Submitted to *INFORMS Journal on Computing*
manuscript

Online Supplement: Knowledge Learning of Insurance Risks Using Dependence Models

Zifeng Zhao
Mendoza College of Business, University of Notre Dame
zzhao2@nd.edu

Peng Shi
Wisconsin School of Business, University of Wisconsin - Madison
pshi@bus.wisc.edu,

Xiaoping Feng
CapitalG, San Francisco, CA
xfeng@uwalumni.com,

---

§A. Dispersion Matrix

In this section, we derive the implied dispersion matrix induced by the latent Gaussian process (3) and (4) of the main text. We derive the result by examining the pairwise correlation within and between geographical clusters:

(a) The correlation between repeated observations for a given policyholder is

$$\mathrm{Corr}(\eta_{ijt}, \eta_{ijt'}) = \gamma^{|t-t'|} \frac{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2}{\sigma_\eta^2 (1-\gamma^2)} = \gamma^{|t-t'|}.$$

(b) The correlation between two different policyholders within the same geographical cluster is

$$\begin{aligned}\mathrm{Corr}(\eta_{ijt}, \eta_{ij't'}) &= \frac{1}{\sigma_\eta^2(1-\gamma^2)} \mathrm{Cov}(w_{ijt}, w_{ij't'}) = \gamma^{|t-t'|} \frac{\sigma_u^2 + \sigma_v^2}{\sigma_\eta^2(1-\gamma^2)} \\ &= \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \gamma^{|t-t'|} = \kappa_1 \gamma^{|t-t'|}.\end{aligned}$$

(c) The correlation between two policyholders from different geographical clusters is

$$\begin{aligned}\mathrm{Corr}(\eta_{ijt}, \eta_{i'j't'}) &= \frac{1}{\sigma_\eta^2(1-\gamma^2)} \mathrm{Cov}(u_{it}, u_{i't'}) = \gamma^{|t-t'|} \frac{\sigma_u^2 \tau_{ii'}}{\sigma_\eta^2(1-\gamma^2)} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \tau_{ii'} \gamma^{|t-t'|} = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2} \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2} \tau_{ii'} \gamma^{|t-t'|} = \kappa_1 \kappa_2 \tau_{ii'} \gamma^{|t-t'|}.\end{aligned}$$

§B. Weighting Scheme for the Composite Likelihood

Denote $d_S(i,i')$ as the Euclidean distance between cluster i and i', and denote $d_T(t,t') = |t-t'|$ as the time lag between time t and t'.

The weighting scheme used in the simulations in Section 4.3 of the main text is

$$\omega(ijt, i'j't') = \begin{cases} 1 & \text{if } i = i', j = j', d_T(t,t') \leq 3 \\ & \text{or } i = i', j \neq j', d_T(t,t') \leq 2 \\ & \text{or } i \neq i', d_S(i,i') \leq 2, d_T(t,t') \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

The weighting scheme used in the insurance dataset is the same as the one in Section 4.3 of the main text, except that we require $d_S(i,i') \leq 50$ kilometers.

§C. Simulation Procedure for the Predictive Distributions

In this section, we show how to simulate observations $\widetilde{\boldsymbol{Y}}_{K^+}$ from the predictive distribution

$$F_{\widetilde{\boldsymbol{Y}}_{K^+}|\boldsymbol{Y}}(\tilde{\boldsymbol{y}}_{K^+}|\tilde{y}_1,\ldots,\tilde{y}_K) = \Pr(\widetilde{\boldsymbol{Y}}_{K^+} \leq \tilde{\boldsymbol{y}}_{K^+}|\widetilde{Y}_1 = \tilde{y}_1,\ldots,\widetilde{Y}_K = \tilde{y}_K),$$

where $\widetilde{\boldsymbol{Y}}_{K^+}$ denotes the vector of outcomes to be predicted. Though $F_{\widetilde{\boldsymbol{Y}}_{K^+}|\boldsymbol{Y}}(\tilde{\boldsymbol{y}}_{K^+}|\tilde{y}_1,\ldots,\tilde{y}_K)$ does not have a closed form solution, an accurate approximation can be made based on a large simulated sample of $\widetilde{\boldsymbol{Y}}_{K^+}|\boldsymbol{Y}$ from it. In this paper, we set the total number of simulation for $\widetilde{\boldsymbol{Y}}_{K^+}|\boldsymbol{Y}$ to be 10,000.

