Minimum Message Length Ridge Regression for Generalized Linear Models

Daniel F. Schmidt and Enes Makalic

Centre for Biostatistics and Epidemiology
The University of Melbourne

26th Australasian Joint Conference on Artificial Intelligence
Dunedin, New Zealand
2013

Outline

Outline

Generalized Linear Models (GLMs) (1)

- We have
  - A vector of targets, $\mathbf{y} \in \mathbb{R}^n$
  - A matrix of features, $\mathbf{X} = (\bar{\mathbf{x}}_1', \ldots, \bar{\mathbf{x}}_n')' \in \mathbb{R}^{n \times p}$

  $\Rightarrow$ Usual problem is to use $\mathbf{X}$ to predict $\mathbf{y}$
- If targets are reals, linear-normal model is a standard choice

$$y_i \sim N(\eta_i, \tau)$$

where

$$\eta_i = \bar{\mathbf{x}}_i \boldsymbol{\beta} + \alpha,$$

is the linear predictor, and

  - $\boldsymbol{\beta} \in \mathbb{R}^p$ are the coefficients
  - $\alpha \in \mathbb{R}$ is the intercept

Generalized Linear Models (GLMs) (1)

- We have
  - A vector of targets, $\mathbf{y} \in \mathbb{R}^n$
  - A matrix of features, $\mathbf{X} = (\bar{\mathbf{x}}_1', \ldots, \bar{\mathbf{x}}_n')' \in \mathbb{R}^{n \times p}$

  $\Rightarrow$ Usual problem is to use $\mathbf{X}$ to predict $\mathbf{y}$
- If targets are reals, linear-normal model is a standard choice

$$y_i \sim N(\eta_i, \tau)$$

where

$$\eta_i = \bar{\mathbf{x}}_i \boldsymbol{\beta} + \alpha,$$

is the linear predictor, and

  - $\boldsymbol{\beta} \in \mathbb{R}^p$ are the coefficients
  - $\alpha \in \mathbb{R}$ is the intercept

Generalized Linear Models (GLMs) (2)

- What if the targets are not continuous variables?
  - Binary data (classification problems)
  - Integers, counts

  $\Rightarrow$ We can use generalized linear models (GLMs)
- GLM framework assumes target distribution satisfies

$$\begin{aligned} \mathrm{E}\left[y_i|\mu_i\right] &= \mu_i \\ \mathrm{var}\left[y_i|\mu_i,\phi\right] &= \phi\, v(\mu_i) \end{aligned}$$

where $\mu_i = f^{-1}(\eta_i)$,

- $f(\cdot)$ is called a link function,
- $v(\cdot)$ is the variance function, and
- ϕ is a dispersion parameter.

Generalized Linear Models (GLMs) (2)

- What if the targets are not continuous variables?
  - Binary data (classification problems)
  - Integers, counts

  $\Rightarrow$ We can use generalized linear models (GLMs)
- GLM framework assumes target distribution satisfies

$$\begin{aligned} \mathrm{E}\left[y_i|\mu_i\right] &= \mu_i \\ \mathrm{var}\left[y_i|\mu_i,\phi\right] &= \phi\, v(\mu_i) \end{aligned}$$

where $\mu_i = f^{-1}(\eta_i)$,

- $f(\cdot)$ is called a link function,
- $v(\cdot)$ is the variance function, and
- ϕ is a dispersion parameter.

Estimating the model (1)

- Let $p(y_i|\boldsymbol{\beta}, \alpha, \phi)$ denote the probability distribution for y_i
- In general $\boldsymbol{\beta}$, α, ϕ are unknown
  $\Rightarrow$ Must be estimated from the data $\mathbf{y}$
- One powerful estimation procedure is ridge regression

$$\left\{\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\phi}\right\} = \underset{\alpha\in\mathbb{R},\phi\in\mathbb{R}_+,\boldsymbol{\beta}\in S(c)}{\arg\min} \left\{-\sum_{i=1}^{n} \log p(y_i|\boldsymbol{\beta}, \alpha, \phi)\right\}$$

where

$$S(c) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \leq c\}$$

- Hyperparameter c controls the amount of regularisation
  $\Rightarrow$ As $c \to \infty$, we recover maximum likelihood

Estimating the model (1)

