The University Of Sheffield.

THE UNIVERSITY OF MELBOURNE

Joint Emotion Analysis via Multi-task Gaussian Processes

Daniel Beck, Trevor Cohn, Lucia Specia

October 28, 2014

Emotion Analysis

Goal

Automatically detect emotions in a text
[Strapparava and Mihalcea, 2008];

Emotion Analysis

Goal

Automatically detect emotions in a text
[Strapparava and Mihalcea, 2008];

Headline	**Fear**	**Joy**	**Sadness**
Storms kill, knock out power, cancel flights	82	0	60
Panda cub makes her debut	0	59	0

Why Multi-task?

- Learn a model that shows sound and interpretable correlations between emotions.

Why Multi-task?

- Learn a model that shows sound and interpretable correlations between emotions.
- Datasets are scarce and small $\rightarrow$ Multi-task models are able to learn from all emotions jointly;

Why Multi-task?

- Learn a model that shows sound and interpretable correlations between emotions.
- Datasets are scarce and small $\rightarrow$ Multi-task models are able to learn from all emotions jointly;
- Annotation scheme is subjective and fine-grained $\rightarrow$ Prone to bias and noise;

Why Multi-task?

- Learn a model that shows sound and interpretable correlations between emotions.
- Datasets are scarce and small $\rightarrow$ Multi-task models are able to learn from all emotions jointly;
- Annotation scheme is subjective and fine-grained $\rightarrow$ Prone to bias and noise;

Disclaimer: this work is not about features (at the moment...)

Multi-task learning and Anti-correlations

Most multi-task models used in NLP assume some degree of correlation between tasks:

Multi-task learning and Anti-correlations

Most multi-task models used in NLP assume some degree of correlation between tasks:

Domain Adaptation: assumes the existence of a "general" domain-independent knowledge in the data.

Multi-task learning and Anti-correlations

Most multi-task models used in NLP assume some degree of correlation between tasks:

Domain Adaptation: assumes the existence of a "general" domain-independent knowledge in the data.

Annotation Noise Modelling: assumes that annotations are noisy deviations from a "ground truth".

Multi-task learning and Anti-correlations

Most multi-task models used in NLP assume some degree of correlation between tasks:

Domain Adaptation: assumes the existence of a "general" domain-independent knowledge in the data.

Annotation Noise Modelling: assumes that annotations are noisy deviations from a "ground truth".

For Emotion Analysis, we need a multi-task model that is able to take into account possible anti-correlations, avoiding negative transfer.

Headline	**Fear**	**Joy**	**Sadness**
Storms kill, knock out power, cancel flights	82	0	60
Panda cub makes her debut	0	59	0

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Mean function

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Kernel function

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})} \quad \text{Prior}$$

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Likelihood

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})}$$

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})}$$

Marginal likelihood

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Posterior $$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})}$$

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int_f p(y_*|\mathbf{x}_*, f, \mathbf{X}, \mathbf{y})p(f|\mathbf{X}, \mathbf{y})df$$

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})}$$

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int_f p(y_*|\mathbf{x}_*, f, \mathbf{X}, \mathbf{y})p(f|\mathbf{X}, \mathbf{y})df$$

Likelihood (test)

Gaussian Processes

Let $(\mathbf{X}, \mathbf{y})$ be the training data and $f(\mathbf{x})$ the latent function that models that data:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, f)p(f)}{p(\mathbf{y}|\mathbf{X})}$$

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int_f p(y_*|\mathbf{x}_*, f, \mathbf{X}, \mathbf{y})p(f|\mathbf{X}, \mathbf{y})df$$

Predictive distribution

GP Regression

Likelihood: In a regression setting, we usually consider a Gaussian likelihood, which allow us to obtain a closed form solution for the test posterior;

[1] AKA Squared Exponential, Gaussian or Exponential Quadratic kernel.

GP Regression

Likelihood: In a regression setting, we usually consider a Gaussian likelihood, which allow us to obtain a closed form solution for the test posterior;

Kernel: Many options available. In this work we use the Radial Basis Function (RBF) kernel[1]:

$$k(\mathbf{x}, \mathbf{x}') = \alpha_f^2 \times exp\left(-\frac{1}{2}\sum_{i=1}^{F}\frac{(x_i - x_i')^2}{l_i}\right)$$

[1] AKA Squared Exponential, Gaussian or Exponential Quadratic kernel.

The Intrinsic Coregionalisation Model

Coregionalisation models extend GPs to vector-valued outputs [Álvarez et al., 2012]. Here we use the *Intrinsic Coregionalisation Model* (ICM):

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{B}_{d,d'}$$

The Intrinsic Coregionalisation Model

Coregionalisation models extend GPs to vector-valued outputs [Álvarez et al., 2012]. Here we use the *Intrinsic Coregionalisation Model* (ICM):

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{B}_{d,d'}$$

Kernel on data points (like RBF, for instance)

The Intrinsic Coregionalisation Model

Coregionalisation models extend GPs to vector-valued outputs [Álvarez et al., 2012]. Here we use the *Intrinsic Coregionalisation Model* (ICM):

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{B}_{d,d'}$$

Coregionalisation matrix: encodes task covariances

The Intrinsic Coregionalisation Model

Coregionalisation models extend GPs to vector-valued outputs [Álvarez et al., 2012]. Here we use the *Intrinsic Coregionalisation Model* (ICM):

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{B}_{d,d'}$$

$\mathbf{B}$ can be parameterised and learned by optimizing the model marginal likelihood.