For the Independence model, the random vector $\widetilde{\boldsymbol{Y}}_{K^+}$ can be simulated individually due to independence. Denote $\widetilde{\boldsymbol{Y}}_{K+j}$ as the jth random variable in $\widetilde{\boldsymbol{Y}}_{K^+}$, we have

$$\widetilde{\boldsymbol{Y}}_{K+j} \sim F_j(\hat{\mu}_j, \hat{p}, \hat{\phi}),$$

where $F_j(\cdot;\hat{\mu}_j,\hat{p},\hat{\phi})$ is the estimated Tweedie cdf of $\widetilde{\boldsymbol{Y}}_{K+j}$.

For the Copula dependence model, the simulation is based on the latent representation of $(\widetilde{\boldsymbol{Y}}_{K^+}, \boldsymbol{Y})$. Let $(\widetilde{\boldsymbol{\eta}}_{K^+}, \boldsymbol{\eta})$ denote the latent Gaussian process that corresponds to $(\widetilde{\boldsymbol{Y}}_{K^+}, \boldsymbol{Y})$, by (3) and (4) of the main text, we have that

$$\begin{pmatrix} \boldsymbol{\eta} \\ \widetilde{\boldsymbol{\eta}}_{K^+} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{12}^T & \hat{\boldsymbol{\Sigma}}_{22} \end{pmatrix}\right],$$

where $\hat{\mathbf{\Sigma}}_{11} = [\hat{\mathbf{\Sigma}}_{ii'}]_{m\times m} \otimes \mathbf{\Omega}_T^{AR}(\hat{\gamma})$ is the correlation matrix of the latent vector $\boldsymbol{\eta}$ corresponding to m clusters of observed loss costs of insured risks over T years, $\hat{\mathbf{\Sigma}}_{22} = [\hat{\mathbf{\Sigma}}_{ii'}]_{\tilde{m}\times\tilde{m}}$ is the correlation matrix of the latent vector $\widetilde{\boldsymbol{\eta}}_{K^+}$ for the renewal and new policyholders in year $T+1$, and $(\hat{\mathbf{\Sigma}}_{12})_{ij}$ is the correlation between the ith random variable in $\boldsymbol{\eta}$ and the jth random variable in $\widetilde{\boldsymbol{\eta}}_{K^+}$, which can be calculated according to (a)$-$(c) in Section §A.

To simulate $\widetilde{\boldsymbol{Y}}_{K^+}|\boldsymbol{Y}$, we first simulate $\widetilde{\boldsymbol{\eta}}_{K^+}|\boldsymbol{\eta}$ via

$$\widetilde{\boldsymbol{\eta}}_{K^+}|\boldsymbol{\eta} \sim N\left(\hat{\mathbf{\Sigma}}_{12}^T\hat{\mathbf{\Sigma}}_{11}^{-1}\boldsymbol{\eta}, \hat{\mathbf{\Sigma}}_{22} - \hat{\mathbf{\Sigma}}_{12}^T\hat{\mathbf{\Sigma}}_{11}^{-1}\hat{\mathbf{\Sigma}}_{12}\right),$$

using the fact that the conditional distribution of a multivariate Gaussian random vector is still multivariate Gaussian. For a systematic overview of simulation for multivariate Gaussian distribution, we refer the readers to the monograph by Genz and Bretz (2009).

