Vector regression without marginal distributions or association structures

Alan Huang

School of Mathematics and Physics
University of Queensland

1 Dec, 2015

Vector regression

Often more than one response variable of interest.

Vector regression

Often more than one response variable of interest.

E.g. (GDP per head, Fertility rate) may be jointly associated with percentage population in urban areas.

Vector regression

Often more than one response variable of interest.

E.g. (GDP per head, Fertility rate) may be jointly associated with percentage population in urban areas.

E.g. Song (2007) models (Burn severity, Incidence of death) jointly as function of age of patient

Vector regression

Often more than one response variable of interest.

E.g. (GDP per head, Fertility rate) may be jointly associated with percentage population in urban areas.

E.g. Song (2007) models (Burn severity, Incidence of death) jointly as function of age of patient

Main obstacle for vector regression – difficult to specify appropriate joint response distributions for the data, especially for vectors of mixed type.

Vector regression

Specialised bivariate models do exist:

Vector regression

Specialised bivariate models do exist:

Continuous–continuous response pairs:

Vector regression

Specialised bivariate models do exist:

Continuous–continuous response pairs: $(Y_1, Y_2 | X_1, X_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

Vector regression

Specialised bivariate models do exist:

Continuous–continuous response pairs: $(Y_1, Y_2 | X_1, X_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

$$\mu_1 = \mu_1(X_1^T \beta_1),\ \mu_2 = \mu_2(X_2^T \beta_2)$$

Vector regression

Specialised bivariate models do exist:

Continuous–continuous response pairs: $(Y_1, Y_2 | X_1, X_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

$\mu_1 = \mu_1(X_1^T \beta_1)$, $\mu_2 = \mu_2(X_2^T \beta_2)$
Σ typically constant for all X_1, X_2,

Vector regression

Specialised bivariate models do exist:

Continuous–continuous response pairs: $(Y_1, Y_2|X_1, X_2) \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

$\mu_1 = \mu_1(X_1^T \beta_1)$, $\mu_2 = \mu_2(X_2^T \beta_2)$
Σ typically constant for all X_1, X_2,
but can be $\Sigma = \Sigma(\mu_1, \mu_2, \gamma)$ in general

Vector regression

Count–count response pairs:

Vector regression

Count–count response pairs: There is no widely-accepted general bivariate Poisson distribution...

Vector regression

Count–count response pairs: There is no widely-accepted general bivariate Poisson distribution... handling both positive and negative correlations.

Vector regression

Binary–continuous response pairs:

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;

Vector regression

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;
- a 2×2 association matrix (not interpretable as correlation matrix) between Y_1 and Y_2 ;

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;
- a 2×2 association matrix (not interpretable as correlation matrix) between Y_1 and Y_2 ;
- a copula function to combine marginal distributions and association matrix into a joint distribution.

Vector regression

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;
- a 2×2 association matrix (not interpretable as correlation matrix) between Y_1 and Y_2 ;
- a copula function to combine marginal distributions and association matrix into a joint distribution.

Mixed response types are particularly difficult to model.

Vector regression

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;
- a 2×2 association matrix (not interpretable as correlation matrix) between Y_1 and Y_2 ;
- a copula function to combine marginal distributions and association matrix into a joint distribution.

Mixed response types are particularly difficult to model. Model misspecification can happen on these three levels.

Vector regression

Binary–continuous response pairs: Perhaps specify (Song, 2007)

- Y_1 marginally Binomial(p_1), Y_2 marginally normal $N(\mu_2, \sigma^2)$;
- a 2×2 association matrix (not interpretable as correlation matrix) between Y_1 and Y_2 ;
- a copula function to combine marginal distributions and association matrix into a joint distribution.

Mixed response types are particularly difficult to model. Model misspecification can happen on these three levels.

The state-of-the-art `vglm` function in the `vgam` R package (Yee, 2015) currently has no scope for handling mixed responses...