- Let $p(y_i|\boldsymbol{\beta}, \alpha, \phi)$ denote the probability distribution for y_i
- In general $\boldsymbol{\beta}$, α, ϕ are unknown
  $\Rightarrow$ Must be estimated from the data $\mathbf{y}$
- One powerful estimation procedure is ridge regression

$$\left\{\hat{\boldsymbol{\beta}}, \hat{\alpha}, \hat{\phi}\right\} = \underset{\alpha\in\mathbb{R},\phi\in\mathbb{R}_+,\boldsymbol{\beta}\in S(c)}{\arg\min} \left\{-\sum_{i=1}^{n} \log p(y_i|\boldsymbol{\beta}, \alpha, \phi)\right\}$$

where

$$S(c) = \left\{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \leq c\right\}$$

- Hyperparameter c controls the amount of regularisation
  $\Rightarrow$ As $c \to \infty$, we recover maximum likelihood

Estimating the model (2)

- Ridge regression requires that c also be estimated
- We therefore need to estimate $\boldsymbol{\beta}$, α, ϕ and c
- Further, we would like to determine which of the p features are associated with the target
- We will use minimum message length to solve these problems

Outline

Minimum Message Length (1)

- Practical implementation of theory of inductive inference inspired by Kolmogorov complexity
  - Model that yields the briefest encoding of data in a hypothetical message is optimal
- The message is composed of two-parts
  - assertion, statement describing a particular model $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$
  - detail, encoding of the data $\mathbf{y}$ using the assertion model $\boldsymbol{\theta}$
- Bayesian procedure, requires prior distributions

Minimum Message Length (2)

- The total length of the two-part message, $I(\boldsymbol{\theta}, \mathbf{y})$, is sum of the lengths of the assertion and the detail

$$I(\boldsymbol{\theta}, \mathbf{y}) = I(\boldsymbol{\theta}) + I(\mathbf{y}|\boldsymbol{\theta})$$

- MML advocates choosing model $\boldsymbol{\theta}$ that minimises the codelength of the hypothetical two-part message

- In this paper we used the Wallace–Freeman approximation to the exact message length

Message Lengths of GLMs (1)

- Exploit Bayesian interpretation of ridge regression
- Need to specify prior distributions

$$\pi(\boldsymbol{\beta}, \alpha|\phi, \lambda) = \pi(\boldsymbol{\beta}|\phi, \lambda)\pi(\alpha|\phi)$$

- We choose

$$\begin{aligned} \pi(\boldsymbol{\beta}|\phi, \lambda) &= \left(\frac{\lambda}{2\pi\phi}\right)^{\frac{k}{2}} \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}} \cdot \exp\left(-\frac{\lambda\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2\phi}\right) \\ \pi(\alpha|\phi) &\propto \frac{1}{\sqrt{\phi}} \end{aligned}$$

  where λ is a regularisation hyperparameter
- Conditioning on ϕ decouples $(\boldsymbol{\beta}, \alpha)$ and ϕ

Message Lengths of GLMs (2)

- Applying Wallace–Freeman formula, plus a suitable "correction" arising because of the choice of priors (details in paper) yields a message length

$$I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi | \lambda; \mathbf{X})$$

- Estimate $\boldsymbol{\beta}$, α and ϕ by minimising the message length

$$\left\{\hat{\boldsymbol{\beta}}_\lambda, \hat{\alpha}_\lambda, \hat{\phi}_\lambda\right\} = \arg\min_{\boldsymbol{\beta},\alpha,\phi} \left\{I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi | \lambda; \mathbf{X})\right\}$$

- Use modified iteratively reweighted least-squares algorithm
  $\Rightarrow$ Decoupling between ϕ and $(\boldsymbol{\beta}, \alpha)$ simplifies procedure

Message Lengths of GLMs (2)

- Applying Wallace–Freeman formula, plus a suitable "correction" arising because of the choice of priors (details in paper) yields a message length

$$I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi | \lambda; \mathbf{X})$$

- Estimate $\boldsymbol{\beta}$, α and ϕ by minimising the message length

$$\left\{\hat{\boldsymbol{\beta}}_\lambda, \hat{\alpha}_\lambda, \hat{\phi}_\lambda\right\} = \underset{\boldsymbol{\beta},\alpha,\phi}{\arg\min} \left\{I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi | \lambda; \mathbf{X})\right\}$$

