Bilateral Ordinal Relevance Multi-instance Regression for Facial Action Unit Intensity Estimation Supplementary Material

Yong Zhang[1,2], Rui Zhao[3], Weiming Dong[1], Bao-Gang Hu[1], and Qiang Ji[3]
[1]National Laboratory of Pattern Recognition, CASIA
[2]University of Chinese Academy of Sciences
[3]Rensselaer Polytechnic Institute
zhangyong201303@gmail.com, zhaor@rpi.edu, weiming.dong@ia.ac.cn,
hubg@nlpr.ia.ac.cn, qji@ecse.rpi.edu

1. Optimization for BORMIR

In Section 3.4 of the main script, we present that the original problem can be decomposed into two subproblems. Here, we show the detailed optimization procedures.

1.1. The Proposed BORMIR

In our BORMIR model, we obtain the intensity estimator by solving the following problem

$$\begin{aligned}\min_{\mathbf{w},\{\boldsymbol{\eta}_i,\boldsymbol{\mu}_i\}_{i=1}^N} \; & L(\mathbf{w},\{\boldsymbol{\alpha}_i\}_{i=1}^N,\mathcal{D}) + \lambda_0 L_0(\mathbf{w},\{\boldsymbol{\beta}_i\}_{i=1}^N,\mathcal{D}) \\ & + \lambda_1 R_1(\mathbf{w},\mathcal{D}) + \lambda_2 R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N,\mathcal{D}) \\ & + \lambda_3 R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N,\mathcal{D}) + \frac{\lambda_4}{2}\|\mathbf{w}\|^2 \\ s.t. \;\; & (\boldsymbol{\eta}_i,\boldsymbol{\mu}_i) \in S^{\eta,\mu}(\boldsymbol{\eta}_i,\boldsymbol{\mu}_i), i = 1,2,...,N, \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha}_i = \mathbf{A}_i\boldsymbol{\eta}_i$, $\boldsymbol{\beta}_i = \mathbf{A}_i^T\boldsymbol{\mu}_i$, and $\lambda_k \geq 0, k = 0,1,2,3,4$, are the hyperparameters. Items in problem (1), including the peak bag label, the valley bag label, the ordinal relevance, the smoothness of intensity, the smoothness of relevance, are listed as follows

$$\begin{aligned} L(\mathbf{w},\{\boldsymbol{\alpha}_i\}_{i=1}^N,\mathcal{D}) &= \frac{1}{2}\sum_{i=1}^N (y_i - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\alpha}_i)^2, \\ L_0(\mathbf{w},\{\boldsymbol{\beta}_i\}_{i=1}^N,\mathcal{D}) &= \frac{1}{2}\sum_{i=1}^N (y_i^0 - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\beta}_i)^2, \\ R_1(\mathbf{w},\mathcal{D}) &= \frac{1}{2}\mathbf{w}^T\mathbf{L}\mathbf{w}, \;\; \mathbf{L} = \sum_{i=1}^N \mathbf{B}_i(\mathbf{D}_i - \mathbf{C}_i)\mathbf{B}_i^T, \\ R_2(\{\boldsymbol{\alpha}_i\}_{i=1}^N,\mathcal{D}) &= \frac{1}{2}\sum_{i=1}^N \boldsymbol{\alpha}_i^T(\mathbf{D}_i - \mathbf{C}_i)\boldsymbol{\alpha}_i, \\ R_2(\{\boldsymbol{\beta}_i\}_{i=1}^N,\mathcal{D}) &= \frac{1}{2}\sum_{i=1}^N \boldsymbol{\beta}_i^T(\mathbf{D}_i - \mathbf{C}_i)\boldsymbol{\beta}_i, \end{aligned} \tag{2}$$

and

$$\begin{aligned} S^{\eta,\mu}(\boldsymbol{\eta}_i, \boldsymbol{\mu}_i) = &\{\boldsymbol{\eta}_i \in \mathbb{R}^{n_i}, \boldsymbol{\mu}_i \in \mathbb{R}^{n_i} | \boldsymbol{\eta}_i \geq \mathbf{0}, \boldsymbol{\mu}_i \geq \mathbf{0}, \\ &\mathbf{e}_i^T(\mathbf{A}_i\boldsymbol{\eta}_i) = 1, \mathbf{e}_i^T(\mathbf{A}_i^T\boldsymbol{\mu}_i) = 1, \\ &\mathbf{V}_i(\mathbf{A}_i\boldsymbol{\eta}_i + \mathbf{A}_i^T\boldsymbol{\mu}_i) = \mathbf{0}\}. \end{aligned}$$

