

Web-based Supplementary Materials for “Haplotype-Based Regression Analysis of Case-Control Studies with Unphased Genotypes and Measurement Errors in Environmental Exposures”

by Iryna Lobach and Raymond J. Carroll
Department of Statistics, Texas A&M University, College Station, TX 77843-3143
iryana@stat.tamu.edu and carroll@stat.tamu.edu

Christine Spinka
Department of Statistics, University of Missouri, Columbia, MO 65211-6100
spinkac@missouri.edu

Mitchell H. Gail and Nilanjan Chatterjee
Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer
Institute, 6120 Executive Blvd, EPS 8038 Rockville MD 20852
gailm@exchange.nih.gov and chattern@mail.nih.gov

Web Appendix A Simulation Results for the Case When Distribution of Environmental Co- variate is Misspecified

Web Table 1: Biases and root mean squared errors for the naive approach that ignores existence of measurement error and our proposed method. Results illustrates sensitivity of the proposed methodology to misspecification of the distribution of environmental covariates. Here we calculated our method under the assumption that the environmental covariate had a normal distribution, while we simulated the environmental covariate from a t -distribution with 10 degrees of freedom. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls, where disease status (D) is binary, environmental variables (X, W) are continuous and the genetic variant h_3 is in the form of diplotype with a multiplicative interaction. The environmental variable is measured with error and the error variance is 0.25, while variance of the environmental covariate is 0.1.

Parameter	True Value	Naive Approach		Proposed Method	
		Bias	RMSE	Bias	RMSE
β_0	-5.000	1.205	1.206	0.230	0.300
β_g	0.693	0.078	0.099	-0.015	0.112
β_x	1.099	-0.794	0.800	-0.036	0.440
β_{xg}	0.693	-0.459	0.467	0.039	0.366
$\text{pr}(h_3)$	0.250	0.005	0.008	0.001	0.011
$\text{pr}(D = 1)$	0.046	-0.032	<0.001	0.008	<0.001
η_1	0.000			0.002	0.027
η_2	0.100			0.005	0.019

Web Appendix B Simulation Results for the Case When Measurement Error Distribution is Estimated

Web Table 2: Biases and root mean squared errors for the naive approach that ignores existence of measurement error and our proposed method. These results illustrate performance of the proposed methodology in the case when measurement error is estimated by replicating 50 randomly selected individuals. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls, where disease status (D) is binary, environmental variables (X, W) are continuous and the genetic variant h_3 is in the form of diplotype with a multiplicative interaction. The environmental variable is measured with error and the error variance is 0.25, while variance of the environmental covariate is 0.1.

Parameter	True Value	Naive Approach		Proposed Method	
		Bias	RMSE	Bias	RMSE
β_0	-5.000	1.192	1.422	0.210	0.229
β_g	0.693	0.055	0.086	-0.007	0.107
β_x	1.099	-0.576	0.408	-0.002	0.688
β_{xg}	0.693	-0.283	0.130	0.040	0.523
$\text{pr}(h_3)$	0.250	0.005	0.008	0.001	0.007
$\text{pr}(D = 1)$	0.046	-0.032	<0.001	0.008	<0.001
η_1	0.000			0.002	0.022
η_2	0.100			0.021	0.067

Web Appendix C Simulation Results for the Case When Environmental Covariate Z Measured Exactly is Present in the Model

Web Table 3: Biases and root mean squared errors for the naive approach that ignores existence of measurement error and our proposed method. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls, where disease status (D) is binary, environmental variables (X, W) are continuous and the genetic variant h_3 is in the form of diplotype with a multiplicative interaction. The error-prone covariate X is simulated from a Normal distribution with zero and variance 0.10. The environmental variable is measured with error and the error variance is 0.25.

Parameter	True Value	Naive Approach		Proposed Method	
		Bias	RMSE	Bias	RMSE
β_0	-5.000	0.026	0.053	-0.444	0.447
β_g	0.693	0.088	0.108	-0.004	0.080
β_x	1.099	-0.783	0.789	0.041	0.301
β_z	0.916	0.296	0.328	0.065	0.226
β_{xg}	0.693	-0.491	0.497	-0.017	0.261
$\text{pr}(h_3)$	0.250	0.000	0.007	0.000	0.007
a_1	0.250			0.003	0.042
σ_x^2	0.100			-0.001	0.101

Web Appendix D Risk Parameter Estimates for Values of Probability of Disease on a Grid

Web Table 4: Risk parameter estimates of the Calcium Data for various values of probability of disease on interval (0.0001,0.0501). The analysis is performed via the naive approach that ignores existence of the measurement error.