- Use modified iteratively reweighted least-squares algorithm
  $\Rightarrow$ Decoupling between ϕ and $(\boldsymbol{\beta}, \alpha)$ simplifies procedure

Message Lengths of GLMs (3)

- Including λ in message length yields a new message length

$$I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi, \lambda; \mathbf{X}) = I(\mathbf{y}, \boldsymbol{\beta}, \alpha, \phi | \lambda; \mathbf{X}) + \frac{1}{2} \log n$$

- Regularisation parameter can be estimated by minimising message length

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}_+}{\arg\min} \left\{ I(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda, \hat{\alpha}_\lambda, \hat{\phi}_\lambda, \lambda; \mathbf{X}) \right\}$$

Message Lengths of GLMs (4)

- Finally we can select which features are associated with $\mathbf{y}$
- Let $\gamma \subset \{1, \ldots, p\}$ denote a list of columns of $\mathbf{X}$
  - Let $\mathbf{X}_\gamma$ denote the submatrix indexed by γ
  - Let $\pi(\gamma)$ be a prior for γ

- Estimate the "best" subset by minimising the message length:

$$\hat{\gamma} = \arg\min_{\gamma} \left\{ I(\mathbf{y}, \hat{\boldsymbol{\beta}}_{\hat{\lambda}}, \hat{\alpha}_{\hat{\lambda}}, \hat{\phi}_{\hat{\lambda}}, \hat{\lambda}; \mathbf{X}_\gamma) - \log \pi(\gamma) \right\}$$

$\Rightarrow$ MML estimates all parameters using a single criterion

Outline

Parameter Estimation (1)

- Parameter estimation experiment
- Compared MML ridge estimates against
  - Maximum likelihood
  - Corrected Akaike information criterion (AIC_c) ridge estimates
- Experimental parameters
  - Distributions: Normal, Binomial, Poisson
  - Sample sizes, $n = \{25, 50, 100, 250\}$
  - Number of features, $p = 10$
  - Feature Toeplitz correlations, $\rho = \{0.1, 0.5, 0.9\}$
  - True coefficients, $\beta_i \sim N(0, 1)$
- Kullback–Leibler divergence used as loss function

Parameter Estimation (1)

- Parameter estimation experiment
- Compared MML ridge estimates against
  - Maximum likelihood
  - Corrected Akaike information criterion (AIC_c) ridge estimates
- Experimental parameters
  - Distributions: Normal, Binomial, Poisson
  - Sample sizes, $n = \{25, 50, 100, 250\}$
  - Number of features, $p = 10$
  - Feature Toeplitz correlations, $\rho = \{0.1, 0.5, 0.9\}$
  - True coefficients, $\beta_i \sim N(0, 1)$
- Kullback–Leibler divergence used as loss function

Parameter Estimation (2)

		$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
	n	AIC_c	MML	AIC_c	MML	AIC_c	MML
Normal	25	0.66	0.47	0.57	0.45	0.31	0.35
	50	0.91	0.70	0.86	0.71	0.64	0.62
	100	0.96	0.85	0.95	0.84	0.83	0.79
	250	0.98	0.94	0.98	0.94	0.93	0.90
Binomial	25	0.03	0.05	0.02	0.06	0.02	0.03
	50	0.56	0.22	0.51	0.19	0.27	0.19
	100	0.82	0.69	0.77	0.64	0.63	0.56
	250	0.96	0.91	0.94	0.89	0.89	0.88
Poisson	25	0.77	0.58	0.83	0.62	0.46	0.39
	50	0.94	0.96	0.96	0.95	0.92	0.89
	100	1.00	1.00	0.99	0.98	0.97	0.97
	250	1.00	1.00	1.00	1.00	1.00	1.00

Medians of the ratios of the Kullback–Leibler (KL) divergences obtained by the MML and AIC_c estimates over the KL divergence obtained by the maximum likelihood estimates.

