Bilinear Text Regression and Applications

Vasileios Lampos
Department of Computer Science
University College London

March, 2015

Outline

$\perp$ **Linear Regression Methods**

$\dashv$ **Bilinear Regression Methods**

$\dashv$ **Applications**

$\models$ **Conclusions**

Recap on regression methods

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

Ordinary Least Squares (OLS)

$$\underset{\boldsymbol{w},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta - \sum_{j=1}^{m} x_{ij} w_j \right)^2$$

or in *matrix form*

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2, \text{ where } \boldsymbol{X}_* = [\boldsymbol{X} \ \operatorname{diag}(\boldsymbol{I})]$$

$$\Rightarrow \boldsymbol{w}_* = \left(\boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{X}_* \right)^{-1} \boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{y}$$

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

Ordinary Least Squares (OLS)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 \Rightarrow \boldsymbol{w}_* = \left(\boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{X}_*\right)^{-1} \boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{y}$$

Why not?

- $\boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{X}_*$ may be singular (thus difficult to invert)
- high-dimensional models difficult to interpret
- unsatisfactory prediction accuracy (estimates have large variance)

Regression basics — Ridge Regression (1/2)

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

Ridge Regression (RR)

$$\boldsymbol{w}_* = \underbrace{\left(\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_* + \boxed{\lambda\boldsymbol{I}}\right)}_{\text{non singular}}^{-1}\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{y}$$ (Hoerl & Kennard, 1970)

$$\underset{\boldsymbol{w},\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(y_i - \beta - \sum_{j=1}^{m} x_{ij}w_j\right)^2 + \lambda\sum_{j=1}^{m} w_j^2\right\}$$

or $\underset{\boldsymbol{w}_*}{\operatorname{argmin}}\left\{\|\boldsymbol{X}_*\boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \lambda\|\boldsymbol{w}\|_{\ell_2}^2\right\}$

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

Ridge Regression (RR)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \boxed{\lambda \|\boldsymbol{w}\|_{\ell_2}^2} \right\}$$

\+ size constraint on the weight coefficients (**regularisation**)
$\rightarrow$ resolves problems caused by collinear variables

\+ less degrees of freedom, better predictive accuracy than OLS

− does **not** perform feature selection (nonzero coefficients)

Regression basics — Lasso

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

$\boldsymbol{\ell_1}$–norm regularisation or **lasso** (Tibshirani, 1996)

$$\underset{\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta - \sum_{j=1}^{m} x_{ij} w_j \right)^2 + \boxed{\lambda \sum_{j=1}^{m} |w_j|} \right\}$$

$$\text{or } \underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \boxed{\lambda \|\boldsymbol{w}\|_{\ell_1}} \right\}$$

− no closed form solution — quadratic programming problem

+ Least Angle Regression explores entire reg. path (Efron *et al.*, 2004)

+ sparse $\boldsymbol{w}$, interpretability, better performance (Hastie *et al.*, 2009)

− if $m > n$, at most n variables can be selected

− strongly corr. predictors → model-inconsistent (Zhao & Yu, 2009)

Regression basics — Lasso for Text Regression

- n-gram frequencies $\boldsymbol{x}_i \in \mathbb{R}^m, \quad i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- flu rates $y_i \in \mathbb{R}, \quad i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}, \quad j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

ℓ_1–norm regularisation or **lasso**

or $$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \boxed{\lambda \|\boldsymbol{w}\|_{\ell_1}} \right\}$$

'unwel', 'temperatur', 'headach', 'appetit', 'symptom', 'diarrhoea', 'muscl', 'feel', ...

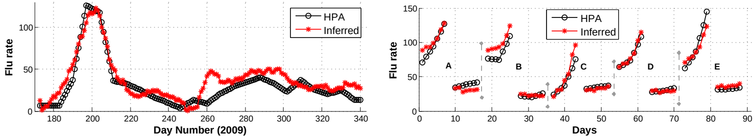


Figure 1 : Flu rate predictions for the UK by applying lasso on Twitter data

(Lampos & Cristianini, 2010)

Regression basics — Elastic Net

- observations $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

[Linear] Elastic Net (LEN)

(Zhou & Hastie, 2005)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \underbrace{\|\boldsymbol{X}_*\boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2}_{\text{OLS}} + \underbrace{\lambda_1 \|\boldsymbol{w}\|_{\ell_2}^2}_{\text{RR reg.}} + \underbrace{\lambda_2 \|\boldsymbol{w}\|_{\ell_1}}_{\text{Lasso reg.}} \right\}$$

+ 'compromise' between ridge regression (handles collinear predictors) and lasso (favours sparsity)
+ entire reg. path can be explored by modifying LAR
+ if $m > n$, number of selected variables not limited to n
− may select redundant variables!