π_d	β_x	β_{h2}	β_{h4}	β_{h5}	β_{xh2}	β_{xh4}	β_{xh5}
0.0001	-0.0852	-0.2088	-0.1663	-0.2769	0.0398	-0.1885	-0.2804
0.0051	-0.0850	-0.2087	-0.1665	-0.2771	0.0402	-0.1880	-0.2806
0.0101	-0.0849	-0.2086	-0.1666	-0.2772	0.0405	-0.1874	-0.2807
0.0151	-0.0848	-0.2085	-0.1668	-0.2773	0.0408	-0.1868	-0.2808
0.0201	-0.0847	-0.2084	-0.1669	-0.2775	0.0411	-0.1862	-0.2809
0.0251	-0.0847	-0.2083	-0.1671	-0.2776	0.0414	-0.1855	-0.2809
0.0301	-0.0846	-0.2083	-0.1672	-0.2778	0.0418	-0.1849	-0.2809
0.0351	-0.0846	-0.2082	-0.1674	-0.2779	0.0421	-0.1842	-0.2809
0.0401	-0.0846	-0.2081	-0.1675	-0.2780	0.0425	-0.1835	-0.2808
0.0451	-0.0845	-0.2080	-0.1677	-0.2782	0.0428	-0.1828	-0.2807
0.0501	-0.0845	-0.2080	-0.1678	-0.2783	0.0431	-0.1820	-0.2806

Web Table 5: Risk parameter estimates of the Calcium Data for various values of probability of disease on interval (0.0001,0.0501). Measurement error variance is 0.10.

π_d	β_x	β_{h2}	β_{h4}	β_{h5}	β_{xh2}	β_{xh4}	β_{xh5}	η_1	η_2
0.0001	-0.0678	-0.1988	-0.2188	-0.3617	0.0334	-0.1645	-0.2503	-0.1053	1.5907
0.0051	-0.0677	-0.1987	-0.2187	-0.3616	0.0337	-0.1638	-0.2499	-0.1059	1.5912
0.0101	-0.0676	-0.1985	-0.2186	-0.3615	0.0339	-0.1630	-0.2494	-0.1064	1.5916
0.0151	-0.0675	-0.1984	-0.2184	-0.3614	0.0342	-0.1622	-0.2490	-0.1070	1.5921
0.0201	-0.0675	-0.1982	-0.2183	-0.3613	0.0344	-0.1613	-0.2485	-0.1075	1.5925
0.0251	-0.0674	-0.1981	-0.2182	-0.3611	0.0347	-0.1605	-0.2480	-0.1081	1.5929
0.0301	-0.0674	-0.1979	-0.2180	-0.3610	0.0349	-0.1597	-0.2476	-0.1086	1.5933
0.0351	-0.0673	-0.1978	-0.2177	-0.3609	0.0352	-0.1588	-0.2471	-0.1091	1.5936
0.0401	-0.0673	-0.1976	-0.2177	-0.3608	0.0354	-0.1580	-0.2466	-0.1096	1.5940
0.0451	-0.0673	-0.1974	-0.2175	-0.3606	0.0357	-0.1571	-0.2461	-0.1101	1.5944

Web Table 6: Risk parameter estimates of the Calcium Data for various values of probability of disease on interval (0.0001,0.0501). Measurement error variance is 0.60.

π_d	β_x	β_{h2}	β_{h4}	β_{h5}	β_{xh2}	β_{xh4}	β_{xh5}	η_1	η_2
0.0001	-0.1365	-0.1820	-0.3736	-0.6523	0.0878	-0.4753	-0.7259	-0.1045	0.6794
0.0051	-0.1360	-0.1818	-0.3701	-0.6471	0.0882	-0.4690	-0.7201	-0.1053	0.6808
0.0101	-0.1357	-0.1816	-0.3670	-0.6424	0.0887	-0.4634	-0.7148	-0.1060	0.6821
0.0151	-0.1355	-0.1814	-0.3642	-0.6382	0.0892	-0.4584	-0.7100	-0.1067	0.6832
0.0201	-0.1353	-0.1812	-0.3616	-0.6343	0.0898	-0.4538	-0.7056	-0.1073	0.6843
0.0251	-0.1353	-0.1811	-0.3593	-0.6307	0.0903	-0.4495	-0.7015	-0.1080	0.6853
0.0301	-0.1352	-0.1809	-0.3570	-0.6273	0.0908	-0.4454	-0.6977	-0.1086	0.6862
0.0351	-0.1353	-0.1807	-0.3548	-0.6241	0.0914	-0.4415	-0.6942	-0.1091	0.6870
0.0401	-0.1354	-0.1805	-0.3527	-0.6211	0.0920	-0.4377	-0.6908	-0.1097	0.6878
0.0451	-0.1355	-0.1803	-0.3508	-0.6182	0.0926	-0.4342	-0.6877	-0.1102	0.6885
0.0501	-0.1357	-0.1800	-0.3488	-0.6154	0.0932	-0.4306	-0.6846	-0.1108	0.6892