1.2. The Optimization of BORMIR

Let $\boldsymbol{\theta}_i = [\boldsymbol{\eta}_i; \boldsymbol{\mu}_i] \in \mathbb{R}^{2n_i}$ by concatenating $\boldsymbol{\eta}_i$ and $\boldsymbol{\mu}_i$. To solve problem (1), we develop an iterative optimization algorithm under the alternating minimization framework [2].

Optimize $\mathbf{w}$, given $\{\boldsymbol{\theta}_i\}_{i=1}^N$ Given $\{\boldsymbol{\theta}_i\}_{i=1}^N$, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ can be computed through $\boldsymbol{\alpha}_i = \mathbf{A}_i\boldsymbol{\eta}_i$ and $\boldsymbol{\beta}_i = \mathbf{A}_i^T\boldsymbol{\mu}_i$. Problem (1) becomes an unconstrained problem with respect to $\mathbf{w}$. The subproblem is

$$\begin{aligned} \min_{\mathbf{w}} \quad & L(\mathbf{w}, \{\boldsymbol{\alpha}_i\}_{i=1}^N, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \{\boldsymbol{\beta}_i\}_{i=1}^N, \mathcal{D}) + \lambda_1 R_1(\mathbf{w}, \mathcal{D}) + \frac{\lambda_4}{2}\|\mathbf{w}\|^2 \\ &= \frac{1}{2}(\mathbf{Y} - \mathbf{X}^T\mathbf{w})^T(\mathbf{Y} - \mathbf{X}^T\mathbf{w}) + \frac{\lambda_0}{2}(\mathbf{Y}_0 - \tilde{\mathbf{X}}^T\mathbf{w})^T(\mathbf{Y}_0 - \tilde{\mathbf{X}}^T\mathbf{w}) + \frac{\lambda_1}{2}\mathbf{w}^T\mathbf{L}\mathbf{w} + \frac{\lambda_4}{2}\mathbf{w}^T\mathbf{w} \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{H}\mathbf{w} + \mathbf{f}^T\mathbf{w}, \end{aligned} \tag{3}$$

where $\mathbf{X} = [\mathbf{B}_1\boldsymbol{\alpha}_1, \mathbf{B}_2\boldsymbol{\alpha}_2, ..., \mathbf{B}_N\boldsymbol{\alpha}_N]$, $\tilde{\mathbf{X}} = [\mathbf{B}_1\boldsymbol{\beta}_1, \mathbf{B}_2\boldsymbol{\beta}_2, ..., \mathbf{B}_N\boldsymbol{\beta}_N]$, $\mathbf{H} = \mathbf{X}\mathbf{X}^T + \lambda_0\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda_1\mathbf{L} + \lambda_4\mathbf{I}$ and $\mathbf{f} = -\mathbf{X}\mathbf{Y} - \lambda_0\tilde{\mathbf{X}}\mathbf{Y}_0$. $\mathbf{Y} = [y_1, y_2, ..., y_N]^T$ and $\mathbf{Y}_0 = [y_1^0, y_2^0, ..., y_N^0]^T$ are the peak and valley bag label vectors of the N training segments. $\mathbf{I}$ is an identity matrix and $\mathbf{H}$ is a positive semi-definite matrix. The subproblem is a standard unconstrained quadratic programming problem. The closed-form solution is

$$\begin{aligned} \mathbf{w}^* &= -\mathbf{H}^{-1}\mathbf{f} \\ &= [\mathbf{X}\mathbf{X}^T + \lambda_0\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \lambda_1\mathbf{L} + \lambda_4\mathbf{I}]^{-1}(\mathbf{X}\mathbf{Y} + \lambda_0\tilde{\mathbf{X}}\mathbf{Y}_0) \end{aligned} \tag{4}$$