Example: Dog Bite Data

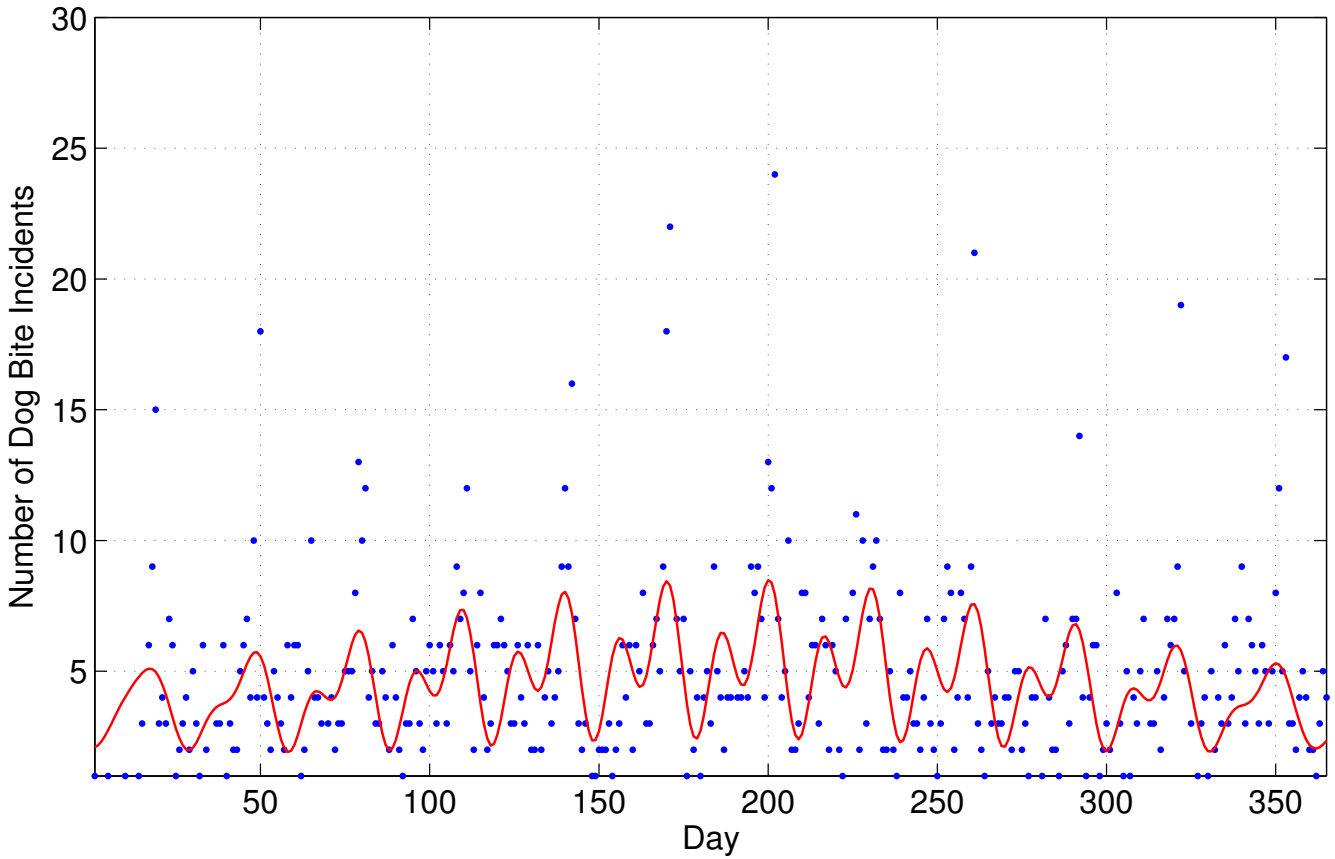
- Conclude with an example on real data
- Dog bite data obtained for a light hearted paper
- Counts of daily dog bite incidents
  - Collected between 1 July 1997 and 30 June 1998
  - Taken from the Australian Institute of Health and Welfare Databse of Australian Hospital Statistics
  - $n = 365$ observations
- Hypothesis: Do dogs bite more frequently on the full moon?
- Our approach: use smoothing with discrete cosine transform bases to look for periodicities
  $\Rightarrow$ Regression where features are DCT bases

Example: Dog Bite Data

- Conclude with an example on real data
- Dog bite data obtained for a light hearted paper
- Counts of daily dog bite incidents
  - Collected between 1 July 1997 and 30 June 1998
  - Taken from the Australian Institute of Health and Welfare Databse of Australian Hospital Statistics
  - $n = 365$ observations
- Hypothesis: Do dogs bite more frequently on the full moon?
- Our approach: use smoothing with discrete cosine transform bases to look for periodicities
  $\Rightarrow$ Regression where features are DCT bases