Would a slightly **different text regression** approach be more suitable for **Social Media** content?

About Twitter (1/2)

Tweet Examples

@PaulLondon: I would strongly support a coalition government. It is the best thing for our country right now. #electionsUK2010

@JohnsonMP: Socialism is something forgotten in our country #supportLabour

@FarageNOT: Far-right 'movements' come along with crises in capitalism #UKIP

@JohnK_1999: RT @HannahB: Stop talking about politics and listen to Justin!! Bieber rules, peace and love ♡ ♡ ♡

The Twitter **basics**:

- 140 characters per status (tweet)
- users follow and be followed
- embedded usage of topics (#elections)
- retweets (**RT**), @replies, @mentions, favourites
- real-time nature
- biased user demographics

About Twitter (2/2)

Tweet Examples

@PaulLondon: I would strongly support a coalition government. It is the best thing for our country right now. #electionsUK2010

@JohnsonMP: Socialism is something forgotten in our country #supportLabour

@FarageNOT: Far-right 'movements' come along with crises in capitalism #UKIP

@JohnK_1999: RT @HannahB: Stop talking about politics and listen to Justin!! Bieber rules, peace and love ♡ ♡ ♡

- contains a **vast amount of information** about various topics
- this information ($\boldsymbol{X}$) can be used to assist **predictions** ($\boldsymbol{y}$)
  (Lampos & Cristianini, 2012; Sakaki *et al.*, 2010; Bollen *et al.*, 2011)

– $f : \boldsymbol{X} \rightarrow \boldsymbol{y}$, f usually formulates a **linear** regression task

– $\boldsymbol{X}$ represents word frequencies only...

+ is it possible to incorporate a **user contribution** somehow?

word selection + user selection

Bi-linear Text Regression

Bilinear Text Regression — The general idea (1/2)

Linear regression: $f\left(\boldsymbol{x}_i\right) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + \beta$

• observations	$\boldsymbol{x}_i \in \mathbb{R}^m$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{X}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$w_j, \beta \in \mathbb{R}$,	$j \in \{1, ..., m\}$	—	$\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

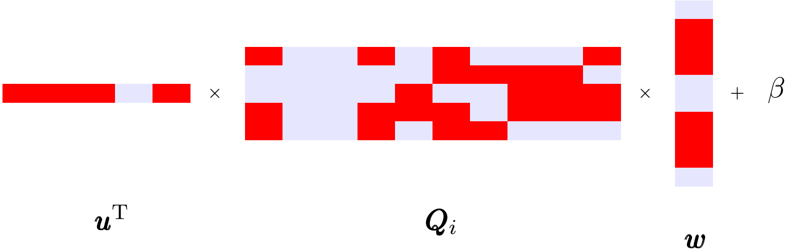
Bilinear regression: $f\left(\boldsymbol{Q}_i\right) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}_i\boldsymbol{w} + \beta$

• users	$p \in \mathbb{Z}^+$			
• observations	$\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{\mathcal{X}}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$u_k, w_j, \beta \in \mathbb{R}$,	$k \in \{1, ..., p\}$ $j \in \{1, ..., m\}$	—	$\boldsymbol{u}$, $\boldsymbol{w}$, β

Bilinear Text Regression — The general idea (2/2)

- users $p \in \mathbb{Z}^{+}$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $u_k, w_j, \beta \in \mathbb{R}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{u}$, $\boldsymbol{w}$, β

$$f\left(\boldsymbol{Q}_i\right) = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta$$



Bilinear Text Regression — Regularisation

- users $p \in \mathbb{Z}^{+}$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $u_k, w_j, \beta \in \mathbb{R}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{u}$, $\boldsymbol{w}$, β