PPCA model

[Bonilla et al., 2008] decomposes $\mathbf{B}$ using PPCA:

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T + \text{diag}(\boldsymbol{\alpha}),$$

PPCA model

[Bonilla et al., 2008] decomposes $\mathbf{B}$ using PPCA:

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T + \text{diag}(\boldsymbol{\alpha}),$$

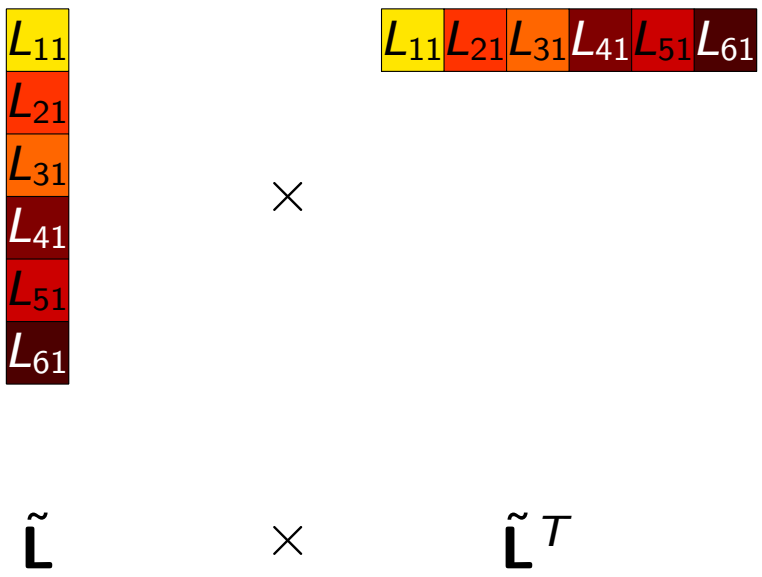
To ensure numerical stability, we employ the incomplete-Cholesky decomposition over $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$:

$$\mathbf{B} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \text{diag}(\boldsymbol{\alpha}),$$

PPCA model

$\tilde{\mathbf{L}}$

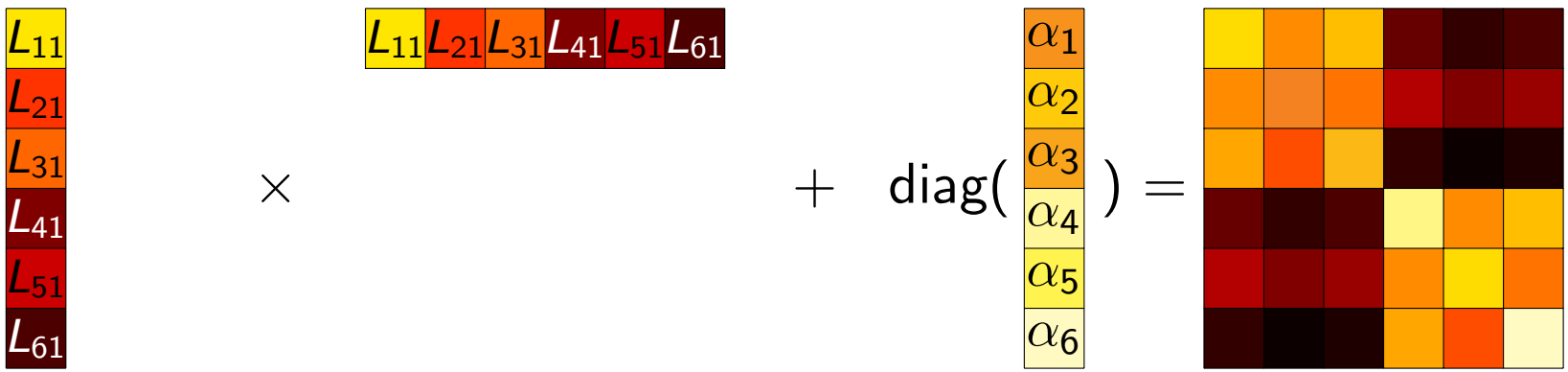
PPCA model


L_{11}
L_{21}
L_{31}
L_{41}
L_{51}
L_{61}
L_{11} L_{21} L_{31} L_{41} L_{51} L_{61}
$\times$
$\tilde{\mathbf{L}}$
$\times$
$\tilde{\mathbf{L}}^T$