A parsimonious approach

Classical assumptions:

- Marginal mean models

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

Our assumptions:

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

Our assumptions:

- Marginal mean models

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

Our assumptions:

- Marginal mean models
- Data come from *some* multivariate exponential family

A parsimonious approach

Classical assumptions:

- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

Our assumptions:

- Marginal mean models
- Data come from *some* multivariate exponential family that needs not be specified

A parsimonious approach

Classical assumptions:

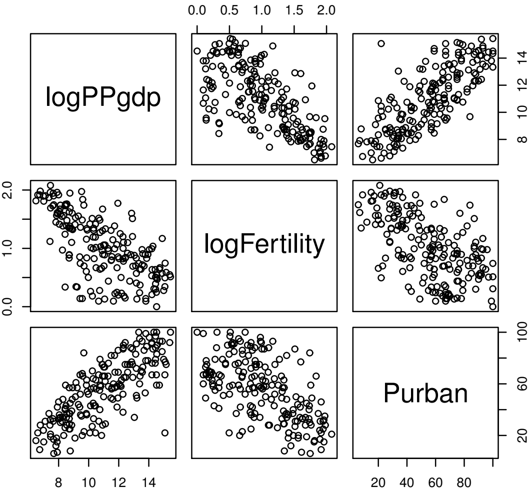
- Marginal mean models
- Marginal distributions /variance functions
- Association matrix
- Copula function

Our assumptions:

- Marginal mean models
- Data come from *some* multivariate exponential family that needs not be specified; parameter space is *all* multivariate exponential families

Example 1: GDP, fertility and urban percentage

Weisberg (2006) describes dataset on GDP per head, fertility rate and percentage of population in urban areas for 193 UN countries.



Example 1: GDP, fertility and urban percentage

We might be interested in how percentage of urban population affects both $(logPPgdp, logFertility)$.

Example 1: GDP, fertility and urban percentage

We might be interested in how percentage of urban population affects both $(logPPgdp, logFertility)$.

Let's specify marginal linear mean model for both responses

$$\begin{aligned} E(\ logPPgdp|Purban) &= \beta_{10} + \beta_{11} Purban \\ E(logFertility|Purban) &= \beta_{20} + \beta_{21} Purban\ . \end{aligned}$$

Example 1: GDP, fertility and urban percentage

We might be interested in how percentage of urban population affects both $(logPPgdp, logFertility)$.

Let's specify marginal linear mean model for both responses

$$\begin{aligned} E(\ logPPgdp|Purban) &= \beta_{10} + \beta_{11}Purban \\ E(logFertility|Purban) &= \beta_{20} + \beta_{21}Purban\ . \end{aligned}$$

We also assume that the joint distributions

$$F(logPPgdp, logFertility|Purban) \sim \text{some bivariate exponential family}$$

Example 1: GDP, fertility and urban percentage

We might be interested in how percentage of urban population affects both $(logPPgdp, logFertility)$.

Let's specify marginal linear mean model for both responses

$$\begin{aligned} E(\ logPPgdp|Purban) &= \beta_{10} + \beta_{11} Purban \\ E(logFertility|Purban) &= \beta_{20} + \beta_{21} Purban\ . \end{aligned}$$

We also assume that the joint distributions

$$F(logPPgdp, logFertility|Purban) \sim \text{some bivariate exponential family}$$

but we do **not** have to specify which particular family – this will be estimated from data using maximum non-parametric likelihood.

