Semiparametric Vector Generalized Linear Models

Estimation and Computation

Gabriel Dennis

Honours Defence

Supervisor: Dr Alan Huang

23rd of June 2021

Project Goals

This project had the following goals

- generalise the semiparametric GLM of Huang (2014) to deal with vector responses
- write code to fit the model
- verify the properties of the model's estimates via simulation

Vector Generalization (VSPGLM)

Generalising the semiparametric GLM Huang (2014) involves writing the joint exponentially tilted density in the following form

$$dF_i(\boldsymbol{y}) = \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}), \quad i = 1, \ldots, n, \boldsymbol{y} \in \mathbb{R}^K, \tag{1}$$

Vector Generalization (VSPGLM)

Generalising the semiparametric GLM Huang (2014) involves writing the joint exponentially tilted density in the following form

$$dF_i(\boldsymbol{y}) = \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}), \quad i = 1, \ldots, n, \boldsymbol{y} \in \mathbb{R}^K, \tag{1}$$

with slight changes to the normalising and tilt constraints

$$b(\boldsymbol{X}_i, \beta, F) = -\log\left\{\int_{\mathcal{Y}} \exp\{\boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y})\right\}, \tag{2}$$

$$\mu_{(k)}(\boldsymbol{X}_{(k)}^T \boldsymbol{\beta}_{(k)}) = \int_{\mathcal{Y}} y_{(k)} \exp\{b_i + \boldsymbol{\theta}_i^T \boldsymbol{y}\} dF(\boldsymbol{y}), \quad k = 1, \ldots, K. \tag{3}$$

Vector Generalization (VSPGLM)

To estimate the reference distribution and mean model parameters the semiparametric log-likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log p_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i. \tag{4}$$

To estimate the reference distribution and mean model parameters the semiparametric log-likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log p_i + b_i + \boldsymbol{\theta}_i^T \boldsymbol{Y}_i. \tag{4}$$

This also results in changes to the empirical normalising and tilt constraints

$$1 = \sum_{i=1}^{n} p_i \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i\}, \quad j = 1, 2, 3, \ldots, n \tag{5}$$

$$\mu_{(k)}(\boldsymbol{X}_{(k)j}^T \boldsymbol{\beta}_{(k)}) = \sum_{i=1}^{n} p_i Y_{(k)i} \exp\{b_j + \boldsymbol{\theta}_j^T \boldsymbol{Y}_i\} \quad j = 1, \ldots, n, k = 1, \ldots, K. \tag{6}$$

Code

The code to fit the model is written in MATLAB using fmincon, and uses a formula syntax to specify constraints across different components mean models.

Code

The code to fit the model is written in MATLAB using fmincon, and uses a formula syntax to specify constraints across different components mean models.

Separate models

```
model = fit_vspglm(["y_1 ~ x_1", "y_2 ~ x_2"], tbl, links);.
```

Code

The code to fit the model is written in MATLAB using fmincon, and uses a formula syntax to specify constraints across different components mean models.

Separate models

```
model = fit_vspglm(["y_1 ~ x_1", "y_2 ~ x_2"], tbl, links);.
```

Different covariates but equal regression coefficients,

```
model = fit_vspglm(["(y_1, y_2) ~ ((x_1&x_2))], tbl, links);,
```

Code

The code to fit the model is written in MATLAB using fmincon, and uses a formula syntax to specify constraints across different components mean models.

Separate models

```
model = fit_vspglm(["y_1 ~ x_1", "y_2 ~ x_2"], tbl, links);.
```

Different covariates but equal regression coefficients,

```
model = fit_vspglm(["(y_1, y_2) ~ ((x_1&x_2))], tbl, links);,
```

same covariates and regression coefficients,

```
model = fit_vspglm(["(y_1, y_2) ~ x"], tbl, links);
```

Simulations

Simulations from non-standard generalized linear mixed models and generalized linear mixed effects models showed comparable parameter coverage and type I error rates to those of Huang (2014) and the multivariate density ratio model of Marchese (2018), using similar sample sizes.

Simulations

Tri-variate Poisson simulation.

$$\alpha \sim \mathcal{N}(0, \sigma_0^2), \boldsymbol{X} \sim \mathrm{U}[-1,1]^3 \quad \sigma \in \mathbb{R}_+$$
$$Y_1|\alpha, \boldsymbol{X}_{(1)} \sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + \alpha))$$
$$Y_2|\alpha, \boldsymbol{X}_{(2)} \sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} + 0.5\alpha))$$
$$Y_3|\alpha, \boldsymbol{X}_{(3)} \sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)} - 0.3\alpha))$$

Table 1: Simulation results using a sample size of $n = 200$ and $N = 1000$ simulations

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	Errors		$p \leqslant 0.05$	CI		
			$\hat{\sigma}$	$\bar{\mathrm{se}}(\hat{\boldsymbol{\beta}})$		90%	95%	99%
0.4	0.40	0.004	0.13	0.12	0.88	0.89	0.94	0.99
-0.8	-0.79	0.002	0.13	0.14	0.99	0.93	0.96	0.99
0	0.001	0.001	0.13	0.12	0.05	0.89	0.95	0.99

Future Work

There are several areas which future work can take place

- reduce the number of parameters optimized over
- increase the numeric stability of the code
- write a R package for the VSPGLM
- verify properties of the VSPGLM theoretically

Bibliography i

Huang, Alan (2014). "Joint estimation of the mean and error distribution in generalized linear models". In: *Journal of the American Statistical Association* 109.505, pp. 186–196.