PPCA model

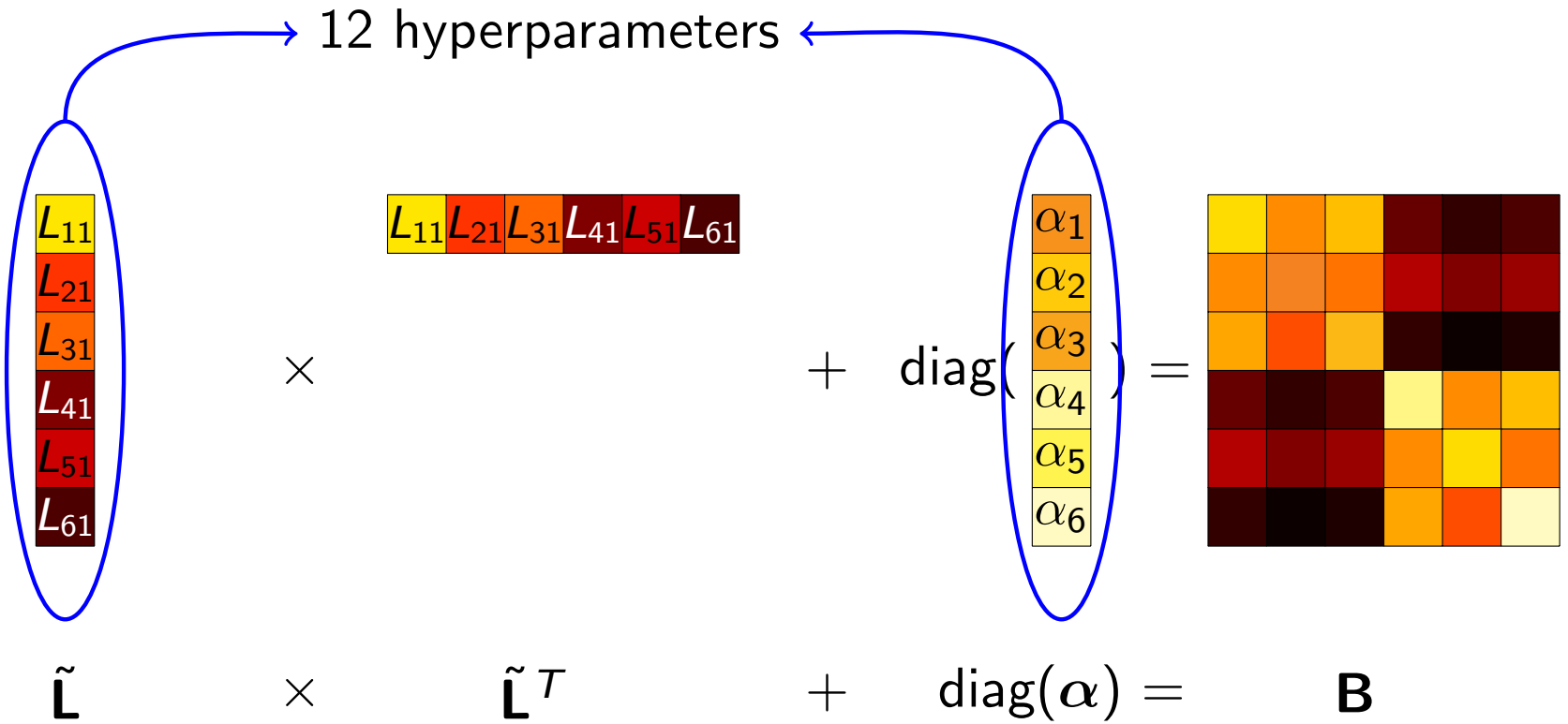
$$\tilde{\mathbf{L}} \times \tilde{\mathbf{L}}^T + \text{diag}(\boldsymbol{\alpha}) =$$

PPCA model

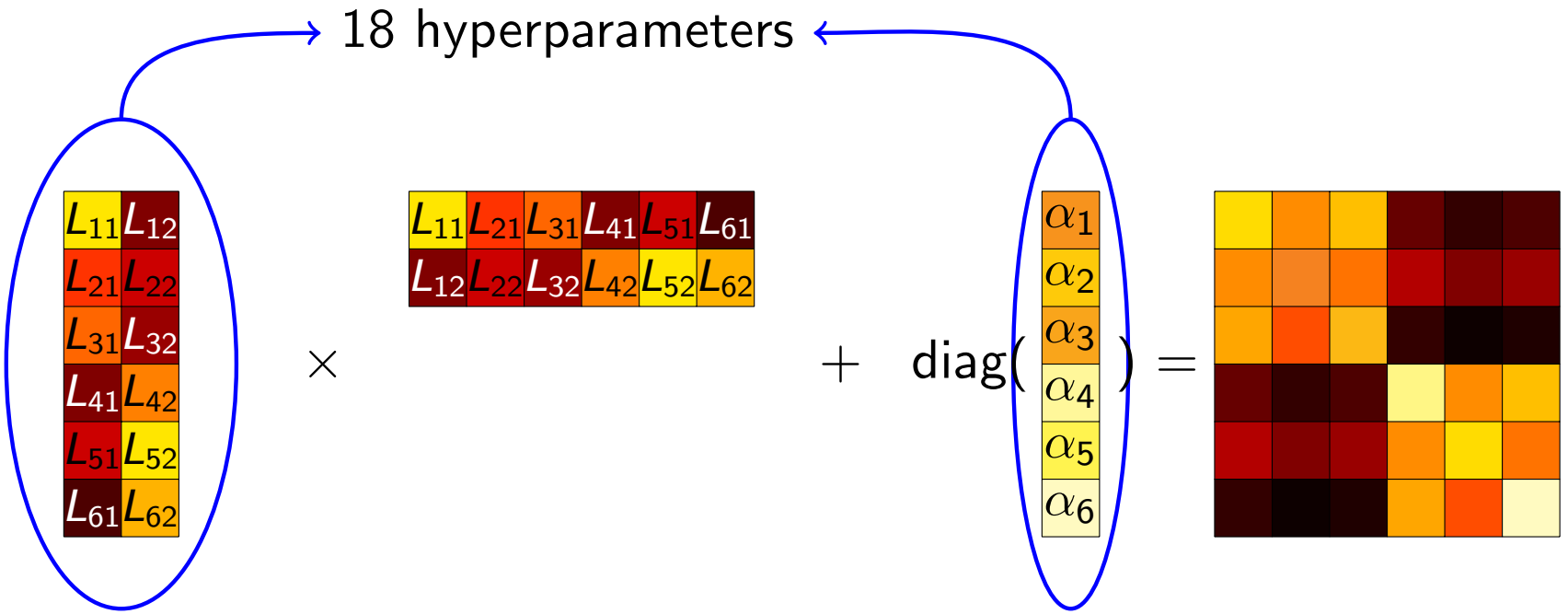


$$\tilde{\mathbf{L}} \times \tilde{\mathbf{L}}^T + \text{diag}(\boldsymbol{\alpha}) = \mathbf{B}$$