Example 1: GDP, fertility and urban percentage

To fit this model, use MATLAB function
`bspglm(y1,y2,x1,x2,link1,link2)`

Example 1: GDP, fertility and urban percentage

To fit this model, use MATLAB function
`bspglm(y1,y2,x1,x2,link1,link2)`

```
[beta, maxloglik, fitted, iter, phat] = bspglm(logPPgdp,
logFertility, Purban, Purban, 'id','id')
```

Example 1: GDP, fertility and urban percentage

To fit this model, use MATLAB function
`bspglm(y1,y2,x1,x2,link1,link2)`

```
[beta, maxloglik, fitted, iter, phat] = bspglm(logPPgdp,
logFertility, Purban, Purban, 'id','id')
```

```
beta{1}
6.9924 0.0730
beta{2}
1.7219 -0.0125
```

Example 1: GDP, fertility and urban percentage

To fit this model, use MATLAB function
`bspglm(y1,y2,x1,x2,link1,link2)`

```
[beta, maxloglik, fitted, iter, phat] = bspglm(logPPgdp,
logFertility, Purban, Purban, 'id','id')
```

```
beta{1}
6.9924 0.0730
beta{2}
1.7219 -0.0125
```

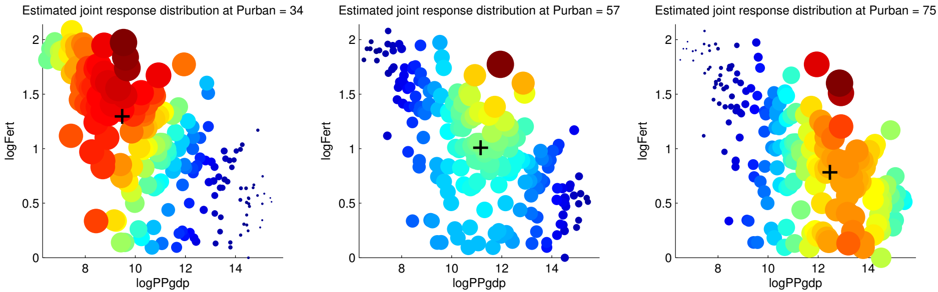
That is,
$$\begin{aligned} \hat{E}(logPPgdp|Purban) &= 6.9924 + 0.0730 * Purban \\ \hat{E}(logFertility|Purban) &= 1.7219 - 0.0125 * Purban \end{aligned}$$

Example 1: GDP, fertility and urban percentage

We can visualise our fitted model using (a primitive) `plot.F()` function.

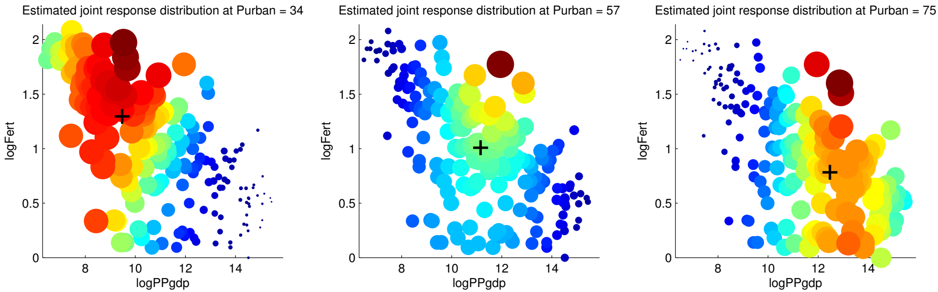
Example 1: GDP, fertility and urban percentage

We can visualise our fitted model using (a primitive) `plot.F()` function.



Example 1: GDP, fertility and urban percentage

We can visualise our fitted model using (a primitive) `plot.F()` function.



Visualising an empirical probability mass function on $\mathbb{R}^2$ is hard...

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y}) \ ,$$

for *some* underlying joint distribution F with density dF.

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y}) \ ,$$

for *some* underlying joint distribution F with density dF.

Underlying distribution F controls the "shape" of the exponential family

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y}) ,$$

for *some* underlying joint distribution F with density dF.

Underlying distribution F controls the "shape" of the exponential family

Includes the multivariate normal as special case.

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y}) \ ,$$

for *some* underlying joint distribution F with density dF.

Underlying distribution F controls the "shape" of the exponential family

Includes the multivariate normal as special case.

F can be zero-inflated, multimodal, mixed measured, etc...