Web Table 7: Risk parameter estimates of the Calcium Data for various values of probability of disease on interval (0.0001,0.0501). Measurement error variance is 0.65.

π_d	β_x	β_{h2}	β_{h4}	β_{h5}	β_{xh2}	β_{xh4}	β_{xh5}	η_1	η_2
0.0001	-0.1506	-0.1770	-0.4287	-0.7581	0.1043	-0.5811	-0.8876	-0.1045	0.5842
0.0051	-0.1499	-0.1769	-0.4228	-0.7482	0.1048	-0.5710	-0.8766	-0.1054	0.5864
0.0101	-0.1494	-0.1767	-0.4180	-0.7399	0.1053	-0.5629	-0.8673	-0.1062	0.5882
0.0151	-0.1492	-0.1765	-0.4140	-0.7325	0.1060	-0.5555	-0.8592	-0.1069	0.5898
0.0201	-0.1490	-0.1763	-0.4098	-0.7260	0.1066	-0.5489	-0.8520	-0.1076	0.5912
0.0251	-0.1489	-0.1754	-0.4048	-0.7168	0.1087	-0.5410	-0.8419	-0.1086	0.5928
0.0301	-0.1491	-0.1759	-0.4031	-0.7150	0.1079	-0.5375	-0.8400	-0.1089	0.5936
0.0351	-0.1492	-0.1757	-0.4000	-0.7100	0.1086	-0.5321	-0.8346	-0.1095	0.5946
0.0401	-0.1494	-0.1755	-0.3970	-0.7054	0.1093	-0.5271	-0.8296	-0.1100	0.5956
0.0451	-0.1496	-0.1752	-0.3942	-0.7011	0.1100	-0.5224	-0.8241	-0.1106	0.5965
0.0501	-0.1499	-0.1750	-0.3916	-0.6971	0.1107	-0.5179	-0.8208	-0.1111	0.5973

Web Table 8: Risk parameter estimates of the Calcium Data for various values of probability of disease on interval (0.0001,0.0501). Measurement error variance is 0.70.

π_d	β_x	β_{h2}	β_{h4}	β_{h5}	β_{xh2}	β_{xh4}	β_{xh5}	η_1	η_2
0.0001	-0.1692	-0.1701	-0.5146	-0.9220	0.1280	-0.7426	-1.1317	-0.1045	0.4850
0.0051	-0.1672	-0.1700	-0.5037	-0.9012	0.1286	-0.7250	-1.1092	-0.1057	0.4889
0.0101	-0.1698	-0.1698	-0.4955	-0.8850	0.1293	-0.7117	-1.0915	-0.1066	0.4917
0.0151	-0.1659	-0.1696	-0.4885	-0.8718	0.1301	-0.7004	-1.0771	-0.1074	0.4940
0.0201	-0.1658	-0.1694	-0.4825	-0.8608	0.1309	-0.6907	-1.0651	-0.1081	0.4960
0.0251	-0.1658	-0.1692	-0.4771	-0.8511	0.1317	-0.6819	-1.0546	-0.1088	0.4976
0.0301	-0.1660	-0.1689	-0.4721	-0.8424	0.1326	-0.6737	-1.0453	-0.1094	0.4992
0.0351	-0.1663	-0.1686	-0.4674	-0.8347	0.1335	-0.6662	-1.0371	-0.1100	0.5005
0.0451	-0.1672	-0.1681	-0.4288	-0.8205	0.1353	-0.6522	-1.0222	-0.1111	0.5030
0.0501	-0.1677	-0.1678	-0.4549	-0.8144	0.1362	-0.6458	-1.0159	-0.1116	0.5040

Web Appendix E Standard Errors and Confidence Intervals of Risk Parameter Estimates of Calcium Data

Web Table 9: Standard errors of risk parameter estimates for the colorectal adenoma study assuming different variances (ξ) of the measurement error. The estimated measurement error variance is $\xi = 0.65$.