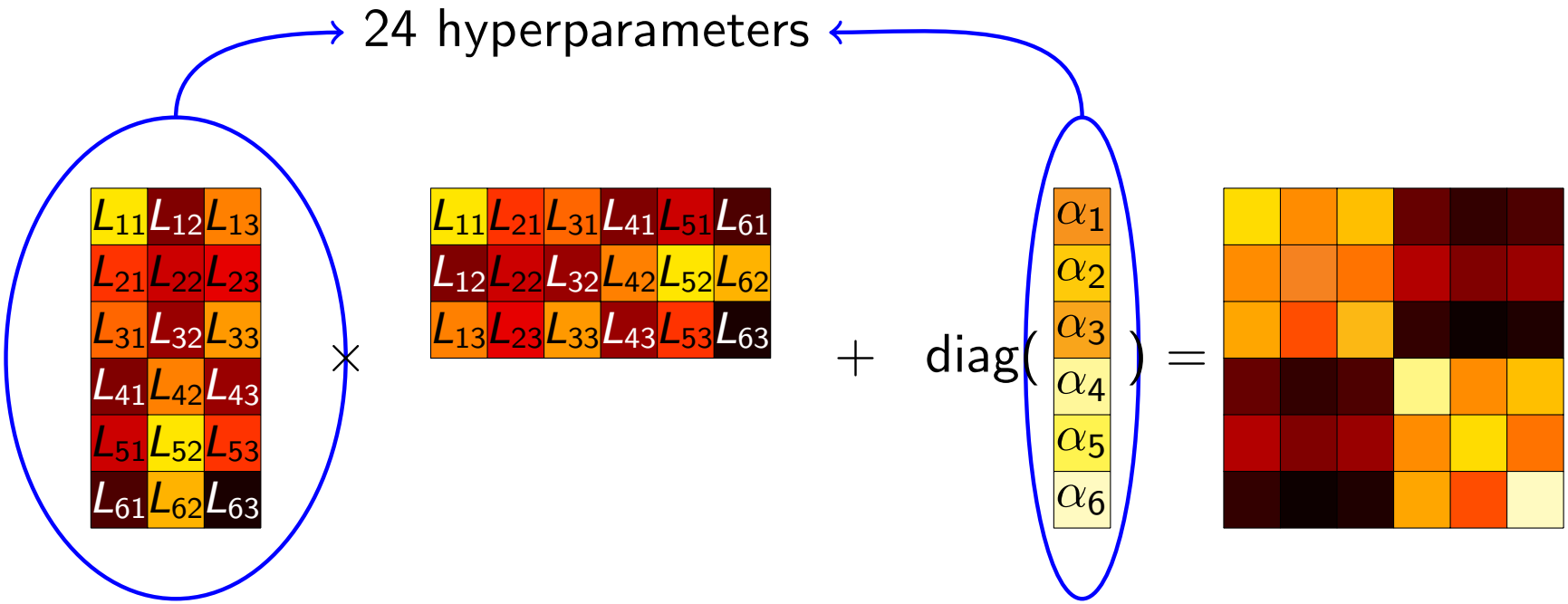
PPCA model


12 hyperparameters
L_{11} L_{21} L_{31} L_{41} L_{51} L_{61}
$\times$
L_{11} L_{21} L_{31} L_{41} L_{51} L_{61}
$+$ diag(α_1 α_2 α_3 α_4 α_5 α_6) $=$
$\tilde{\mathbf{L}}$ $\times$ $\tilde{\mathbf{L}}^T$ $+$ diag($\boldsymbol{\alpha}$) $=$ $\mathbf{B}$

PPCA model


18 hyperparameters
L_{11} L_{12}
L_{21} L_{22}
L_{31} L_{32}
L_{41} L_{42}
L_{51} L_{52}
L_{61} L_{62}
×
L_{11} L_{21} L_{31} L_{41} L_{51} L_{61}
L_{12} L_{22} L_{32} L_{42} L_{52} L_{62}
+ diag(
α_1
α_2
α_3
α_4
α_5
α_6
) =
$\tilde{\mathbf{L}}$ × $\tilde{\mathbf{L}}^T$ + diag($\boldsymbol{\alpha}$) = $\mathbf{B}$

PPCA model


24 hyperparameters
L_{11} L_{12} L_{13}
L_{21} L_{22} L_{23}
L_{31} L_{32} L_{33}
L_{41} L_{42} L_{43}
L_{51} L_{52} L_{53}
L_{61} L_{62} L_{63}
×
L_{11} L_{21} L_{31} L_{41} L_{51} L_{61}
L_{12} L_{22} L_{32} L_{42} L_{52} L_{62}
L_{13} L_{23} L_{33} L_{43} L_{53} L_{63}
+ diag(
α_1
α_2
α_3
α_4
α_5
α_6
) =
$\tilde{\mathbf{L}}$ × $\tilde{\mathbf{L}}^T$ + diag($\boldsymbol{\alpha}$) = $\mathbf{B}$

Experimental Setup

- Dataset: SEMEval2007 "Affective Text" [Strapparava and Mihalcea, 2007];

Experimental Setup

- Dataset: SEMEval2007 "Affective Text" [Strapparava and Mihalcea, 2007];
- 1000 News headlines, each one annotated with 6 scores [0-100], one for emotion;