Optimize $\{\boldsymbol{\theta}_i\}_{i=1}^N$, given $\mathbf{w}$ Given $\mathbf{w}$, problem (1) can be decomposed into independent subproblems with respect to each $\boldsymbol{\theta}_i$. Each problem is a quadratic programming problem with linear constraints. Let $\mathbf{F}_i = \mathbf{D}_i - \mathbf{C}_i$. The objective is

$$\begin{aligned} & L(\mathbf{w}, \boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \boldsymbol{\beta}_i, \mathcal{D}) + \lambda_2 R_2(\boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_3 R_2(\boldsymbol{\beta}_i, \mathcal{D}) \\ = \ & \frac{1}{2}(y_i - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\alpha}_i)^2 + \frac{\lambda_0}{2}(y_i^0 - \mathbf{w}^T\mathbf{B}_i\boldsymbol{\beta}_i)^2 + \frac{1}{2}\boldsymbol{\alpha}_i^T\mathbf{F}_i\boldsymbol{\alpha}_i + \frac{1}{2}\boldsymbol{\beta}_i^T\mathbf{F}_i\boldsymbol{\beta}_i \\ = \ & \frac{1}{2}(y_i - \mathbf{w}^T\mathbf{B}_i\mathbf{A}_i\boldsymbol{\eta}_i)^2 + \frac{\lambda_0}{2}(y_i^0 - \mathbf{w}^T\mathbf{B}_i\mathbf{A}_i^T\boldsymbol{\mu}_i)^2 + \frac{\lambda_2}{2}(\mathbf{A}_i\boldsymbol{\eta}_i)^T\mathbf{F}_i(\mathbf{A}_i\boldsymbol{\eta}_i) + \frac{\lambda_3}{2}(\mathbf{A}_i^T\boldsymbol{\mu}_i)^T\mathbf{F}_i(\mathbf{A}_i^T\boldsymbol{\mu}_i). \end{aligned}$$

Let $\mathbf{M}_i = \mathbf{w}^T\mathbf{B}_i\mathbf{A}_i$ and $\mathbf{N}_i = \mathbf{w}^T\mathbf{B}_i\mathbf{A}_i^T$. It becomes

$$\begin{aligned} & L(\mathbf{w}, \boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_0 L_0(\mathbf{w}, \boldsymbol{\beta}_i, \mathcal{D}) + \lambda_2 R_2(\boldsymbol{\alpha}_i, \mathcal{D}) + \lambda_3 R_2(\boldsymbol{\beta}_i, \mathcal{D}) \\ = \ & \frac{1}{2}\boldsymbol{\eta}_i^T(\mathbf{M}_i^T\mathbf{M}_i + \lambda_2\mathbf{A}_i^T\mathbf{F}_i\mathbf{A}_i)\boldsymbol{\eta}_i + \frac{1}{2}\boldsymbol{\mu}_i^T(\lambda_0\mathbf{N}_i^T\mathbf{N}_i + \lambda_3\mathbf{A}_i\mathbf{F}_i\mathbf{A}_i^T)\boldsymbol{\mu}_i - y_i\mathbf{M}_i\boldsymbol{\eta}_i - \lambda_0 y_i^0\mathbf{N}_i\boldsymbol{\mu}_i + const \\ = \ & \frac{1}{2}\boldsymbol{\theta}_i^T\mathbf{H}\boldsymbol{\theta}_i + \mathbf{f}^T\boldsymbol{\theta}_i + const, \end{aligned} \tag{5}$$

where

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^T\mathbf{M} + \lambda_2\mathbf{A}_i^T\mathbf{F}_i\mathbf{A}_i & \mathbf{0} \\ \mathbf{0} & \lambda_0\mathbf{N}_i^T\mathbf{N}_i + \lambda_3\mathbf{A}_i\mathbf{F}_i\mathbf{A}_i^T \end{bmatrix}, \ \mathbf{f} = \begin{bmatrix} -y_i\mathbf{M}_i^T \\ -\lambda_0 y_i^0\mathbf{N}_i^T \end{bmatrix}.$$