Parameter	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
β_{h_2}	0.106	0.118	0.119	0.120
β_{h_4}	0.130	0.151	0.153	0.155
β_{h_5}	0.135	0.169	0.172	0.175
β_x	0.068	0.158	0.167	0.176
β_{xh_2}	0.084	0.189	0.200	0.211
β_{xh_4}	0.090	0.192	0.203	0.213
β_{xh_5}	0.100	0.203	0.213	0.224

Web Table 10: Bootstrap standard errors of risk parameter estimates for the colorectal adenoma study assuming different variances (ξ) of the measurement error. The estimated measurement error variance is $\xi = 0.65$.

Parameter	$\xi = 0.10$	$\xi = 0.60$	$\xi = 0.65$	$\xi = 0.70$
β_{h_2}	0.078	0.146	0.134	0.216
β_{h_4}	0.112	0.196	0.306	0.136
β_{h_5}	0.131	0.319	0.335	0.272
β_x	0.059	0.392	0.174	0.370
β_{xh_2}	0.117	0.159	0.242	0.299
β_{xh_4}	0.166	0.259	0.252	0.353
β_{xh_5}	0.172	0.358	0.424	0.505

Web Table 11: Lower (LB) and upper (UB) bounds of 95% Wald confidence intervals of risk parameter estimates for the colorectal adenoma study assuming different variances (ξ) of the measurement error. The estimated measurement error variance is $\xi = 0.65$.

Parameter	$\xi = 0.10$		$\xi = 0.60$		$\xi = 0.65$		$\xi = 0.70$	
	LB	UB	LB	UB	LB	UB	LB	UB
β_{h_2}	-0.394	0.021	-0.391	0.070	-0.410	0.056	-0.372	0.099
β_{h_4}	-0.446	0.065	-0.667	-0.075	-0.662	0.196	-0.842	-0.233
β_{h_5}	-0.631	-0.103	-0.991	-0.330	-1.095	-0.422	-1.281	-0.595
β_x	-0.201	0.065	-0.450	0.170	-0.478	0.177	-0.531	0.161
β_{xh_2}	-0.125	0.204	-0.239	0.510	-0.287	0.496	-0.190	0.635
β_{xh_4}	-0.350	0.001	-0.896	-0.142	-0.979	-0.184	-1.230	-0.395
β_{xh_5}	-0.433	-0.039	-0.142	-1.111	-1.306	-0.471	-1.562	-0.685

Web Table 12: 95% Bootstrap confidence intervals of risk parameter estimates for the colorectal adenoma study assuming different variances (ξ) of the measurement error. The estimated measurement error variance is $\xi = 0.65$.

Parameter	$\xi = 0.10$		$\xi = 0.60$		$\xi = 0.65$		$\xi = 0.70$	
	LB	UB	LB	UB	LB	UB	LB	UB
β_{h_2}	-0.371	0.035	-0.423	0.056	-0.413	0.063	-0.438	0.064
β_{h_4}	-0.516	0.105	-0.878	0.068	-1.221	-0.036	-1.083	-0.132
β_{h_5}	-0.751	-0.046	-1.436	-0.128	-1.701	-0.234	-1.851	-0.260
β_x	-0.199	0.056	-0.424	0.178	-0.477	0.221	-0.546	0.280
β_{xh_2}	-0.083	0.182	-0.289	0.674	-0.395	0.789	-0.353	0.828
β_{xh_4}	-0.407	0.035	-0.975	0.061	-1.221	-0.016	-1.368	-0.080
β_{xh_5}	-0.557	-0.011	-1.762	-0.114	-2.114	-0.183	-2.163	-0.211

Web Appendix F Coverage Probabilities of Wald and LR Confidence Intervals in Binary Case

Web Table 13: Coverage probabilities of the 95% Wald and LR confidence intervals for gene-environment interaction parameter. Disease status (D), the genetic variant (G), and the environmental covariate (X) are binary. The environmental variable is measured with error, with misclassification probabilities being 0.20 for exposed and 0.10 for non-exposed subjects. The results are based on a simulation study with 500 replications for 1000 cases and 1000 controls.

	n = 400	n = 2000
True value of β_{xg}	0.693	0.693
Mean of $\hat{\beta}_{xg}$ over all simulated datasets	0.848	0.692
Median of $\hat{\beta}_{xg}$ over all simulated datasets	0.695	0.669
Coverage of the Wald CI	0.937	0.931
Coverage of the Likelihood Ratio CI	0.954	0.949

Web Appendix G Wald and LR-type Confidence Intervals for Colorectal Adenoma Study

Web Table 14: Wald and ARLR confidence intervals for β_{xh4} for the Colorectal Adenoma Study assuming various measurement error variances ξ . The estimated measurement error variance is 0.65. The analysis is performed for Female subjects.