The jth random variable $\widetilde{\boldsymbol{Y}}_{K^+j}$ in $\widetilde{\boldsymbol{Y}}_{K^+}$ is then generated via

$$\widetilde{\boldsymbol{Y}}_{K^+j} \sim F_j^{-1}\left(\Phi(\widetilde{\boldsymbol{\eta}}_{K^+j}); \hat{\mu}_j, \hat{p}, \hat{\phi}\right),$$

where $F_j(\cdot; \hat{\mu}_j, \hat{p}, \hat{\phi})$ is the estimated Tweedie cdf of $\widetilde{\boldsymbol{Y}}_{K^+j}$.

Remark: Note that we do not observe $\boldsymbol{\eta}$ directly, we instead observe a one-sided truncated version of $\boldsymbol{\eta}$ through $\boldsymbol{Y}$ since $\boldsymbol{Y}_i = F_i^{-1}(\Phi(\boldsymbol{\eta}_i))$, where $\boldsymbol{Y}_i$ and $\boldsymbol{\eta}_i$ denote the ith random variable of $\boldsymbol{Y}$ and $\boldsymbol{\eta}$, and $F_i(\cdot)$ is the true Tweedie cdf of $\boldsymbol{Y}_i$. Here, for computational efficiency, we approximate $\boldsymbol{\eta}$ via $\boldsymbol{\eta}_i \approx \Phi^{-1}(F_i(\boldsymbol{Y}_i; \hat{\mu}_i, \hat{p}, \hat{\phi}))$, where $F_i(\cdot; \hat{\mu}_i, \hat{p}, \hat{\phi})$ is the estimated Tweedie cdf of $\boldsymbol{Y}_i$.

§D. Additional Simulations

In this section, we conduct additional numerical experiments to examine the performance of MCL and TCL under the equal cluster size setting, where we keep the simulation setting the same as in Section 4.3 of the main text except that we now set the cluster size to be equal among all spatial clusters, with $n = 3$ or 6. For $n = 3$, we further conduct simulations with number of years of observations $T = 10$.

The simulation results are summarized in Tables S.1 to S.6: Tables S.1-S.2 give the result for $n = 3, T = 5$, Tables S.3-S.4 report the result for $n = 6, T = 5$ and Tables S.5-S.6 give the result for $n = 3, T = 10$. The finding is consistent with the result presented in Section 4.3 of the main text under the unequal cluster size setting, i.e., both MCL and TCL give

Table S.1 Performance of MCL and TCL for marginal parameters under cluster size $n=3$ and time $T=5$ (Remark: MCL and TCL have the same marginal estimation result.)

$m=10\times 10$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.312	1.043	1.039	3.073	1.033	1.561	0.508	-1.564	-1.021	1.018	1.494	5.018
Bias	-0.188	0.043	0.039	0.073	0.033	0.061	0.008	-0.064	-0.021	0.018	-0.006	0.018
S.D.	0.683	0.346	0.319	0.322	0.314	0.300	0.123	0.314	0.253	0.093	0.018	0.133
$m=20\times 20$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.451	1.002	1.006	3.015	1.008	1.510	0.500	-1.500	-1.014	1.005	1.497	5.012
Bias	-0.049	0.002	0.006	0.015	0.008	0.010	-0.000	-0.000	-0.014	0.005	-0.003	0.012
S.D.	0.406	0.169	0.151	0.166	0.152	0.146	0.060	0.140	0.121	0.053	0.007	0.072
$m=30\times 30$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.462	1.014	1.006	3.017	1.011	1.514	0.503	-1.505	-1.008	1.003	1.497	5.012
Bias	-0.038	0.014	0.006	0.017	0.011	0.014	0.003	-0.005	-0.008	0.003	-0.003	0.012
S.D.	0.282	0.121	0.112	0.122	0.111	0.105	0.040	0.101	0.088	0.034	0.006	0.053

Table S.2 Performance of MCL and TCL for dependence parameters under cluster size $n=3$ and time $T=5$