The linear constraints can be rewritten as

$$S^\theta(\boldsymbol{\theta}_i) = \{\boldsymbol{\theta}_i \in \mathbb{R}^{2n_i} | \boldsymbol{\theta}_i \geq \mathbf{0}, \begin{bmatrix} \mathbf{e}_i^T\mathbf{A}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_i^T\mathbf{A}_i^T \\ \mathbf{V}_i\mathbf{A}_i & \mathbf{V}_i\mathbf{A}_i^T \end{bmatrix} \boldsymbol{\theta}_i = \begin{bmatrix} 1 \\ 1 \\ \mathbf{0} \end{bmatrix}\}.$$

The subproblem becomes

$$\min_{\theta_i} \ \frac{1}{2}\boldsymbol{\theta}_i^T \mathbf{H}\boldsymbol{\theta}_i + \mathbf{f}^T\boldsymbol{\theta}_i$$
$$s.t. \ \ \boldsymbol{\theta}_i \in S^{\theta}(\boldsymbol{\theta}_i) \tag{6}$$

The subproblem is a standard quadratic programming problem with linear constraints, which can be solved efficiently with existing solvers. In this paper, we adopt the interior-point algorithm [8] to solve problem (6).

2. Relevance analysis

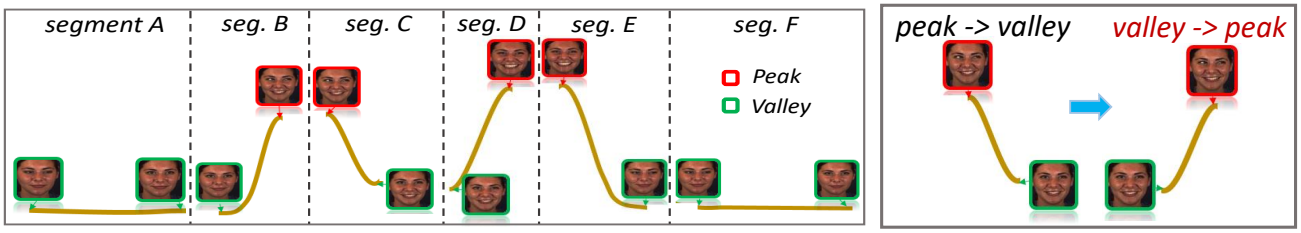
The relevance is learned only for training frames and is not involved during testing. Hence, we analyze relevance on the training set of FERA 2015. We studied (i) the correlation between relevance and the ground truth intensity and (ii) the correlation between relevance and the output of a simple classifier. We use logistic regression model as the classifier which is trained to tell peak frames apart from valley frames. PCC is used as the measure. Rel: the 'peak relevance', MO: the intensity prediction by our model, CO: the output probability of the classifier, GT: the ground truth. As shown in Table 1, Rel has the correlation with GT and CO to some extent, but is not strongly correlated. Because the relevance is the importance of frames in a segment, which is not equivalent to the intensity label. However, MO and CO are highly correlated with GT since they are directly associated with the intensity.

Table 1. Correlation analysis on FERA 2015. PCC is the measure.

AU	6	10	12	14	17
Rel & GT	0.28	0.26	0.39	0.34	0.36
Rel & CO	0.26	0.26	0.37	0.18	0.31
MO & GT	0.84	0.81	0.92	0.64	0.67
CO & GT	0.62	0.61	0.69	0.47	0.58

3. Splitting of Sequences and Arrangement of frame order

In Section 3 of the main script, we describe the splitting of sequences. Here, we give a detailed illustration. We use the definition of 'peak' and 'valley' from [6] which differs from onset/apex/offset. Given peak/valley frames, Fig. 1(a) gives an illustration of splitting a sequence into segments. We have three types of segments: (i) from valley to peak, (ii) from peak to valley and (iii) constant intensity. **We rearrange only the frame order of segments of type (ii)** as shown in Fig. 1(b).



(a) (b)

Figure 1. (a) Segment splitting. (b) Order arrangement.