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T\mathbf{y}\right] dF(\mathbf{y}) \ ,$$

for *some* underlying joint distribution F with density dF.

Underlying distribution F controls the "shape" of the exponential family

Includes the multivariate normal as special case.

F can be zero-inflated, multimodal, mixed measured, etc...

Contains infinitely many models, one for each underlying distribution F.

What is the model?

We assume that response vector $\mathbf{Y}$ given covariates X come from *some* **multivariate exponential family**, that is,

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y}) \ ,$$

for *some* underlying joint distribution F with density dF.

Underlying distribution F controls the "shape" of the exponential family

Includes the multivariate normal as special case.

F can be zero-inflated, multimodal, mixed measured, etc...

Contains infinitely many models, one for each underlying distribution F.

Key innovation: We leave underlying joint distribution F unspecified in the model, to be estimated non-parametrically from data.

The model

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y})$$

Canonical parameter vector $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(X; \beta, F)$ controls the mean of $F(\mathbf{y}|X)$:

The model

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T\mathbf{y}\right] dF(\mathbf{y})$$

Canonical parameter vector $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(X; \beta, F)$ controls the mean of $F(\mathbf{y}|X)$:

$$E(\mathbf{Y}|X) = \frac{\int \mathbf{y} \exp\left[\boldsymbol{\theta}^T\mathbf{y}\right] dF(\mathbf{y})}{\int_{\mathbb{R}^d} \exp\left[\boldsymbol{\theta}^T\mathbf{y}\right] dF(\mathbf{y})}$$

The model

$$dF(\mathbf{y}|X) \propto \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y})$$

Canonical parameter vector $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(X; \beta, F)$ controls the mean of $F(\mathbf{y}|X)$:

$$E(\mathbf{Y}|X) = \frac{\int \mathbf{y} \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y})}{\int_{\mathbb{R}^d} \exp\left[\boldsymbol{\theta}^T \mathbf{y}\right] dF(\mathbf{y})} = \begin{pmatrix} \mu_1(X_1^T \beta_1) \\ \vdots \\ \mu_d(X_d^T \beta_d) \end{pmatrix}$$

Estimation and inference via empirical likelihood

To estimate the underlying joint distribution F, we replace F with a set of probability masses $\{p_1, \ldots, p_n\}$ on the observed support $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$.

Estimation and inference via empirical likelihood

To estimate the underlying joint distribution F, we replace F with a set of probability masses $\{p_1, \ldots, p_n\}$ on the observed support $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$.

We then maximise the empirical likelihood in both β and p.

Estimation and inference via empirical likelihood

To estimate the underlying joint distribution F, we replace F with a set of probability masses $\{p_1, \ldots, p_n\}$ on the observed support $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_n\}$.

We then maximise the empirical likelihood in both β and p.

Completely nonparametric – no smoothing parameters or choice of bases

Estimation and inference via empirical likelihood

To estimate the underlying joint distribution F, we replace F with a set of probability masses $\{p_1, \dots, p_n\}$ on the observed support $\{\mathbf{Y}_1, \dots, \mathbf{Y}_n\}$.

We then maximise the empirical likelihood in both β and p.

Completely nonparametric – no smoothing parameters or choice of bases

Retains properties of parametric maximum likelihood estimation:

- consistency;
- asymptotic efficiency;
- asymptotic normality;
- χ^2 likelihood ratio tests. (see Huang, 2015 for more details).

Example 2: burns injury data

Song (2007) examines the relationship between patient *age* (in months) and a continuous–binary response vector *(burn severity, incidence of death)*.

Example 2: burns injury data

Song (2007) examines the relationship between patient *age* (in months) and a continuous–binary response vector *(burn severity, incidence of death)*.