	Multilevel (MCL)					Two-stage (TCL)				
$m=10\times 10$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.462	0.464	0.719	8.212	1.9	0.455	0.466	0.702	8.341	12.6
Bias	-0.038	-0.036	0.019	-0.788	–	-0.045	-0.034	0.002	-0.659	–
S.D.	0.050	0.057	0.157	3.669	–	0.059	0.056	0.158	3.464	–
$m=20\times 20$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.485	0.487	0.704	8.798	8.9	0.482	0.488	0.700	8.798	53.0
Bias	-0.015	-0.013	0.004	-0.202	–	-0.018	-0.012	0.000	-0.202	–
S.D.	0.030	0.031	0.082	2.445	–	0.036	0.030	0.082	2.299	–
$m=30\times 30$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.493	0.491	0.697	8.722	29.0	0.491	0.491	0.693	8.774	156.5
Bias	-0.007	-0.009	-0.003	-0.278	–	-0.009	-0.009	-0.007	-0.226	–
S.D.	0.020	0.023	0.052	1.645	–	0.024	0.022	0.052	1.518	–

Table S.3 Performance of MCL and TCL for marginal parameters under cluster size $n=6$ and time $T=5$ (Remark: MCL and TCL have the same marginal estimation result.)

$m=10\times 10$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.409	1.024	1.017	3.037	1.023	1.532	0.504	-1.540	-1.037	1.007	1.495	5.008
Bias	-0.091	0.024	0.017	0.037	0.023	0.032	0.004	-0.040	-0.037	0.007	-0.005	0.008
S.D.	0.608	0.230	0.215	0.228	0.209	0.198	0.089	0.198	0.185	0.074	0.012	0.118
$m=20\times 20$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.475	1.006	1.004	3.010	1.006	1.508	0.502	-1.515	-1.004	1.002	1.497	5.009
Bias	-0.025	0.006	0.004	0.010	0.006	0.008	0.002	-0.015	-0.004	0.002	-0.003	0.009
S.D.	0.345	0.115	0.109	0.126	0.103	0.104	0.047	0.105	0.094	0.039	0.006	0.069
$m=30\times 30$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.452	1.000	1.002	3.013	1.003	1.506	0.504	-1.507	-1.006	1.005	1.497	5.016
Bias	-0.048	-0.000	0.002	0.013	0.003	0.006	0.004	-0.007	-0.006	0.005	-0.003	0.016
S.D.	0.245	0.076	0.072	0.090	0.069	0.072	0.030	0.071	0.065	0.027	0.005	0.049

consistent parameter estimation, while MCL achieves similar estimation accuracy as TCL with around only $1/7\sim 1/5$ computational time of TCL.

We further analyze the computational complexity of MCL in terms of the number of clusters m, the number of years T and the cluster size n. Based on the weighting scheme in Section §B, simple algebra shows that the computational complexity of the multilevel like-

Table S.4 Performance of MCL and TCL for dependence parameters under cluster size $n=6$ and time $T=5$

	Multilevel (MCL)					Two-stage (TCL)				
$m=10\times10$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.465	0.468	0.725	7.795	5.3	0.450	0.472	0.702	8.017	47.9
Bias	-0.035	-0.032	0.025	-1.205	–	-0.050	-0.028	0.002	-0.983	–
S.D.	0.043	0.047	0.139	3.425	–	0.061	0.046	0.143	3.228	–
$m=20\times20$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.488	0.488	0.700	8.607	31.9	0.483	0.489	0.693	8.680	225.7
Bias	-0.012	-0.012	0.000	-0.393	–	-0.017	-0.011	-0.007	-0.320	–
S.D.	0.027	0.028	0.066	2.200	–	0.038	0.027	0.066	2.082	–
$m=30\times30$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.494	0.492	0.704	8.489	152.3	0.489	0.493	0.700	8.487	703.9
Bias	-0.006	-0.008	0.004	-0.511	–	-0.011	-0.007	-0.000	-0.513	–
S.D.	0.021	0.020	0.041	1.348	–	0.029	0.017	0.042	1.228	–

Table S.5 Performance of MCL and TCL for marginal parameters under cluster size $n=3$ and time $T=10$

(Remark: MCL and TCL have the same marginal estimation result.)