Marchese, Scott (2018). "Semiparametric Regression Methods for Mixed Type Data Analysis". Thesis.

Trivariate Poisson Model

$$
\begin{aligned}
\alpha &\sim \mathcal{N}(0, \sigma_0^2), \quad \sigma \in \mathbb{R}_+ \\
\boldsymbol{X} &\sim \mathrm{U}[-1, 1]^3 \\
Y_1 | \alpha, \boldsymbol{X}_{(1)} &\sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(1)}^T \boldsymbol{\beta}_{(1)} + 0.4\alpha)) \\
Y_2 | \alpha, \boldsymbol{X}_{(2)} &\sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(2)}^T \boldsymbol{\beta}_{(2)} - 0.5\alpha)) \\
Y_3 | \alpha, \boldsymbol{X}_{(3)} &\sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(3)}^T \boldsymbol{\beta}_{(3)}))
\end{aligned}
$$

Trivariate Poisson Simulation Results

Table 2: Simulation results for trivariate Poisson model using a sample size of $n = 200$ and $N = 1000$ simulations

$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	Errors		$p \leqslant 0.05$	CI		
			$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$		90%	95%	99%
0.4	0.40	0.004	0.13	0.12	0.88	0.89	0.94	0.99
-0.8	-0.79	0.002	0.13	0.14	0.99	0.93	0.96	0.99
0	0.001	0.001	0.13	0.12	0.05	0.89	0.95	0.99

Trivariate Mixed Effects Simulation

$$
\begin{aligned}
\alpha &\sim \mathcal{N}(0,\ \sigma_0^2) \\
\boldsymbol{X} &\sim \mathrm{U}[-1,1]^3 \\
\boldsymbol{Y}_1|\boldsymbol{X}_{(1)},\alpha &\sim \mathcal{N}(\boldsymbol{X}_{(1)}^T\boldsymbol{\beta}_{(1)}+\alpha,\ \sigma_1^2) \\
\boldsymbol{Y}_2|\boldsymbol{X}_{(2)},\alpha &\sim \mathrm{Pois}(\exp(\boldsymbol{X}_{(2)}^T\boldsymbol{\beta}_{(2)}-0.5\alpha)) \\
\boldsymbol{Y}_3|\boldsymbol{X}_{(3)},\alpha &\sim \mathrm{Gamma}(\lambda,\ \exp(\boldsymbol{X}_{(3)}^T\boldsymbol{\beta}_{(3)}+0.4\alpha))
\end{aligned}
$$

Trivariate Mixed Effects Simulation Results

Table 3: Simulation results for trivariate mixed effects model using a sample size of $n = 200$ and $N = 1000$ simulations

Margin	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$	$\lvert\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rvert$	Errors		$p \leqslant 0.05$	CI		
				$\hat{\sigma}$	$\bar{\text{se}}(\hat{\boldsymbol{\beta}})$		90%	95%	99%
Normal	1	1.005	0.0058	0.17	0.18	1	0.90	0.95	0.99
Poisson	-0.5	-0.49	0.0005	0.13	0.13	0.96	0.83	0.88	0.93
Gamma	0.4	0.39	0.004	0.12	0.13	0.85	0.89	0.93	0.97

Multivariate Normal Simulation Results

Table 4: Simulation results for bivariate normal model using sample size of $n = 200$ and $N = 1000$ simulations

β	$\hat{\beta}$	$\lvert\beta - \hat{\beta}\rvert$	Errors		$p \leqslant 0.05$	CI		
			$\hat{\sigma}$	$\bar{\text{se}}(\hat{\beta})$		90%	95%	99%
-1	-0.99	0.0004	0.11	0.11	1	0.91	0.95	0.99
0	-0.0004	0.0004	0.11	0.11	0.05	0.90	0.95	0.99
0.5	0.49	0.009	0.14	0.13	0.95	0.88	0.94	0.98
2.2	2.19	0.003	0.14	0.14	1	0.90	0.94	0.98

F simulation results

Table 5: Type 1 errors at significance levels of $0.10, 0.05$ and 0.01 using sample sizes of $n = 75, 150$, and $N = 3000$ simulations

	Type 1 Errors		
n	0.10	0.05	0.01
75	0.119	0.060	0.012
150	0.097	0.050	0.013