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \psi(\boldsymbol{u}, \theta_u) + \psi(\boldsymbol{w}, \theta_w) \right\}$$

$\psi(\cdot)$: **regularisation function** with a set of hyper-parameters (θ)

- if $\psi(\boldsymbol{v}, \lambda) = \lambda \|\boldsymbol{v}\|_{\ell_1}$ Bilinear Lasso
- if $\psi(\boldsymbol{v}, \lambda_1, \lambda_2) = \lambda_1 \|\boldsymbol{v}\|_{\ell_2}^2 + \lambda_2 \|\boldsymbol{v}\|_{\ell_1}$ Bilinear Elastic Net (**BEN**)

(Lampos *et al.*, 2013)

Bilinear Elastic Net (BEN)

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \lambda_{u_1} \|\boldsymbol{u}\|_{\ell_2}^2 + \lambda_{u_2} \|\boldsymbol{u}\|_{\ell_1} + \lambda_{w_1} \|\boldsymbol{w}\|_{\ell_2}^2 + \lambda_{w_2} \|\boldsymbol{w}\|_{\ell_1} \right\}$$

BEN's objective function

- **Bi-convexity**: fix $\boldsymbol{u}$, learn $\boldsymbol{w}$ and vv
- Iterating through convex optimisation tasks: **convergence** (Al-Khayyal & Falk, 1983; Horst & Tuy, 1996)
- **FISTA** (Beck & Teboulle, 2009) in **SPAMS** (Mairal *et al.*, 2010): Large-scale optimisation solver, quick convergence

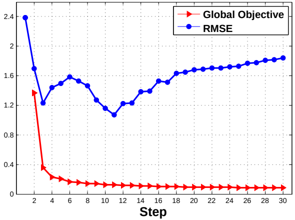


Figure 2 : Objective function value and RMSE (on hold-out data) through the model's iterations

Multi-Task Learning

Multi-Task Learning

What

- Instead of learning/optimising a single task (one target variable)
- ... optimise multiple tasks jointly

Why (Caruana, 1997)

- improves **generalisation performance** exploiting domain-specific information of **related** tasks
- a good choice for under-sampled distributions — knowledge transfer
- application-driven reasons (e.g. explore **interplay** between political parties)

How

- Multi-task regularised regression

The $\boldsymbol{\ell_{2,1}}$-norm regularisation

$$\|\boldsymbol{W}\|_{2,1} = \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_{\ell_2}\,, \text{ where } \boldsymbol{W}_j \text{ denotes the } j\text{-th row}$$

$\boldsymbol{\ell_{2,1}}$-norm regularisation

$$\underset{\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{\ell_F}^2 + \lambda \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_{\ell_2} \right\}$$

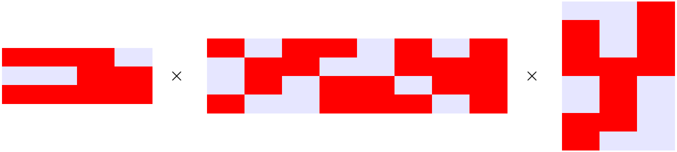
- multi-task learning: instead of $\boldsymbol{w} \in \mathbb{R}^m$, learn $\boldsymbol{W} \in \mathbb{R}^{m\times\tau}$, where τ is the number of tasks
- $\ell_{2,1}$-norm regularisation, i.e. the sum of $\boldsymbol{W}$'s row ℓ_2-norms (Argyriou *et al.*, 2008; Liu *et al.*, 2009) extends the notion of **group lasso** (Yuan & Lin, 2006)
- group lasso: instead of single variables, selects groups of variables
- 'groups' now become the τ-dimensional rows of $\boldsymbol{W}$

Bilinear + Multi-Task Learning

Bilinear Multi-Task Learning

- tasks $\tau \in \mathbb{Z}^+$
- users $p \in \mathbb{Z}^+$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $\boldsymbol{y}_i \in \mathbb{R}^{\tau}$, $i \in \{1, ..., n\}$ — $\boldsymbol{Y}$
- weights, bias $\boldsymbol{u}_k, \boldsymbol{w}_j, \boldsymbol{\beta} \in \mathbb{R}^{\tau}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{\beta}$

$$f\left(\boldsymbol{Q}_i\right) = \operatorname{tr}\left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{W}\right) + \boldsymbol{\beta}$$