	Wald CI	LR-type CI
$\xi = 0.10$	(-0.416,-0.090)	(-0.460, 0.119)
$\xi = 0.60$	(-0.972, 0.175)	(-1.220, 0.388)
$\xi = 0.65$	(-1.143, 0.198)	(-1.462, 0.420)
$\xi = 0.70$	(-1.394, 0.230)	(-1.912, 0.583)

Web Appendix H Technical Derivations

A.1 Proof of (3)

The proof of (3) is straightforward. Note that

$$\begin{aligned}
& \text{pr}(D = d, H^{\text{dip}} = h^{\text{dip}}, X = x, W = w | R = 1, Z = z) \\
& \propto \text{pr}(D = d, H^{\text{dip}} = h^{\text{dip}}, X = x, W = w, R = 1 | Z = z) \\
& \propto \frac{n_d}{\pi_d} \left[1 + \sum_{j=1}^K \exp\{\beta_{0j} + m(h^{\text{dip}}, x, z, \beta)\} \right]^{-1} \exp[I_{(d \geq 1)}(d) \{\beta_{0d} + m(h^{\text{dip}}, x, z, \beta)\}] \\
& \quad \times Q(h^{\text{dip}}, \Theta) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) \\
& \propto \frac{n_0}{\pi_0} S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) \\
& = \frac{S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta)}{\sum_{d_*} \sum_{h_*^{\text{dip}}} \int S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d_*, h_*^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) dw dx} \\
& = \frac{S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta)}{\sum_{d_*} \sum_{h_*^{\text{dip}}} \int S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x | z, \eta) dx}.
\end{aligned}$$

Equation (3) now follows by appropriate summation over $h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}$ and integration over x .

A.2 Proof of Theorem 1

The proof consists of two steps. The first shows that the estimating equation has mean zero when evaluated at the true parameters. We then show that the estimating function evaluated at the true parameters has a covariance matrix of the form $\mathcal{I} - \Lambda$.

A.2.1 Unbiasedness of the Estimating Function

We first consider the derivative with respect to Ω . Denote the first partial derivative of $S(d, h^{\text{dip}}, x, z, \Omega)$ with respect to Ω by $S_\Omega(d, h^{\text{dip}}, x, z, \Omega)$. The semiparametric profile likelihood score for $\mathcal{B}(\Omega^T, \eta^T)^T$ is the derivative of the logarithm of (3) with respect to Ω and is given as

$$n^{-1} \sum_{i=1}^n \{C_1(D_i, Z_i, W_i, G) - C_2(Z_i)\},$$

where $C_1(\bullet) = \{A_1^T(\bullet), B_1^T(\bullet)\}$, $C_2(\bullet) = \{A_2^T(\bullet), B_2^T(\bullet)\}$,

$$A_1(d, z, w, g) = \frac{\sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \int S_\Omega(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) dx}{\sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \int S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w | d, h^{\text{dip}}, x, z, \xi) f_X(x | z, \eta) dx},$$

$$A_2(z) = \frac{\int \sum_{d_*} \sum_{h_*^{\text{dip}}} S_\Omega(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x | z, \eta) dx}{\int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x | z, \eta) dx},$$

and where $B_1(\bullet)$ and $B_2(\bullet)$ are defined by replacing $S_\Omega(\bullet)$ by $S(\bullet)s_X(x|z, \eta)$, where $s_X(x|z, \eta) = \partial \log\{f_X(x|z, \eta)\}/\partial \eta$.