$m=10\times10$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.410	1.020	1.019	3.040	1.027	1.528	0.504	-1.531	-1.026	1.008	1.496	5.012
Bias	-0.090	0.020	0.019	0.040	0.027	0.028	0.004	-0.031	-0.026	0.008	-0.004	0.012
S.D.	0.493	0.265	0.241	0.244	0.232	0.227	0.093	0.225	0.187	0.067	0.012	0.094
$m=20\times20$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.500	0.999	1.002	3.007	0.998	1.507	0.500	-1.505	-1.007	1.000	1.498	5.007
Bias	-0.000	-0.001	0.002	0.007	-0.002	0.007	0.000	-0.005	-0.007	-0.000	-0.002	0.007
S.D.	0.291	0.133	0.115	0.120	0.112	0.113	0.043	0.102	0.092	0.036	0.006	0.056
$m=30\times30$	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	p	$\log\phi$
Mean	2.453	1.006	1.005	3.016	1.008	1.510	0.501	-1.506	-1.009	1.004	1.498	5.014
Bias	-0.047	0.006	0.005	0.016	0.008	0.010	0.001	-0.006	-0.009	0.004	-0.002	0.014
S.D.	0.203	0.084	0.078	0.078	0.072	0.073	0.032	0.074	0.065	0.024	0.005	0.043

Table S.6 Performance of MCL and TCL for dependence parameters under cluster size $n=3$ and time $T=10$

	Multilevel (MCL)					Two-stage (TCL)				
$m=10\times10$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.476	0.480	0.720	8.465	3.9	0.472	0.482	0.707	8.558	27.7
Bias	-0.024	-0.020	0.020	-0.535	–	-0.028	-0.018	0.007	-0.442	–
S.D.	0.035	0.045	0.113	3.227	–	0.042	0.043	0.116	2.975	–
$m=20\times20$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.492	0.492	0.701	8.867	25.2	0.490	0.493	0.700	8.840	137.9
Bias	-0.008	-0.008	0.001	-0.133	–	-0.010	-0.007	0.000	-0.160	–
S.D.	0.021	0.025	0.057	1.866	–	0.024	0.024	0.059	1.778	–
$m=30\times30$	γ	κ_1	κ_2	α	Time (min)	γ	κ_1	κ_2	α	Time (min)
Mean	0.496	0.496	0.704	8.759	132.7	0.498	0.497	0.696	8.947	526.5
Bias	-0.004	-0.004	0.004	-0.241	–	-0.002	-0.003	-0.004	-0.053	–
S.D.	0.015	0.015	0.039	1.169	–	0.018	0.014	0.040	1.065	–

lihood functions $l_i(\boldsymbol{\theta}_i;\boldsymbol{y}), i=0,1,2,3$ of MCL (see detailed definition on page 16) is roughly linear w.r.t. the number of clusters m and number of years T. As can be seen from Tables 3, S.2 S.4 and S.6, empirically, the computational time of MCL does increase approximately at a linear rate as m grows. Comparing the computational time for MCL in Table S.2 ($n=3, T=5$) and Table S.6 ($n=3, T=10$), the computational time of MCL increases roughly at a linear rate as T grows. Under the equal cluster size setting, it is straightforward to show that the computational complexity of likelihood functions $l_i(\boldsymbol{\theta}_i;\boldsymbol{y}), i=0,1,2$ is linear and $l_3(\boldsymbol{\theta}_i;\boldsymbol{y})$ is roughly quadratic w.r.t. the cluster size n. Comparing the computational time for MCL in Table S.2 ($n=3, T=5$) and Table S.4 ($n=6, T=5$), empirically, we do observe an approximately quadratic relationship between the computational time and the cluster size n.

§E. Auto Insurance Case Study

This section provides an additional case study to demonstrate the generalizability and superiority of the proposed copula model for insurance risk learning/prediction.