$\boldsymbol{U}^{\mathrm{T}}$ $\times$ $\boldsymbol{Q}_i$ $\times$ $\boldsymbol{W}$

Bilinear Group $\ell_{2,1}$ (BGL) (1/2)

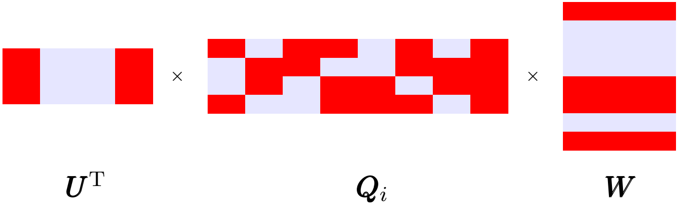
- tasks $\tau \in \mathbb{Z}^{+}$
- users $p \in \mathbb{Z}^{+}$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $\boldsymbol{y}_i \in \mathbb{R}^{\tau}$, $i \in \{1, ..., n\}$ — $\boldsymbol{Y}$
- weights, bias $\boldsymbol{u}_k, \boldsymbol{w}_j, \boldsymbol{\beta} \in \mathbb{R}^{\tau}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{\beta}$

$$\underset{\boldsymbol{U},\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\boldsymbol{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_2 \right\}$$

- BGL can be broken into 2 convex tasks: first learn $\{\boldsymbol{W}, \boldsymbol{\beta}\}$, then $\{\boldsymbol{U}, \boldsymbol{\beta}\}$ and vv + iterate through this process

Bilinear Group $\ell_{2,1}$ (BGL) (2/2)

$$\underset{\boldsymbol{U},\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\boldsymbol{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_2 \right\}$$



- a feature (user/word) is selected for **all tasks** (not just one), but possibly with different weights
- especially useful in the **domain of politics** (e.g. user pro party A, against party B)

Voting Intention Modelling

(Lampos *et al.*, 2013)

Political Opinion/Voting Intention Mining — Brief Recap

Primary papers

- predict the result of an election via Twitter (Tumasjan *et al.*, 2010)
- model socio-political sentiment polls (O'Connor *et al.*, 2010)
- above 2 failed on 2009 US congr. elections (Gayo-Avello, 2011)
- desired properties of such models (Metaxas *et al.*, 2011)

Features

- lexicon-based, e.g. using LIWC (Tausczik & Pennebaker, 2010)
- task-specific keywords (names of parties, politicians)
- tweet volume

reviewed in (Gayo-Avello, 2013)

– political **descriptors change** in time, differ per country
– **personalised** modelling (present in actual polls) missing
– **multi-task** learning?

Voting Intention Modelling — Data (United Kingdom)

- 42K users distributed proportionally to regional population figures
- 60m tweets from 30/04/2010 to 13/02/2012
- $80,976$ unigrams (word features)
- 240 voting intention polls (YouGov)
- 3 parties: Conservatives (**CON**), Labour Party (**LAB**), Liberal Democrats (**LIB**)
- main language: English

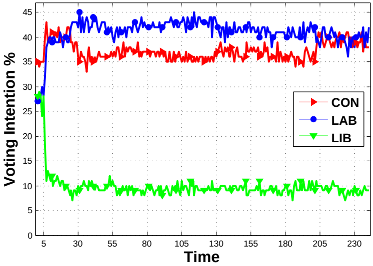


Figure 3 : Voting intention time series for the UK (YouGov)

Voting Intention Modelling — Data (Austria)

- 1.1K users manually selected by Austrian political analysts (SORA)
- 800K tweets from 25/01 to 01/12/2012
- $22,917$ unigrams (word features)
- 98 voting intention polls from various pollsters
- 4 parties: Social Democratic Party (**SPÖ**), People's Party (**ÖVP**), Freedom Party (**FPÖ**), Green Alternative Party (**GRÜ**)
- main language: German