It is useful to note that the density of Z and (W, G, Z) given $D = d$ can be written as

$$[Z|D = d] = f_Z(z) \frac{n_0}{\pi_0 n_d} \int \sum_{h_*^{\text{dip}}} S(d, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) dx; \quad (\text{A.1})$$

$$\begin{aligned} [W, G, Z|D = d] &= f_Z(z) \frac{n_0}{\pi_0 n_d} \int \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) \\ &\quad \times f_X(x|z, \eta) dx. \end{aligned} \quad (\text{A.2})$$

Then it follows from (A.1) that

$$\begin{aligned} \text{E}\{A_2(Z)\} &= \sum_{d_*} \frac{n_{d_*}}{n} \text{E}\{A_2(Z)|D = d_*\} \\ &= \int \sum_{d_*} \frac{n_0}{n\pi_0} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) A_2(z) dx dz \\ &= \int \sum_{d_*} \frac{n_0}{n\pi_0} \sum_{h_*^{\text{dip}}} S_\Omega(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) dx dz. \end{aligned}$$

It is also straightforward using (A.2) to show that

$$\begin{aligned} \text{E}\{A_1(D, Z, W, G)\} &= \sum_{d_*} \frac{n_{d_*}}{n} \text{E}\{A_1(D, Z, W, G|D = d_*)\} \\ &= \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{h_*^{\text{dip}}} S_\Omega(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) dx dz, \end{aligned}$$

thus showing that the top part of (3) has mean zero. That the bottom part also has mean zero is shown similarly.

Much the same argument holds for the estimating function for ξ . Define $\mathcal{C}(w|d, h^{\text{dip}}, x, z, \xi)$ to be the derivative of $\log\{f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi)\}$ with respect to ξ , and define

$$\begin{aligned} A_\xi(d, w, z, g) &= \frac{\sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \int S(d, h^{\text{dip}}, x, z, \Omega) \mathcal{C}(w|d, h^{\text{dip}}, x, z, \xi) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx}{\sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \int S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx}. \end{aligned}$$

Then it is easy to show that

$$\begin{aligned} \text{E}\{A_\xi(D, W, Z, G)\} &= \sum_{d_*} \frac{n_{d_*}}{n} \text{E}\{A_\xi(D, Z, W, G|D = d_*)\} \\ &= \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) \\ &\quad \times \left\{ \int f_{\text{mem}}(w|d_*, h_*^{\text{dip}}, x, z, \xi) \mathcal{C}(w|d_*, h_*^{\text{dip}}, x, z, \xi) dw \right\} dx dz = 0, \end{aligned}$$

the interior integral being equal to zero by standard likelihood results.

A.2.2 Covariance Matrix of the Estimating Function

As described above, the estimating equation is given as (3). Define $C_3(d) = \{A_3^T(d), B_3^T(d)\}^T = E[\{C_1(D, Z, W, G) - C_2(Z)\} | D = d]$. Then, when evaluated at the true parameters, the estimating function takes the form

$$n^{-1/2} \sum_{i=1}^n \{C_1(D_i, Z_i, W_i, G) - C_2(Z_i) - C_3(D_i)\},$$

which is a sum of independent, mean zero random variables. It follows directly that, when evaluated at the true parameters, the estimating function has covariance matrix

$$\Sigma_* = n^{-1} \sum_{i=1}^n E \left[\{C_1(D_i, Z_i, W_i, G) - C_2(Z_i)\} \{C_1(D_i, Z_i, W_i, G) - C_2(Z_i)\}^T \right] - \Lambda. \quad (\text{A.3})$$

Make the definitions

$$\begin{aligned} \mathcal{Q}_1(d, g, w, z, \mathcal{B}, \xi) &= \int \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \left\{ S_\Omega^T(d, h^{\text{dip}}, x, z, \Omega), S(d, h^{\text{dip}}, x, z, \Omega) s_X^T(x|z, \eta) \right\}^T \\ &\quad \times f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx; \\ \mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi) &= \int \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, \xi) f_X(x|z, \eta) dx; \\ \mathcal{Q}_3(z, \mathcal{B}, \xi) &= \int \sum_{d_*} \sum_{h_*^{\text{dip}}} \left\{ S_\Omega^T(d_*, h_*^{\text{dip}}, x, z, \Omega), S(d_*, h_*^{\text{dip}}, x, z, \Omega) s_X^T(x|z, \eta) \right\}^T \\ &\quad \times f_X(x|z, \eta) dx; \\ \mathcal{Q}_4(z, \mathcal{B}, \xi) &= \int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) dx. \end{aligned}$$