4. AU intensity estimation on the PAIN database

In Section 4.2 of the main script, to make the comparison to MI-DORF, we perform pain intensity estimation on the PAIN database. We also present the performance of different methods for AU intensity estimation on PAIN. The results are shown in Table 2. Our method outperforms the competing methods in PCC and ICC. The MAE of our method is the second best. Though LT is slightly better than our method in MAE, it performs poor in PCC and ICC.

Table 2. Comparison of different methods for AU intensity estimation on the PAIN database. Bracketed and bold numbers represent the best performance; bold numbers represent the second best.

	AU	4	6	7	9	10	12	20	25	26	Avg
PCC	SOVR [1]	**.640**	.521	.241	.137	**.535**	.561	.179	**.521**	-.054	.365
	RVR [3]	[**.655**]	**.559**	**.379**	.188	[**.570**]	[**.595**]	[**.334**]	[**.538**]	-.063	**.417**
	LT [4]	-.071	.173	.068	.000	.000	-.013	.045	.056	**.214**	.052
	DSRVM [5]	-.012	.446	.314	.074	.495	.359	-.099	.213	-.119	.186
	MIR [7]	.357	.479	.247	.035	.012	.484	.169	.490	-.004	.252
	OSVR [9]	.573	.541	.247	.327	.527	**.593**	.212	.503	-.114	.379
	BORMIR	.582	[**.564**]	**.490**	[**.495**]	.435	.585	**.278**	.507	**.118**	[**.450**]
ICC	SOVR [1]	**.522**	.508	.212	.125	**.437**	.554	.137	[**.511**]	-.043	.329
	RVR [3]	[**.555**]	[**.538**]	**.315**	.188	[**.447**]	**.556**	[**.236**]	**.499**	-.041	**.366**
	LT [4]	.043	.075	.005	-.039	-.005	.040	.039	.011	**.074**	.027
	DSRVM [5]	-.013	.400	.130	-.013	.133	.249	-.022	.186	-.017	.115
	MIR [7]	.277	.447	.196	.052	.015	.458	.105	.457	-.017	.221
	OSVR [9]	.470	**.529**	.236	**.274**	.413	[**.578**]	.181	.481	-.107	.340
	BORMIR	.445	.501	[**.407**]	[**.466**]	.315	.535	**.211**	.479	[**.097**]	[**.384**]
MAE	SOVR [1]	2.00	1.12	2.54	2.61	1.76	1.03	2.43	**1.03**	2.22	1.86
	RVR [3]	**1.44**	**1.00**	1.60	1.83	1.38	0.96	1.36	[**0.91**]	1.29	1.31
	LT [4]	[**1.43**]	1.15	[**1.19**]	[**1.38**]	**1.31**	1.12	[**1.01**]	1.20	[**1.00**]	[**1.20**]
	DSRVM [5]	1.86	1.02	**1.22**	1.84	[**1.07**]	1.06	1.56	1.06	1.39	1.34
	MIR [7]	2.80	1.11	3.06	5.33	5.25	**0.94**	3.04	1.30	2.66	2.83
	OSVR [9]	1.87	1.05	2.14	2.03	2.00	1.00	2.13	1.05	1.89	1.68
	BORMIR	1.47	[**0.94**]	1.31	**1.66**	1.64	[**0.93**]	**1.04**	1.09	**1.06**	**1.24**

References

[1] W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *ICML*, 2005. 4

[2] I. Csisz, G. Tusnády, et al. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1984. 2

[3] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *ISVC*, 2012. 4

[4] S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *CVPR*, 2015. 4

[5] S. Kaltwang, S. Todorovic, and M. Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *TPAMI*, 2016. 4

[6] M. Mavadati, P. Sanger, and M. H. Mahoor. Extended disfa dataset: Investigating posed and spontaneous facial expressions. In *CVPRW*, 2016. 3

[7] K. L. Wagstaff and T. Lane. *Salience assignment for multiple-instance regression*. Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2007. 4

[8] S. J. Wright. *Primal-dual interior-point methods*. SIAM, 1997. 3

[9] R. Zhao, Q. Gan, S. Wang, and Q. Ji. Facial expression intensity estimation using ordinal information. In *CVPR*, 2016. 4