Appendix to "Tweedie's Compound Poisson Model With Grouped Elastic Net"

Wei Qian, Yi Yang and Hui Zou
School of Statistics
University of Minnesota

A. Additional Numerical Results of Simulation Example 1

We repeat the procedures in section 4.1 with smaller sample size ($n = 100, 200$ instead of 500), and the results appear to be close to that of sample size 500 in this example. The results are summarized in Tables A1-A3.

B. Misspecification of Link Function

we use the log link in our implementation to ensure that the estimated conditional mean $\hat{\mu}(\mathbf{x})$ is always positive, and at the same time, impose a multiplicative structure. In the context of insurance premium estimation, the log link is much used as it decomposes the expected premium into different risk factors in a multiplicative fashion (when only main effects are considered), and generates convenient model for intuitive interpretation. However, if the true data is not generated by the log link, building models with the log link is subject to estimation bias. Inspired by the Associate Editor's comment, we intend to illustrate this issue by a data simulation setting similar to that of Case 3 in Example 1 except for the link function. Let $\eta := 0.3 + \sum_{j=1}^{6} p_1(T_j)$. Consider the following two scenarios for the true link functions, from which the data is generated: (A) $\mu = \frac{[\exp(\eta)+1]^r - 1}{r}$; (B) $\mu = \exp(\text{sgn}(\eta)|\eta|^r)$, where $r > 0$ and $\text{sgn}(\cdot)$ is the sign function. Clearly, when $r = 1$, both scenarios reduce to the log link. When $r \neq 1$, however, the presumed log-link in our algorithm implementation does not match the true link function, resulting in the misspecification of the link function. To generate data, we choose $r = 0.5, 0.7, 1, 1.3, 1.5$, and assume $\omega = 0$. The corresponding data

Table A1: (Case 1) Averaged simulation results based on 100 runs.

	block-		coefficient-		Gini index
	C	IC	C	IC	
oracle	3	0	9	0	-
$n = 100, \omega = 0$					
lasso	2.87	1.57	5.59	1.94	0.871 (0.018)
grouped lasso	2.93	0.98	8.79	2.94	0.889 (0.015)
grouped elastic net	2.94	1.50	8.82	4.50	0.891 (0.015)
$n = 100, \omega = 0.5$					
lasso	2.81	1.58	5.08	1.89	0.839 (0.024)
grouped lasso	2.93	1.34	8.79	4.02	0.877 (0.011)
grouped elastic net	2.95	1.87	8.85	5.61	0.882 (0.011)
$n = 200, \omega = 0$					
lasso	2.84	1.13	5.55	1.37	0.908 (0.023)
grouped lasso	2.90	0.65	8.70	1.95	0.925 (0.017)
grouped elastic net	2.96	1.13	8.88	3.39	0.948 (0.004)
$n = 200, \omega = 0.5$					
lasso	2.78	1.65	5.46	1.84	0.891 (0.024)
grouped lasso	2.94	1.04	8.82	3.12	0.938 (0.006)
grouped elastic net	2.98	1.42	8.94	4.26	0.941 (0.006)

each has 18 predictor terms, and 6 of them are relevant. We use the lasso method to obtain coefficient-C, coefficient-IC and the mean squared estimation error of η (MSE_η). Based on the results summarized in Table A4 and Figure A1, even though the implemented algorithm may still identify the relevant variables, the estimation of η becomes unreliable when the true link function is very different from the log-link. We leave the further investigation regarding robust estimation for the misspecified link function to the future.

Table A2: (Case 2) Averaged simulation results based on 100 runs.

	block-		coefficient-		Gini index
	C	IC	C	IC	
oracle	3	0	9	0	-
$n = 100, \omega = 0$					
lasso	1.61	0.82	2.13	1.08	0.738 (0.017)
grouped lasso	1.42	0.48	4.26	1.44	0.733 (0.020)
grouped elastic net	2.81	0.72	8.43	2.16	0.746 (0.017)
$n = 100, \omega = 0.5$					
lasso	1.67	0.92	2.28	1.11	0.758 (0.019)
grouped lasso	1.55	0.65	4.65	1.95	0.766 (0.018)
grouped elastic net	2.83	0.92	8.49	2.76	0.766 (0.018)
$n = 200, \omega = 0$					
lasso	1.88	0.30	2.59	0.30	0.819 (0.016)
grouped lasso	1.78	0.18	5.34	0.54	0.825 (0.015)
grouped elastic net	2.97	0.30	8.91	0.90	0.825 (0.015)
$n = 200, \omega = 0.5$					
lasso	1.83	0.70	2.46	0.74	0.817 (0.016)
grouped lasso	1.73	0.43	5.19	1.29	0.823 (0.015)
grouped elastic net	2.89	0.55	8.67	1.65	0.823 (0.015)

C. Partial Fits in Real Data Example

Based on the auto insurance claim data example in section 5, Figure A2 shows the partial fits of the variables selected by the Tweedie model with grouped elastic net (GrpNet).