Then it is easy to show that (A.3) can be rewritten as

$$\begin{aligned} \Sigma_* &= \mathcal{A}_1 - \mathcal{A}_2 - \Lambda; \\ \mathcal{A}_1 &= \frac{n_0}{n\pi_0} \int \sum_{d_*} \sum_{g_*} \frac{\mathcal{Q}_1(d_*, g_*, w, z, \mathcal{B}, \xi) \mathcal{Q}_1^T(d_*, g_*, w, z, \mathcal{B}, \xi)}{\mathcal{Q}_2(d_*, g_*, w, z, \mathcal{B}, \xi)} dw f_Z(z) dz; \\ \mathcal{A}_2 &= \frac{n_0}{n\pi_0} \int \frac{\mathcal{Q}_3(z, \mathcal{B}, \xi) \mathcal{Q}_3^T(z, \mathcal{B}, \xi)}{\mathcal{Q}_4(z, \mathcal{B}, \xi)} f_Z(z) dz. \end{aligned}$$

We claim that $\mathcal{I} = \mathcal{A}_1 - \mathcal{A}_2$. By a direct calculation, $\mathcal{I} = \mathcal{I}_1 - \mathcal{I}_2$, where using (A.1),

$$\begin{aligned} \mathcal{I}_2 &= - \sum_{d_*} \frac{n_{d_*}}{n} E \left[\frac{\partial}{\partial \mathcal{B}^T} \left\{ \frac{\mathcal{Q}_3(Z, \mathcal{B}, \xi)}{\mathcal{Q}_4(Z, \mathcal{B}, \xi)} \Big| D = d \right\} \right] \\ &= - \frac{n_0}{n\pi_0} \frac{\partial^2}{\partial \mathcal{B} \partial \mathcal{B}^T} \int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) dx dz + \mathcal{A}_2. \end{aligned}$$

In addition, using (A.2), we find that

$$\begin{aligned} \mathcal{I}_1 &= - \sum_{d_*} \frac{n_{d_*}}{n} E \left[\frac{\partial}{\partial \mathcal{B}^T} \left\{ \frac{\mathcal{Q}_1(d_*, G, W, Z, \mathcal{B}, \xi)}{\mathcal{Q}_2(d_*, G, W, Z, \mathcal{B}, \xi)} \Big| D = d \right\} \right] \\ &= - \frac{n_0}{n\pi_0} \frac{\partial^2}{\partial \mathcal{B} \partial \mathcal{B}^T} \int \sum_{d_*} \sum_{h_*^{\text{dip}}} S(d_*, h_*^{\text{dip}}, x, z, \Omega) f_X(x|z, \eta) f_Z(z) dx dz + \mathcal{A}_1, \end{aligned}$$

completing the proof.

A.3 Proof of Theorem 2

The estimating function for \mathcal{B} can be written in the form

$$0 = \sum_{i=1}^n \sum_{j=1}^M I_{(m_i=j)}(m_i) \mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}),$$

where

$$\mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) = \begin{bmatrix} A_1(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) - A_2(Z_i, \mathcal{B}) - A_3(D_i, \mathcal{B}) \\ A_4(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) \end{bmatrix},$$

$A_2(\bullet)$ and $A_3(\bullet)$ are independent of m and are given in the sections A.2.1 and A.2.2,

$$A_1(d, z, w, g, m, \mathcal{B}) = \mathcal{Q}_1(d, g, w, z, \mathcal{B}, \xi) \{ \mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi) \}^{-1},$$

and

$$A_4(d, z, w, g, m, \mathcal{B}) = \int \sum_{h^{\text{dip}} \in \mathcal{H}_G^{\text{dip}}} \frac{\partial}{\partial \xi^{\text{T}}} \log \{ f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, m, \xi) \} \{ \mathcal{Q}_2(d, g, w, z, \mathcal{B}, \xi) \}^{-1} \\ \times S(d, h^{\text{dip}}, x, z, \Omega) f_{\text{mem}}(w|d, h^{\text{dip}}, x, z, m, \xi) f_X(x|z, \eta) dx,$$

where $\mathcal{Q}_1(\bullet)$ and $\mathcal{Q}_2(\bullet)$ are defined in the section A.2.2. The expectation of the right hand side of (6) is

$$\sum_{j=1}^M p(j) \text{E} \left\{ \sum_{i=1}^n \mathcal{C}(D_i, Z_i, W_i, m_i = j, \mathcal{B}) \right\} = 0,$$

since we have shown that the expectation is zero if the same number of replicates are used. Similarly, $-(\text{Hessian})$ of the right hand side of (6) is