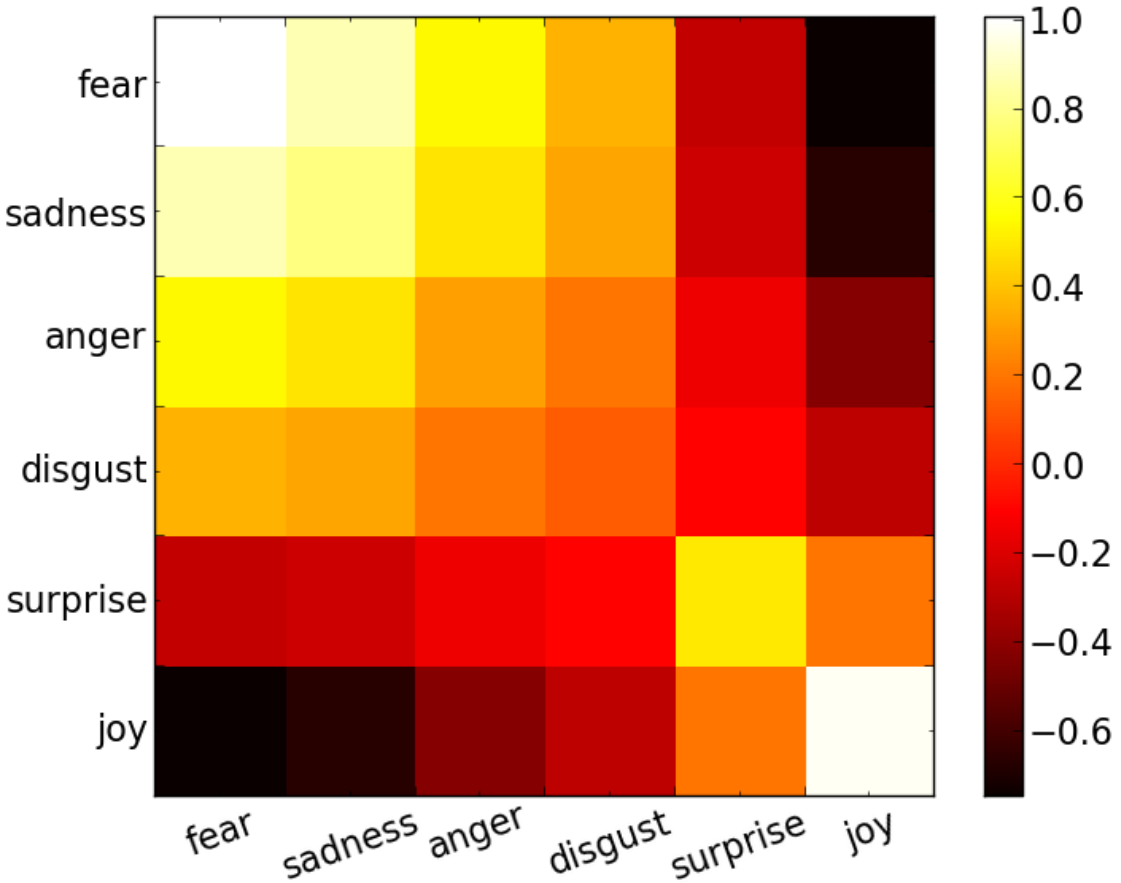
Experimental Setup

- Dataset: SEMEval2007 "Affective Text" [Strapparava and Mihalcea, 2007];
- 1000 News headlines, each one annotated with 6 scores [0-100], one for emotion;
- Bag-of-words representation as features;

Experimental Setup

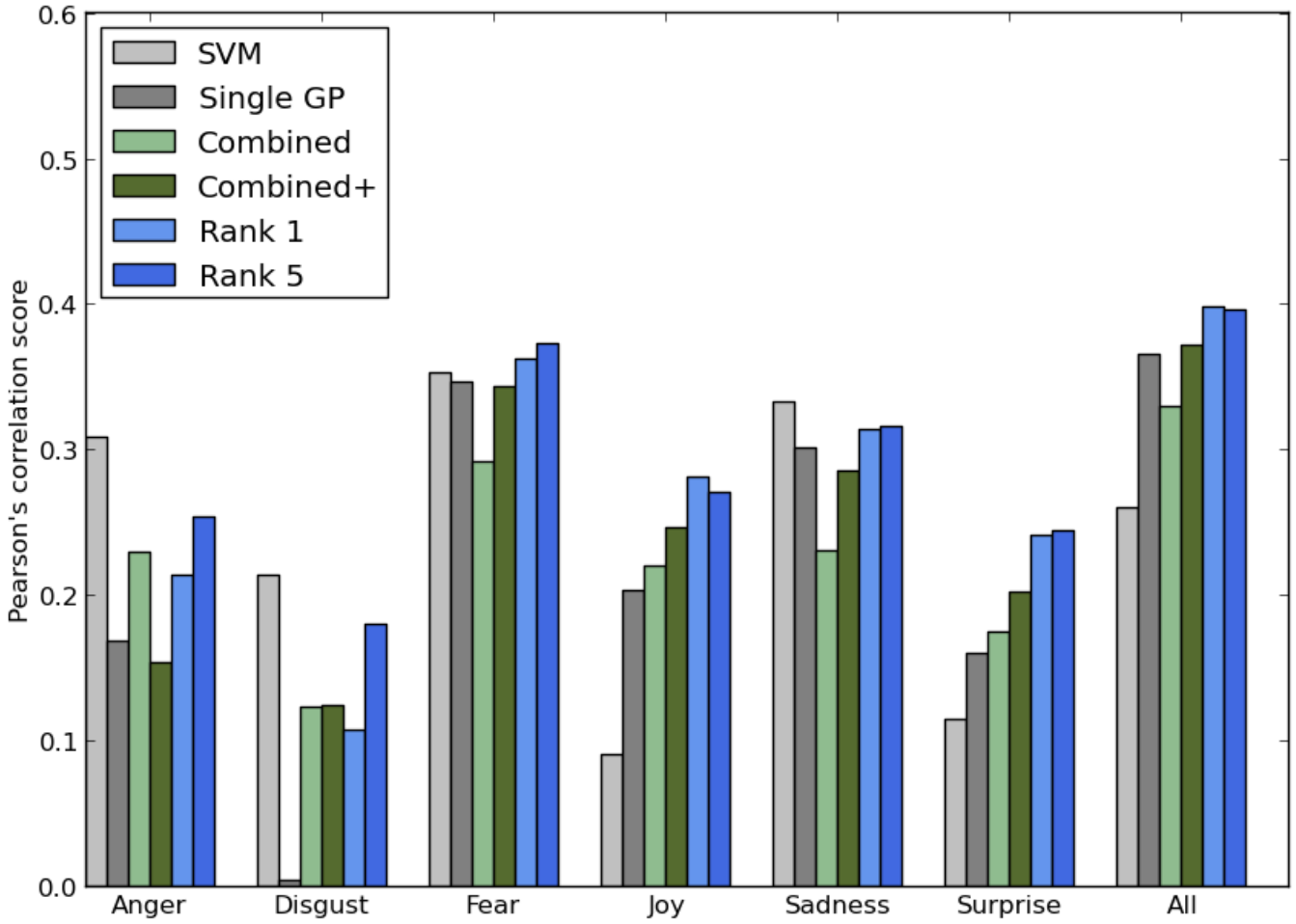
- Dataset: SEMEval2007 "Affective Text" [Strapparava and Mihalcea, 2007];
- 1000 News headlines, each one annotated with 6 scores [0-100], one for emotion;
- Bag-of-words representation as features;
- Pearson's correlation score as evaluation metric;