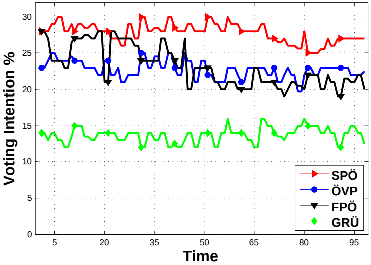


Figure 4 : Voting intention time series for Austria

Voting Intention Modelling — Evaluation

- 10-fold validation
  - train a model using data based on a set of contiguous polls $\mathcal{A}$
  - test on the next $\mathcal{D} = 5$ polls
  - expand training set to $\{\mathcal{A} \cup \mathcal{D}\}$, test on the next $|\mathcal{D}'| = 5$ polls
- **realistic scenario**: train on past, predict future polls
- overall we test predictions on 50 polls (in each case study)

Baselines

- $\mathbf{B}_{\boldsymbol{\mu}}$: constant prediction based on $\mu(\boldsymbol{y})$ in the training set
- $\mathbf{B}_{\mathbf{last}}$: constant prediction based on $\mathrm{last}(\boldsymbol{y})$ in the training set
- **LEN**: (linear) Elastic Net prediction (using word frequencies)

Voting Intention Modelling — Performance tables

Average RMSEs on the voting intention percentage predictions in the 10-step validation process

Table 1 : UK case study

	CON	**LAB**	**LIB**	$\boldsymbol{\mu}$
B$_{\mu}$	2.272	1.663	1.136	1.69
B$_{\text{last}}$	2	2.074	1.095	1.723
LEN	3.845	2.912	2.445	3.067
BEN	1.939	1.644	1.136	1.573
BGL	**1.785**	**1.595**	**1.054**	**1.478**

Table 2 : Austrian case study

	SPÖ	**ÖVP**	**FPÖ**	**GRÜ**	$\boldsymbol{\mu}$
B$_{\mu}$	1.535	1.373	3.3	1.197	1.851
B$_{\text{last}}$	**1.148**	1.556	**1.639**	1.536	1.47
LEN	1.291	1.286	2.039	**1.152**	1.442
BEN	1.392	1.31	2.89	1.205	1.699
BGL	1.619	**1.005**	1.757	1.374	**1.439**

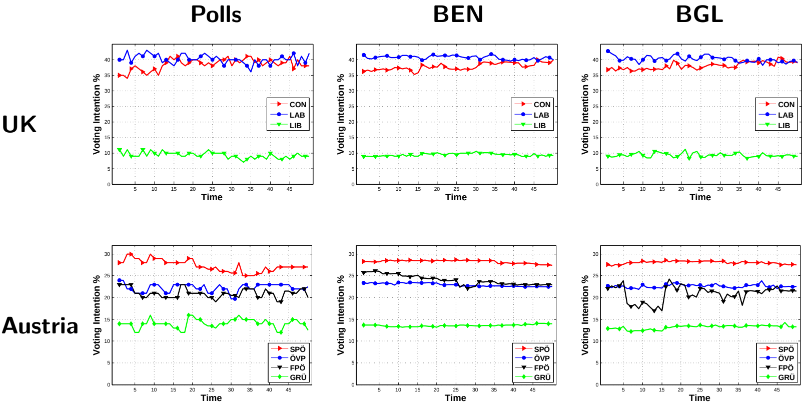


Figure 5 : Performance figures for BEN and BGL in the UK/Austria case studies

Voting Intention Modelling — Qualitative evaluation

Party	Tweet	Score	Author
CON	PM in friendly chat with top EU mate, Sweden's Fredrik Reinfeldt, before family photo	1.334	Journalist
	Have Liberal Democrats broken electoral rules? Blog on Labour complaint to cabinet secretary	-0.991	Journalist
LAB	I am so pleased to hear Paul Savage who worked for the Labour group has been Appointed the Marketing manager for the baths hall GREAT NEWS	-0.552	Politician (Labour)
LBD	RT *@user*: Must be awful for TV bosses to keep getting knocked back by all the women they ask to host election night (via *@user*)	0.874	LibDem MP
SPÖ	Inflationsrate in Ö. im Juli leicht gesunken: von 2,2 auf 2,1%. Teurer wurde Wohnen, Wasser, Energie. **Translation:** *Inflation rate in Austria slightly down in July from 2,2 to 2,1%. Accommodation, Water, Energy more expensive.*	0.745	Journalist
ÖVP	kann das buch "res publica" von johannes #voggenhuber wirklich empfehlen! so zum nachdenken und so... #europa #demokratie **Translation:** *can really recommend the book "res publica" by johannes #voggenhuber! Food for thought and so on #europe #democracy*	-2.323	User
FPÖ	Neue Kampagne der #Krone zur #Wehrpflicht: "GIB BELLO EINE STIMME!" **Translation:** *New campaign by the #Krone on #Conscription: "GIVE WOOFY A VOICE!"*	7.44	Political satire
GRÜ	Protestsong gegen die Abschaffung des Bachelor-Studiums Internationale Entwicklung: <link> #IEbleibt #unibrennt #uniwut **Translation:** *Protest songs against the closing-down of the bachelor course of International Development: <link> #IDremains #uniburns #unirage*	1.45	Student Union