$$- \sum_{i=1}^n \sum_{j=1}^M I_{(m_i=j)}(m_i) \frac{\partial}{\partial \mathcal{B}^{\text{T}}} \mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}),$$

and this has expectation $\sum_{j=1}^M p(j) \mathcal{I}_j = \mathcal{I}$. Finally, the covariance matrix of the right hand side of (6) is

$$\text{E} \left\{ \sum_{i=1}^n \sum_{j=1}^M I_{(m_i=j)}(m_i) \mathcal{C}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) \mathcal{C}^{\text{T}}(D_i, Z_i, W_i, G_i, m_i, \mathcal{B}) \right\} \\ = \sum_{j=1}^M p(j) \Sigma_j = \sum_{j=1}^M p(j) (\mathcal{I}_j - \Lambda_j) = \mathcal{I} - \sum_{j=1}^M p(j) \Lambda_j.$$

This then shows (7).

A.4 Proof of Theorem 3

The proof consists of two steps. First we will show that the limiting distribution of the likelihood ratio test statistic is of the form (13). We then will show that it is distributed as a weighted sum of χ_1^2 random variables.

A.4.1 Step 1

Using a typical likelihood ratio argument, by Taylor expansion we have

$$2 \{ \mathcal{L}(\mathcal{B}_0) - \mathcal{L}(\widehat{\mathcal{B}}) \} = (\mathcal{B}_0 - \widehat{\mathcal{B}})^T \mathcal{L}_{\mathcal{B}\mathcal{B}}(\theta_*) (\mathcal{B}_0 - \widehat{\mathcal{B}}),$$

where θ_* is between θ_0 and $\widehat{\theta}$. Now use (13) and (13), so that

$$\begin{aligned} 2 \{ \mathcal{L}(\mathcal{B}_0) - \mathcal{L}(\widehat{\mathcal{B}}) \} &= \{ n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \}^T \mathcal{I} \{ n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0) \} + o_p(1) \\ &= \mathcal{V}^T \mathcal{I} \mathcal{V} + o_p(1). \end{aligned}$$

A.4.2 Step 2

Since the covariance matrix \mathcal{S}^{-1} is symmetric and positive definite, using Cholesky decomposition it can be factored as $\mathcal{S}^{-1} = LL^T$ where L is a lower-triangular matrix.

Define P to be an orthogonal matrix of eigenvectors of LIL^T and Λ is a diagonal matrix of eigenvalues of LIL^T . Since LIL^T is square and symmetric, Singular Value Decomposition can be applied to it in the following manner: $P^T LIL^T P = \Lambda$. Let $\mathcal{V}_1 = L^{-1}\mathcal{V}$ and $\mathcal{V}_2 = P\mathcal{V}_1$. Note that the distribution of $\mathcal{V}^T \mathcal{I} \mathcal{V}$ is the same as the distribution of $\mathcal{V}_2^T \Lambda \mathcal{V}_2$. It can be easily seen that \mathcal{V}_2 has a limiting Normal(0, I) distribution, where I is an identity matrix.

The fact that the quadratic form $\mathcal{V}_2^T E \mathcal{V}_2$ is distributed as $\sum_{i=1}^k \lambda_i Z_i^2$ completes the proof.

A.5 Proof of Theorem 4

Following ideas the of Roy (1957) it is readily seen that the likelihood ratio takes the following form

$$\begin{aligned} \mathcal{L}_n(\widehat{\mathcal{B}}) - \mathcal{L}_n(\delta_0, \widehat{\gamma}) &= (\widehat{\mathcal{B}} - \mathcal{B})^T \mathcal{L}_{\mathcal{B}\mathcal{B}}(\mathcal{B}_*) (\widehat{\mathcal{B}} - \mathcal{B}) \\ &\quad - (\widehat{\gamma} - \gamma)^T \mathcal{L}_{\mathcal{B}\mathcal{B}}(\delta_0, \gamma_*) (\widehat{\gamma} - \gamma), \end{aligned} \tag{A.4}$$

where \mathcal{B}_* is a point between \mathcal{B} and $\widehat{\mathcal{B}}$, likewise γ_* is a point between γ and $\widehat{\gamma}$. Using arguments of Roy (1975) and Wald (1943), it can be seen that (A.4) for large samples is equivalent to $\{n^{-1/2}(\widehat{\delta} - \delta_0)\}^T \mathcal{J} \{n^{-1/2}(\widehat{\delta} - \delta_0)\}$. Applying arguments used while proving the Theorem 1 we arrive at (14).