Learned Task Covariances



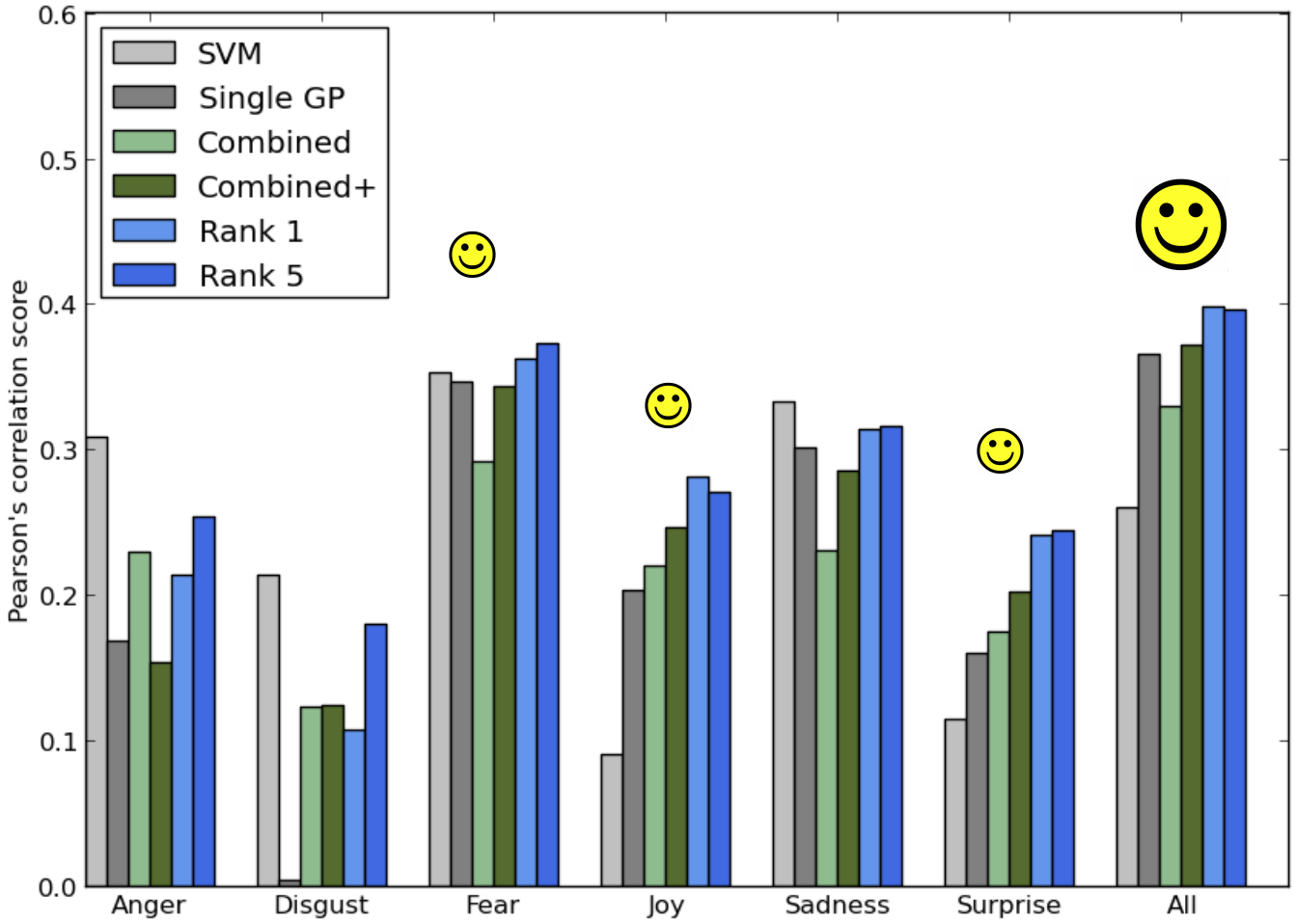
100 sentences

Prediction Results



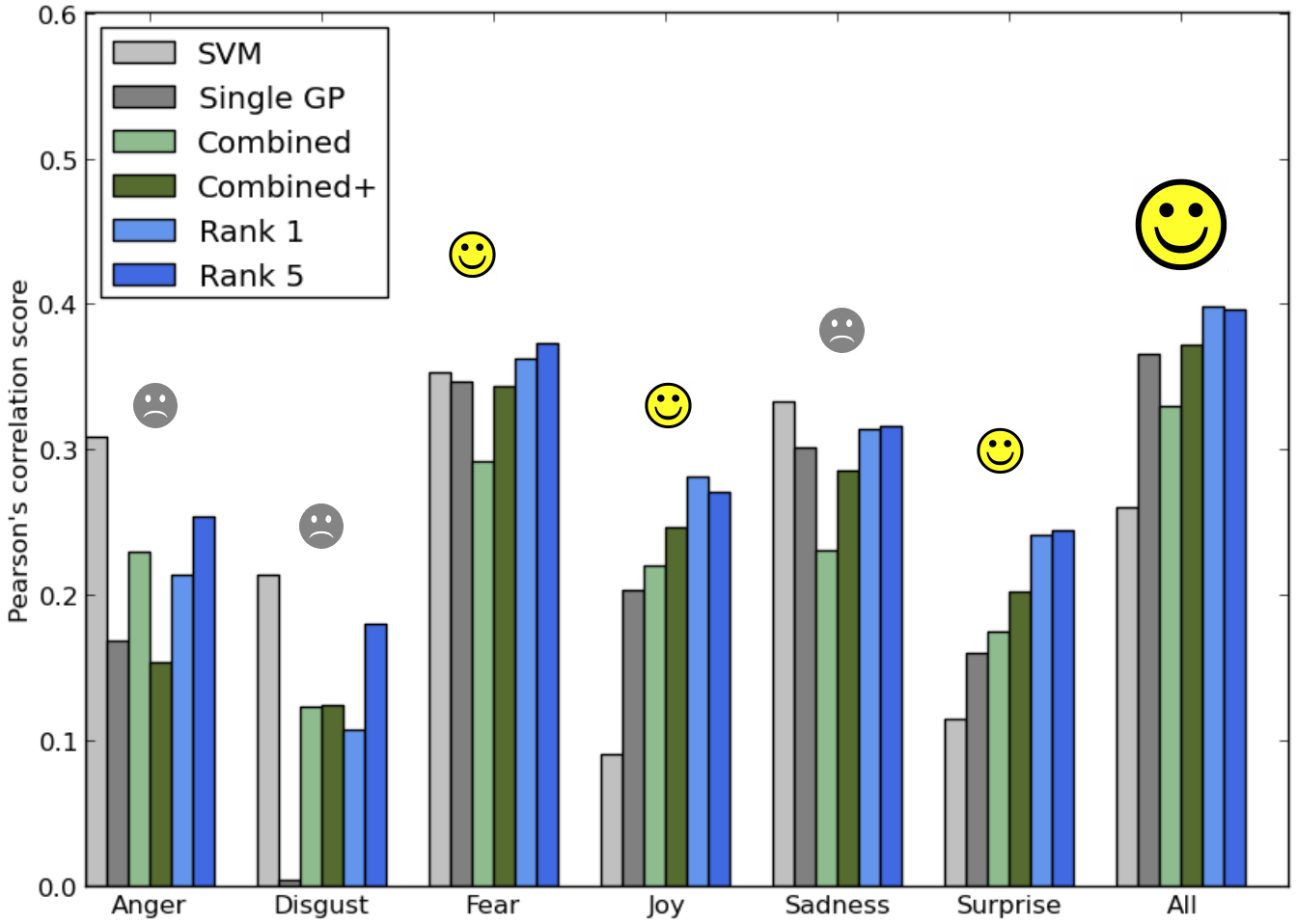
Split: 100/900

Prediction Results



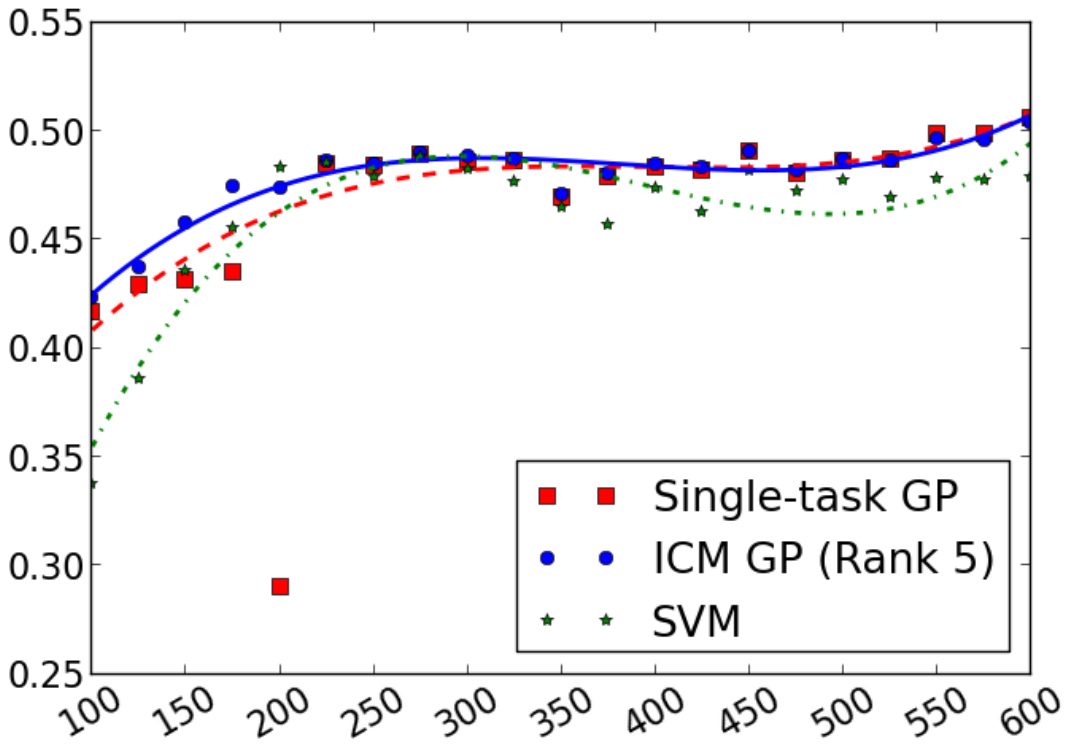
Split: 100/900

Prediction Results



Split: 100/900

Training Set Size Influence



Split: 100+/100

Conclusions

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;
- For small datasets, it also outperforms single-task baselines;

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;
- For small datasets, it also outperforms single-task baselines;

Future Work

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;
- For small datasets, it also outperforms single-task baselines;

Future Work

- Modelling the label distribution (different priors, different likelihoods)

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;
- For small datasets, it also outperforms single-task baselines;

