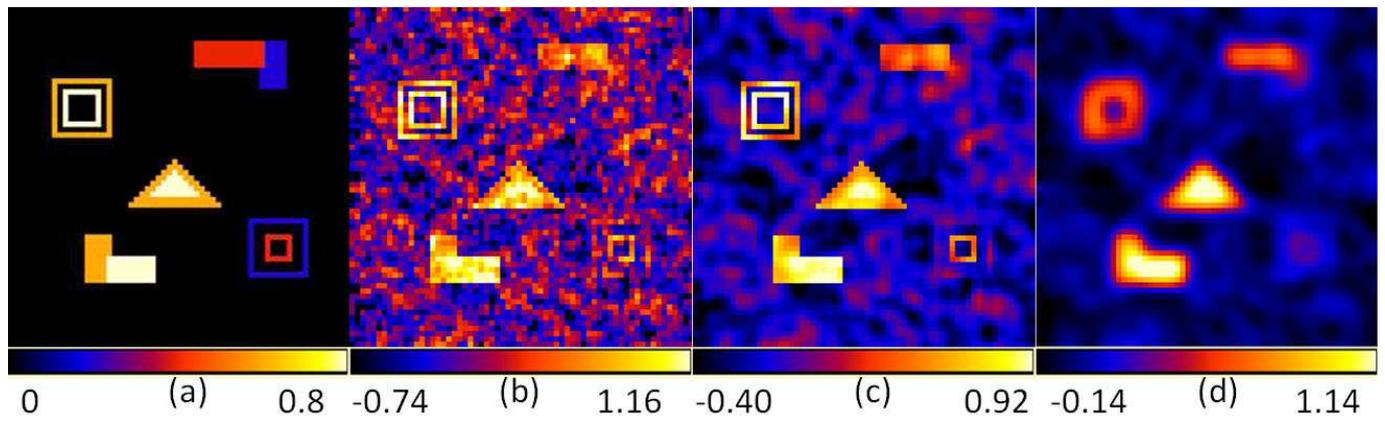


Fig. 6. Selected results for the comparison of the conventional analyses and MARM obtained from a simulated phantom data set with $N(0, 1)$ noises and $n = 60$. Panel (a) is the ground truth image of five ROIs with black, blue, red, yellow, and white color representing $\beta_2(d)=0, 0.2, 0.4, 0.6,$ and $0.8,$ respectively. Results from (b) the voxel-wise method without the initial smoothing step; (c) MARM with $S = 10$; and (d) the conventional analyses.