Table A3: (Case 3) Averaged simulation results based on 100 runs.

	block-		coefficient-		Gini index
	C	IC	C	IC	
oracle	6	0	6	0	-
$n = 100, \omega = 0$					
lasso	5.86	0.68	5.73	4.29	0.541 (0.009)
grouped lasso	5.81	0.65	5.81	13.57	0.525 (0.010)
grouped elastic net	5.97	1.12	5.97	15.30	0.540 (0.007)
$n = 100, \omega = 0.5$					
lasso	5.31	0.79	4.66	4.30	0.360 (0.015)
grouped lasso	4.76	0.48	4.76	10.96	0.317 (0.014)
grouped elastic net	5.52	0.91	5.52	13.77	0.369 (0.010)
$n = 200, \omega = 0$					
lasso	6.00	0.78	5.99	4.26	0.591 (0.004)
grouped lasso	5.99	0.94	5.99	14.80	0.580 (0.004)
grouped elastic net	6.00	1.32	6.00	15.96	0.582 (0.004)
$n = 200, \omega = 0.5$					
lasso	5.92	0.72	5.80	4.34	0.416 (0.006)
grouped lasso	5.85	0.68	5.85	13.74	0.403 (0.006)
grouped elastic net	6.00	0.99	6.00	14.97	0.416 (0.004)

Table A4: Averaged simulation results for link function misspecification based on 100 runs.

Scenario	r	coefficient-		MSE_η
		C	IC	
A	0.5	5.90	4.04	0.457
	0.7	5.97	4.20	0.313
	1.0	6.00	4.41	0.160
	1.3	6.00	4.60	0.162
	1.5	6.00	4.59	0.256
B	0.5	5.88	4.04	0.546
	0.7	5.94	4.12	0.391
	1.0	6.00	4.41	0.160
	1.3	6.00	5.44	0.165
	1.5	6.00	6.13	0.534

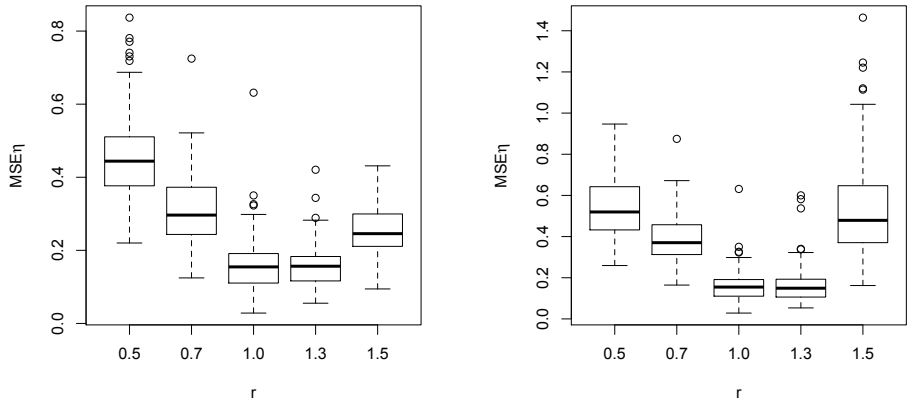

Figure A1: Boxplots of MSE_η for link function misspecification. Left panel: scenario (i). Right panel: scenario (ii).

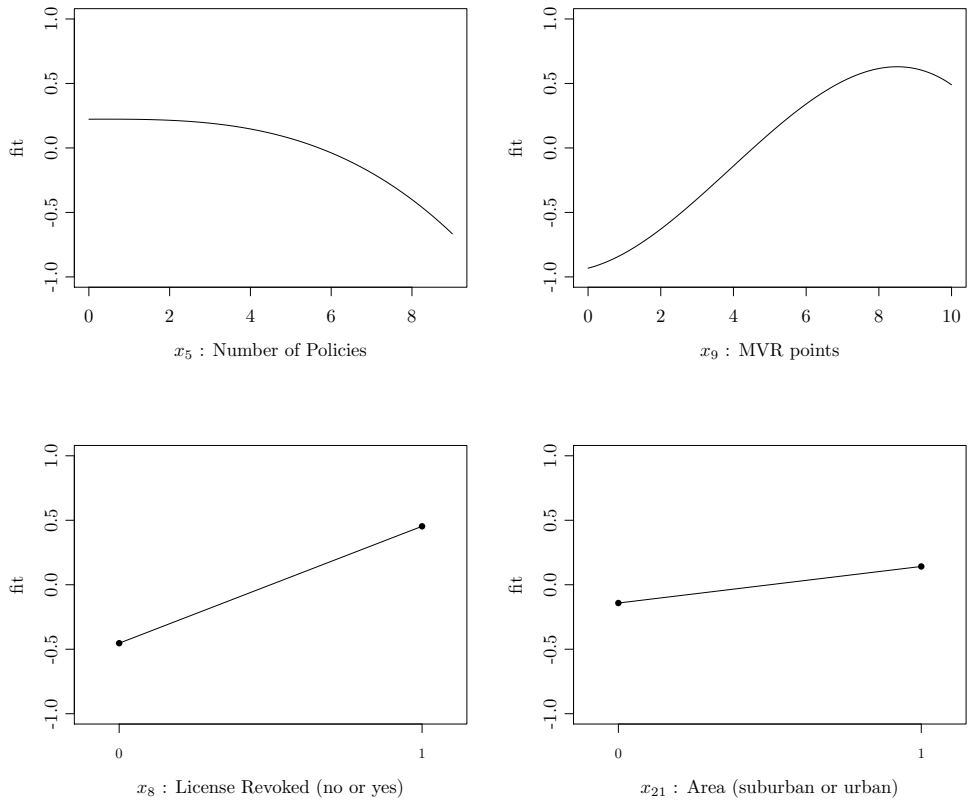

Figure A2: The partial fits of the selected variables by the GrpNet model for the auto insurance claim data.