Future Work

- Modelling the label distribution (different priors, different likelihoods)
- Multiple multi-task levels (for example, MTurk data [Snow et al., 2008]);

Conclusions and Future Work

Conclusions

- The proposed model is able to learn sensible correlations and anti-correlations;
- For small datasets, it also outperforms single-task baselines;

Future Work

- Modelling the label distribution (different priors, different likelihoods)
- Multiple multi-task levels (for example, MTurk data [Snow et al., 2008]);
- Other multi-task GP models [Álvarez et al., 2012, Hensman et al., 2013];

The University Of Sheffield.

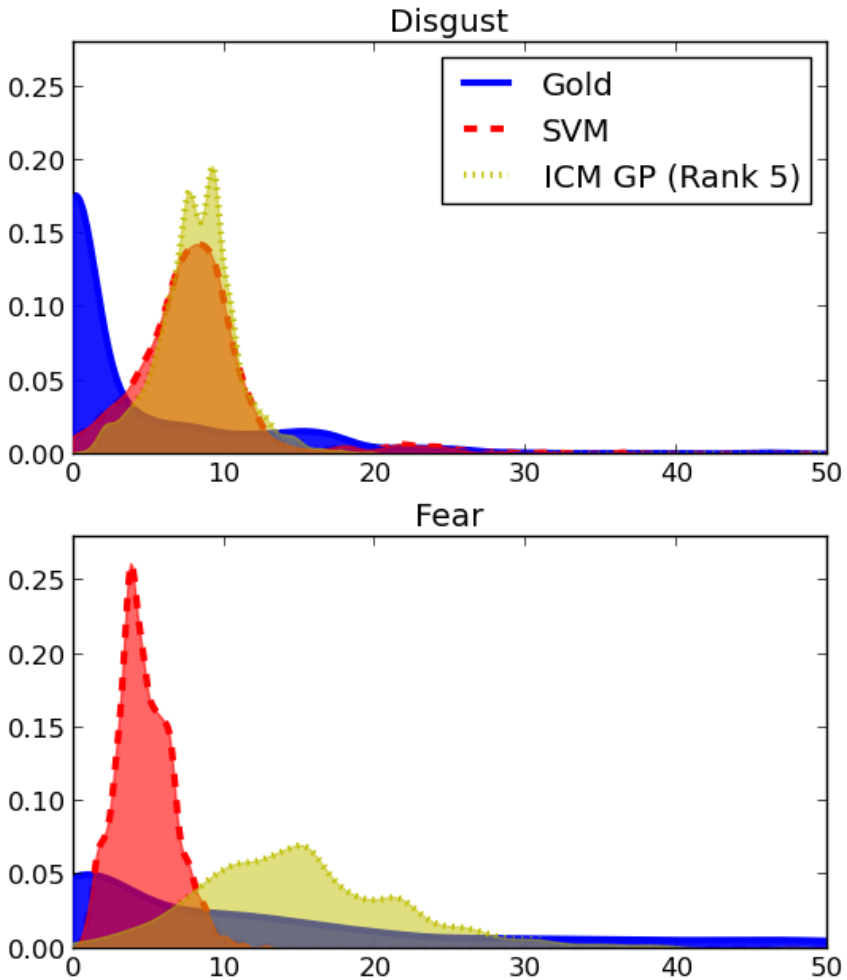
THE UNIVERSITY OF MELBOURNE

Joint Emotion Analysis via Multi-task Gaussian Processes

Daniel Beck, Trevor Cohn, Lucia Specia

October 28, 2014

Error Analysis




Álvarez, M. A., Rosasco, L., and Lawrence, N. D. (2012).
Kernels for Vector-Valued Functions: a Review.
Foundations and Trends in Machine Learning, pages 1–37.


Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I. (2008).
Multi-task Gaussian Process Prediction.
Advances in Neural Information Processing Systems.


Cohn, T. and Specia, L. (2013).
Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation.
In *Proceedings of ACL*.


Hensman, J., Lawrence, N. D., and Rattray, M. (2013).
Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters.
BMC Bioinformatics, 14:252.


Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008).
Cheap and Fast - But is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks.

In *Proceedings of EMNLP*.


Strapparava, C. and Mihalcea, R. (2007).
SemEval-2007 Task 14 : Affective Text.
In *Proceedings of SEMEVAL*.


Strapparava, C. and Mihalcea, R. (2008).
Learning to identify emotions in text.
In *Proceedings of the 2008 ACM Symposium on Applied Computing*.