Data

We consider the personal automobile insurance market where multiple vehicles within a single household are covered by an umbrella policy. We examine a portfolio of such insurance policies obtained from a private property-casualty insurer. The portfolio consists of 253 households of which each owns 3 cars. All households are covered by a comprehensive policy that provides coverage for both collision and third party liabilities. The portfolio is observed for 4 years from 2003 to 2006. The outcome of interest is a vehicle's annual loss cost. Similar to the government property insurance dataset in the main text, a large proportion (92.01%) of the annual loss cost of a vehicle is zero. The mean of the annual loss cost is 352.40 and the standard deviation is 9856.68.

The data demonstrates a hierarchical structure where each policy covers multiple cars within the same household and each car is repeatedly observed over time. Thus, there are two sources of dependence in the data: 1. the within-household dependence among vehicles from a given household, and 2. the temporal dependence for a given vehicle.

Model

We model the vehicle annual loss costs via the proposed copula regression model in the main text. Specifically, the marginal distribution of univariate loss cost is specified as the Tweedie generalized linear model, and the temporal and within-household dependence among loss costs is specified using a Gaussian copula. Further, loss costs between different households are assumed to be independent.

This model can be regarded as a special case of the spatio-temporal copula model presented in the main text. Relating to the case study in the main text, one can think of a household as a cluster consisting of multiple vehicles, comparing to a geographical region consisting of several government entities, and we assume independence between households as opposed to the spatial dependence between geographical regions.

The copula model allows the insurer to learn the vehicle risk using dependence from two dimensions. The first is the within-household dependence among vehicles in a given household, and the second is the temporal dependence for a given vehicle. In comparison, the traditional experience rating currently used in the insurance industry only utilizes the temporal dependence for each vehicle and ignores the within-household dependence.

Risk Prediction

To demonstrate the superior performance of the proposed copula model over the traditional experience rating approach in risk learning, we conduct a comparative analysis for risk prediction similar to the one in Section 6.1 of the main text.

Specifically, the data are split into a training set for developing the model and a test set for out-of-sample prediction performance evaluation. To construct the training data, we randomly select 2 cars from each household, and use their observations in 2003-2005 for model development. This way, the constructed training data has both the within-household dependence among vehicles in a given household and the temporal dependence for a given vehicle. The data observed in 2006 is reserved as the test set.

We examine the prediction performance of the two methods for both renewal and new vehicles in year 2006. The ordered Lorenz curves are presented in Figure S.1 and the associated Gini indices are reported in Table S.7. For both groups, the pure premium calculated from the independence model is used as the base premium, which can be regarded as the manual premium with no experience rating implemented.

For the group of renewal vehicles, we consider two versions of experience rating as competing premium: the traditional experience rating that allows risk learning over temporal dependence only and the copula model that allows risk learning over both temporal and within-household dependence. The Gini indices are 43.38 and 28.24 respectively for the copula approach and the traditional experience rating approach, suggesting that by utilizing within-household dependence, the insurer can learn additional information about the risk of renewal vehicles compared with the traditional experience rating which only employs the temporal dependence.