Table 3 : Scored tweet examples from both case studies using BGL

Extracting Socioeconomic Patterns from the News

(Lampos *et al.*, 2014)

Socioeconomic Patterns — Data

News Summaries

- Open Europe Think Tank: summaries of news articles on EU or member countries (focus on politics, perhaps right-wing biased!)
- from February 2006 to mid-November 2013
  1913 days or 94 months or **8 years**
- involving $\mathbf{435}$ international **news outlets**
- extracted $8,413$ unigrams and $19,045$ bigrams

Socioeconomic Indicators

- EU Economic Sentiment Indicator (**ESI**)
  $\rightarrow$ predictor for future economic developments (Gelper & Croux, 2010)
  $\rightarrow$ consists of 5 weighted confidence sub-indicators:
  - industrial (40%), services (30%), consumer (20%) construction (5%), retail trade (5%)
- **EU Unemployment** — seasonally adjusted ratio of the non employed over the entire EU labour force

Socioeconomic Patterns — Task description

- **qualitative differences** to voting intention modelling
  - aim is **NOT** to predict socioeconomic indicators
  - characterise news by conducting a supervised analysis on them **driven by** socioeconomic factors
- use predictive performance as an **informal guarantee** that the model is reasonable
- the better the predictive performance, the more trustful the extracted patterns should be

Slightly modified **BEN**

$$\underset{\boldsymbol{o}\geq 0,\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{o}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \lambda_{o_1} \|\boldsymbol{o}\|_{\ell_2}^2 + \lambda_{o_2} \|\boldsymbol{o}\|_{\ell_1} + \lambda_{w_1} \|\boldsymbol{w}\|_{\ell_2}^2 + \lambda_{w_2} \|\boldsymbol{w}\|_{\ell_1} \right\}$$

- $\min(\boldsymbol{o}) \geq 0$ to enhance weight interpretability for both news outlets and n-grams

Socioeconomic Patterns — Predictive performance

- similar evaluation as in voting intention prediction
- differences: time frame is now a month, train using a moving window of 64 contiguous months, test on the next 3 months
- make predictions for a total of 30 months

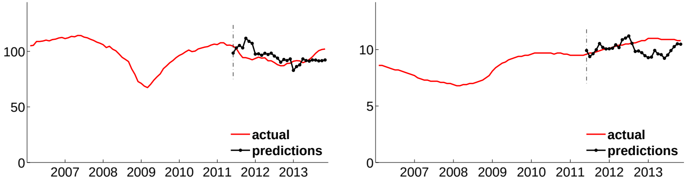


Figure 6 : Monthly rates of EU-wide ESI (right) and Unemployment (left) together with BEN's predictions for the last 30 months

	ESI	**Unemployment**
LEN	9.253 (9.89%)	0.9275 (8.75%)
BEN	$\mathbf{8.209}$ (8.77%)	$\mathbf{0.9047}$ (8.52%)

Table 4 : 10-fold validation average RMSEs (and error rates) for LEN and BEN on ESI and unemployment rates prediction

Socioeconomic Patterns — Qualitative analysis (ESI)

Figure 7 : Visualisation of BEN's outputs for EU's ESI in the last fold (i.e. model trained on 64 months up to August 2013). The word cloud depicts the top-60 positively and negatively weighted n-grams (120) in total together with the top-30 outlets.

Socioeconomic Patterns — Qualitative analysis (Unempl.)

Figure 8 : Visualisation of BEN's outputs for EU-Unemployment in the last fold (i.e. model trained on 64 months up to August 2013). The word cloud depicts the top-60 positively and negatively weighted n-grams (120) in total together with the top-30 outlets.