Model:

$$E(\text{burn severity}|\text{age}) = \beta_{10} + \beta_{11}\text{age}$$
$$P(\text{death}|\text{age}) = \frac{\exp(\beta_{20} + \beta_{21}\text{age})}{1 + \exp(\beta_{20} + \beta_{21}\text{age})}$$

(burn severity, death|age) $\sim$ **some** bivariate exponential family

Example 2: burns injury data

Song (2007) examines the relationship between patient *age* (in months) and a continuous–binary response vector *(burn severity, incidence of death)*.

Model:

$$\begin{aligned} E(\text{burn severity}|\text{age}) &= \beta_{10} + \beta_{11}\text{age} \\ P(\text{death}|\text{age}) &= \frac{\exp(\beta_{20} + \beta_{21}\text{age})}{1 + \exp(\beta_{20} + \beta_{21}\text{age})} \end{aligned}$$

$$(\text{burn severity, death}|\text{age}) \sim \textbf{some} \text{ bivariate exponential family}$$

We fit this using

```
[beta1, maxloglik1] = bspglm(burn, death, age, age,
'id','logit')
```

Example 2: burns injury data

The fitted model is

$$\begin{aligned}\hat{E}(\text{burn severity}|\text{age}) &= 6.631 + 0.003 \text{ age} \\ \hat{P}(\text{death}|\text{age}) &= \frac{\exp(-3.737 + 0.044 \text{ age})}{1 + \exp(-3.737 + 0.044 \text{ age})}\end{aligned}$$

Example 2: burns injury data

The fitted model is

$$\begin{aligned}\hat{E}(\text{burn severity}|\text{age}) &= 6.631 + 0.003 \text{ age} \\ \hat{P}(\text{death}|\text{age}) &= \frac{\exp(-3.737 + 0.044 \text{ age})}{1 + \exp(-3.737 + 0.044 \text{ age})}\end{aligned}$$

Song (2007) interested in testing whether age is related to burn severity, $H_0 : \beta_{11} = 0$.

Example 2: burns injury data

The fitted model is

$$\begin{aligned}\hat{E}(\text{burn severity}|\text{age}) &= 6.631 + 0.003 \text{ age} \\ \hat{P}(\text{death}|\text{age}) &= \frac{\exp(-3.737 + 0.044 \text{ age})}{1 + \exp(-3.737 + 0.044 \text{ age})}\end{aligned}$$

Song (2007) interested in testing whether age is related to burn severity, $H_0 : \beta_{11} = 0$.

Fit model *without* age for burn severity,

```
[beta0, maxloglik0] = bspglm(burn, death, 1, age,
'id','logit')
```

Example 2: burns injury data

The fitted model is

$$\begin{aligned}\hat{E}(\text{burn severity}|\text{age}) &= 6.631 + 0.003 \text{ age} \\ \hat{P}(\text{death}|\text{age}) &= \frac{\exp(-3.737 + 0.044 \text{ age})}{1 + \exp(-3.737 + 0.044 \text{ age})}\end{aligned}$$

Song (2007) interested in testing whether age is related to burn severity, $H_0 : \beta_{11} = 0$.

Fit model *without* age for burn severity,

```
[beta0, maxloglik0] = bspglm(burn, death, 1, age,
'id','logit')
```

The p-value for the test is

$$P(\chi_1^2 \geq 2(\text{maxloglik1} - \text{maxloglik0}))$$

Example 2: burns injury data

The fitted model is

$$\begin{aligned}\hat{E}(\text{burn severity}|\text{age}) &= 6.631 + 0.003 \text{ age} \\ \hat{P}(\text{death}|\text{age}) &= \frac{\exp(-3.737 + 0.044 \text{ age})}{1 + \exp(-3.737 + 0.044 \text{ age})}\end{aligned}$$

Song (2007) interested in testing whether age is related to burn severity, $H_0 : \beta_{11} = 0$.