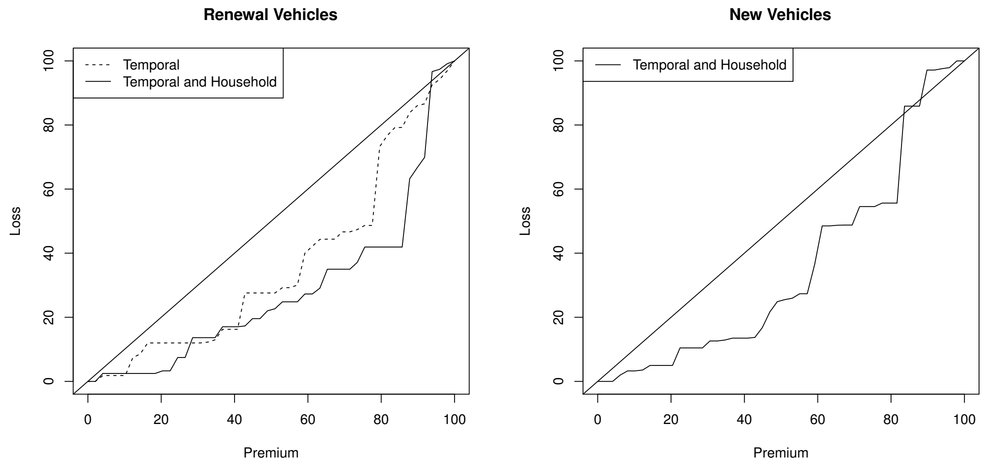

Figure S.1 Ordered Lorenz curves for the portfolio of renewal and new vehicles.

For the group of new vehicles, the experience rating via copula is used as the competing premium. Note that for a new vehicle, the traditional experience rating is equivalent to the base premium provided by the independence model, as the insurer does not have prior observations of new vehicles and thus cannot learn their risks from the past. In contrast, the proposed copula model allows the insurer to learn the risks from existing vehicles within the same household. The Gini index given by the copula model is 28.20, further indicating significant improvement of risk prediction given by the copula model over the traditional experience rating.

Table S.7 Gini indices for hold-out sample validation. The standard error is reported in the brackets.

Renewal vehicle		New vehicle
temporal	temporal+household	temporal+household
28.24	43.38	28.20
(12.60)	(13.79)	(13.91)

This result is consistent with the result in Section 6.1 of the main text. The message conveyed here is that with multidimensional customer learning based on the proposed copula dependence model, the insurers can better understand the risks of new vehicles by incorporating the claim experience of related existing vehicles within the same household, thus achieves better ratemaking performance than using the traditional experience rating currently used in the insurance industry.

§F. Definition of the ordered Lorenz curve and the Gini index

In this section, we give a formal definition of the ordered Lorenz curve and the Gini index. For more details, we refer to Frees et al. (2011).

The essential idea of the ordered Lorenz curve is to measure the discrepancy between the premium and loss distributions. Let $B(\boldsymbol{x})$ be the base premium and $P(\boldsymbol{x})$ be the competeing premium, both depending on a set of exogenous rating variables $\boldsymbol{x}$. The ordered premium and loss distributions are defined based on the relativity $R(\boldsymbol{x}) = P(\boldsymbol{x})/B(\boldsymbol{x})$ as:

$$\hat{H}_P(s) = \frac{\sum_{i=1}^n B(\boldsymbol{x}_i) I(R(\boldsymbol{x}_i) \le s)}{\sum_{i=1}^n B(\boldsymbol{x}_i)} \quad \text{and} \quad \hat{L}_P(s) = \frac{\sum_{i=1}^n y_i I(R(\boldsymbol{x}_i) \le s)}{\sum_{i=1}^n y_i},$$

where y_i denotes the actual loss cost and $\boldsymbol{x}_i$ denotes the exogenous rating variable of the ith policyholder.

The ordered Lorenz curve is the plot of $\left(\hat{H}_P(s), \hat{L}_P(s)\right)$ by varying s from 0 to ∞. The 45-degree line, known as the line of equality, indicates the percentage of losses equals the percentage of premiums. An ordered Lorenz curve below the line of equality suggests that the insurer could look to the competing premium to identify more profitable contracts. The associated Gini index is defined as twice the area between the ordered Lorenz curve and the line of equality and ranges over $(-1, 1)$, or ranges over $(-100, 100)$ if the unit is percentage as the way we report in the main text and the case study. It is calculated as:

$$Gini(H_P(s), L_P(s)) = 2 \int_0^\infty (H_P(s) - L_P(s)) dH_P(s).$$

References

Frees EW, Meyers G, Cummings AD (2011) Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association* 106(495):1085–1098.

Genz A, Bretz F (2009) *Computation of Multivariate Normal and t Probabilities* (Springer-Verlag Berlin Heidelberg), 1st edition.