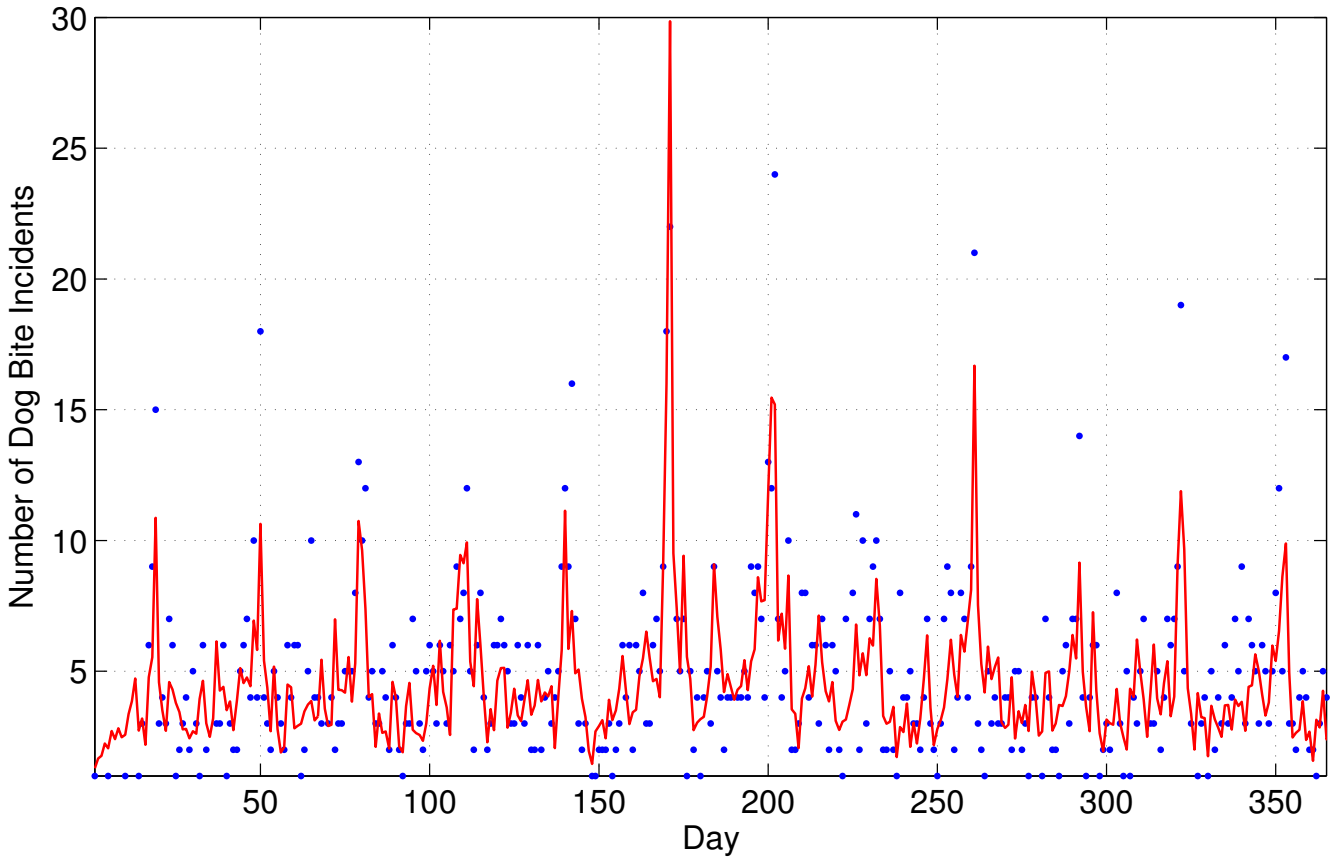
Example: Dog Bite Data (1)

- Gaussian regression model: $I = 1,001.32$ (4 retained features)



Example: Dog Bite Data (2)

- Poisson regression model: $I = 947.55$ (23 retained features)



Example: Dog Bite Data (3)

- Poisson regression model fits better
  $\Rightarrow$ Count data, so not unexpected
- Interestingly, strongest feature has period of 31 days!
  - Moon was in first quarter (first half-moon) at start of time series
- So maybe dogs bite more on half-moons? :)

Conclusion

- Potential future applications/extensions:
  - GLM Kernel machines (MML "support vector machines")
  - MML GLM LASSO regression
  - Testing for associations/interactions in genetic data
- Questions?