Fit model *without* age for burn severity,

```
[beta0, maxloglik0] = bspglm(burn, death, 1, age,
'id','logit')
```

The p-value for the test is

$$P(\chi^2_1 \geq 2(\text{maxloglik1} - \text{maxloglik0})) = 0.436.$$

Example 2: burns injury data

Can also test the compound hypothesis that age has no relationship with both burn severity and incidence of death, $H_0 : \beta_{11} = \beta_{21} = 0$.

Example 2: burns injury data

Can also test the compound hypothesis that age has no relationship with both burn severity and incidence of death, $H_0 : \beta_{11} = \beta_{21} = 0$.

Fit model without age for *both* components,

```
[beta00, maxloglik00] = bspglm(burn, death, 1, 1,
'id','logit')
```

Example 2: burns injury data

Can also test the compound hypothesis that age has no relationship with both burn severity and incidence of death, $H_0 : \beta_{11} = \beta_{21} = 0$.

Fit model without age for *both* components,
`[beta00, maxloglik00] = bspglm(burn, death, 1, 1, ‘id’,‘logit’)`

The p-value for the test is

$$P(\chi^2_2 \geq 2(\mathrm{maxloglik1} - \mathrm{maxloglik00}))$$

Example 2: burns injury data

Can also test the compound hypothesis that age has no relationship with both burn severity and incidence of death, $H_0 : \beta_{11} = \beta_{21} = 0$.

Fit model without age for *both* components,

```
[beta00, maxloglik00] = bspglm(burn, death, 1, 1,
'id','logit')
```

The p-value for the test is

$$P(\chi^2_2 \geq 2(\text{maxloglik1} - \text{maxloglik00})) < 0.001.$$

Example 2: burns injury data

Can also test the compound hypothesis that age has no relationship with both burn severity and incidence of death, $H_0 : \beta_{11} = \beta_{21} = 0$.

Fit model without age for *both* components,
```
[beta00, maxloglik00] = bspglm(burn, death, 1, 1,
'id','logit')
```

The p-value for the test is

$$P(\chi^2_2 \geq 2(\text{maxloglik1} - \text{maxloglik00})) < 0.001.$$

So, incidence of death is associated with age, but burn severity is not.

Why multivariate exponential families?

Satisfies two basic properties that any vector regression model should satisfy (Song, 2007):

Why multivariate exponential families?

Satisfies two basic properties that any vector regression model should satisfy (Song, 2007):

1. Closed under marginalization: all lower-dimensional regression models have same distributional form.

Why multivariate exponential families?

Satisfies two basic properties that any vector regression model should satisfy (Song, 2007):

1. Closed under marginalization: all lower-dimensional regression models have same distributional form.

2. Arbitrary associations: allows for both positive and negative associations between components of $\mathbf{Y}$.

Why multivariate exponential families?

3. Nonconstant variance-covariance structure is the norm.

Why multivariate exponential families?

3. Nonconstant variance-covariance structure is the norm.

4. Gourieroux et al (1984): Regardless of data-generating mechanism, any exponential family likelihood always produces strongly consistent estimates of mean parameters.

Why multivariate exponential families?

3. Nonconstant variance-covariance structure is the norm.

4. Gourieroux et al (1984): Regardless of data-generating mechanism, any exponential family likelihood always produces strongly consistent estimates of mean parameters. Exponential family likelihoods are the only ones that can do this...!

Why multivariate exponential families?

3. Nonconstant variance-covariance structure is the norm.

4. Gourieroux et al (1984): Regardless of data-generating mechanism, any exponential family likelihood always produces strongly consistent estimates of mean parameters. Exponential family likelihoods are the only ones that can do this...!

5. Hiejima (1997): Any mean-variance relationship can be approximated asymptotically well by some exponential family.

References

Huang, A. (2015) *Vector generalized linear models without marginal distributions and association structures*, submitted.

Song, P. X-K. (2007) *Correlated Data Analysis: Modeling, Analytics and Applications*, Springer Series in Statistics.

Yee, T (2015) *Vector Generalized Linear and Additive Effects*, Springer Series in Statistics.