Conclusions

+ introduced a new class of methods for **bilinear text regression**
+ directly applicable to Social Media content
+ or other types of textual content such as news articles
+ **better predictive performance** than the linear alternative (in the investigated case studies)
+ extended to **bilinear multi-task learning**

To do

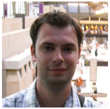
– investigate finer grained modelling settings by applying different regularisation functions (or different combinations of them)

– further understand the properties of bilinear versus linear text regression, e.g. when and why is it a good choice or how different combinations of regularisation settings affect performance

– task-specific improvements

In collaboration with

Trevor Cohn, University of Melbourne



Daniel Preoţiuc-Pietro, University of Pennsylvania

Sina Samangooei, Amazon Research

Douwe Gelling, University of Sheffield

Thank you

Any questions?

Download the slides from

http://www.lampos.net/research/talks-posters

References I

Al-Khayyal and Falk. **Jointly Constrained Biconvex Programming**. MOR, 1983.

Argyriou, Evgeniou and Pontil. **Convex multi-task feature learning**. Machine Learning, 2008.

Beck and Teboulle. **A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems**. J. Imaging Sci., 2009.

Bermingham and Smeaton. **On using Twitter to monitor political sentiment and predict election results**. SAAIP, 2011.

Bollen, Mao and Zeng. **Twitter mood predicts the stock market**. JCS, 2011.

Caruana. **Multitask Learning**. Machine Learning, 1997.

Efron, Hastie, Johnstone and Tibshirani. **Least Angle Regression**. The Annals of Statistics, 2004.

Gayo-Avello. **A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data**. SSCR, 2013.

Gayo-Avello, Metaxas and Mustafaraj. **Limits of Electoral Predictions using Twitter**. ICWSM, 2011.

Gelper and Croux. **On the construction of the European Economic Sentiment Indicator**. OBES, 2010.

Hastie, Tibshirani and Friedman. **The Elements of Statistical Learning**. 2009.

Hoerl and Kennard. **Ridge regression: Biased estimation for nonorthogonal problems**. Technometrics, 1970.

References II

Horst and Tuy. **Global Optimization: Deterministic Approaches**. 1996.

Lampos and Cristianini. **Tracking the flu pandemic by monitoring the Social Web**. CIP, 2010.

Lampos and Cristianini. **Nowcasting Events from the Social Web with Statistical Learning**. ACM TIST, 2012.

Lampos, Preoţiuc-Pietro and Cohn. **A user-centric model of voting intention from Social Media**. ACL, 2013.

Lampos, Preoţiuc-Pietro, Samangooei, Gelling and Cohn. **Extracting Socioeconomic Patterns from the News: Modelling Text and Outlet Importance Jointly**. ACL LACSS, 2014.

Liu, Ji and Ye. **Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization**. UAI, 2009.

Mairal, Jenatton, Obozinski and Bach. **Network Flow Algorithms for Structured Sparsity**. NIPS, 2010.

Metaxas, Mustafaraj and Gayo-Avello. **How (not) to predict elections**. SocialCom, 2011.

O'Connor, Balasubramanyan, Routledge and Smith. **From Tweets to polls: Linking text sentiment to public opinion time series**. ICWSM, 2010.

Pirsiavash, Ramanan and Fowlkes. **Bilinear classifiers for visual recognition**. NIPS, 2009.

Quesada and Grossmann. **A global optimization algorithm for linear fractional and bilinear programs**. JGO, 1995.

References III

Sakaki, Okazaki and Matsuo. **Earthquake shakes Twitter users: real-time event detection by social sensors**. WWW, 2010.

Tausczik and Pennebaker. **The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods**. JLSP, 2010.

Tibshirani. **Regression Shrinkage and Selection via the LASSO**. JRSS, 1996.

Tumasjan, Sprenger, Sandner and Welpe. **Predicting elections with Twitter: What 140 characters reveal about political sentiment**. ICWSM, 2010.

Yuan and Lin. **Model selection and estimation in regression with grouped variables**. JRSS, 2006.

Zhao and Yu. **On model selection consistency of LASSO**. JMLR, 2006.

Zhou and Hastie. **Regularization and variable selection via the elastic net**. JRSS, 2005.