Convex Relaxation of Mixture Regression with Efficient Algorithms

Authors

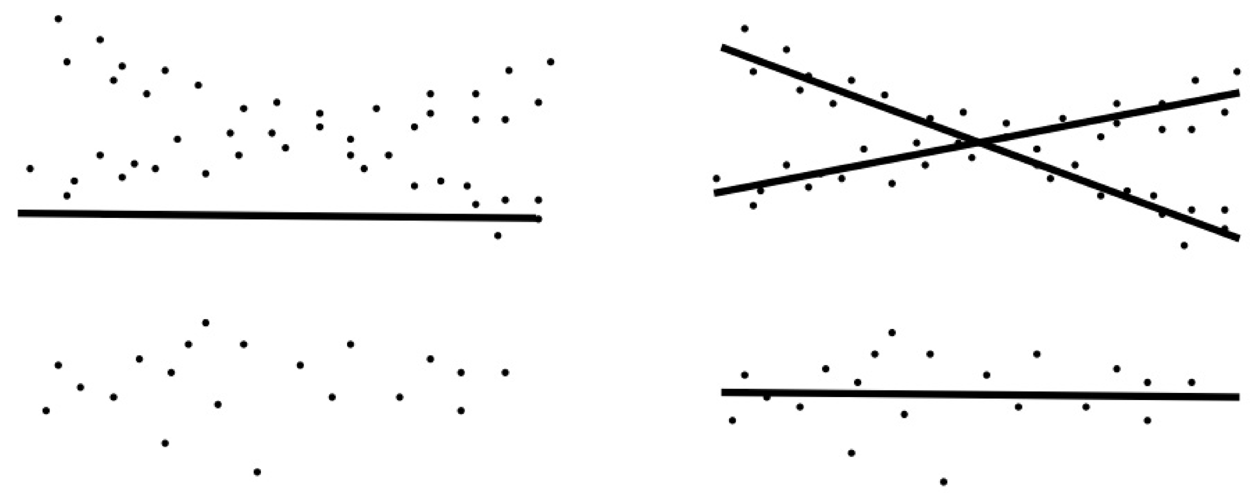
Novi Quadrianto[1] | Tibério S. Caetano[1] | John Lim[1] | Dale Schuurmans[2]

1: NICTA & Australian National University | 2: University of Alberta

Abstract

- We develop a convex relaxation of maximum a posteriori estimation of a mixture of regression model.
- We reformulate the relaxation problem to eliminate the need for general semidefinite programming.
- We provide two reformulations that admit fast algorithms. The first is a max-min spectral reformulation exploiting quasi-Newton descent. The second is a min-min reformulation consisting of fast alternating steps of closed-form updates.
- We provide experiments in a real problem of motion segmentation from video data.

Mixture of Regression Problem



Linear regression (left) and mixture of regression (right); Gaffney & Smyth 1999

Given

- a labeled training set $\mathcal{D}$ comprising t input-output pair $\{(x_1, y_1), \ldots, (x_t, y_t)\}$;
- assumption that the output variable y is generated by a mixture of k components;
- no information about which component of the mixture generates each output variable y_i.

Find

- a regression model $f : \mathcal{X} \to \mathcal{Y}$ for each of the mixture component.

The Model

Denote

- $X \in \mathbb{R}^{t \times n}$ as a matrix of input and $y \in \mathbb{R}^{t \times 1}$ as a vector of output variables;
- $\Pi \in \{0, 1\}^{t \times k}, \Pi\mathbf{1} = \mathbf{1}$ and $\max(\mathrm{diag}(\Pi^T \Pi)) \leq \gamma t$ (bounding the size of the largest component) as the hidden assignment matrix.

Assume

- a linear regression model $y_i | x_i, \pi_i = \psi_i w + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$ w.r.t a feature representation $\psi_i = \pi_i \otimes x_i$.

Gaussian Likelihood

With the Gaussian noise model , our likelihood is then

$$p(y_i | x_i, \pi_i; w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\psi_i w - y_i)^2\right].$$

Log-Posterior Optimization

With an additional assumption of Gaussian prior on the parameter w, the minimization of (negative) log-posterior is now in the form of

$$\min_{\Pi, w} \left[\frac{1}{2\sigma^2} w^T \Psi^T \Psi w - \frac{1}{\sigma^2} y^T \Psi w + \frac{\alpha}{2} w^T w\right], \tag{1}$$

s.t. constraints on the assignment matrix Π.

Semidefinite Relaxation

Problem (1) can be relaxed to

$$\min_{M: \mathrm{tr}\, M = t, \gamma t I \succcurlyeq M \succcurlyeq 0} \max_{c} \left[-\frac{1}{2}\sigma^2 c^T c - \frac{1}{2\alpha}\left(\frac{y}{\sigma^2} - c\right)^T M \odot XX^T \left(\frac{y}{\sigma^2} - c\right)\right]. \tag{2}$$

(Sketchy) Steps

- Compute convex conjugate of the log-partition function (i.e. $\frac{1}{2\sigma^2} w^T \Psi^T \Psi w$).
- Relax the constraints on the assignment matrix

$$\begin{aligned} &\{\Pi\Pi^T : \Pi \in \{0,1\}^{t \times k}, \Pi\mathbf{1} = \mathbf{1}, \max(\mathrm{diag}(\Pi^T \Pi)) \leq \gamma t\} \\ \subseteq \; &\{M : M \in \mathbb{R}^{t \times t}, \mathrm{tr}\, M = t, \gamma t I \succcurlyeq M \succcurlyeq 0\}. \end{aligned}$$

Note : The relaxation of Problem (1) to (2) only refers to loosening the set of feasible solutions and no relaxation has been introduced in the objective function. However, solving (2) directly is prohibitive for medium to large scale problems as it requires a general semidefinite solver.

Algorithm 1: Max-Min Reformulation

The "hammer" (Overton & Womersley 1993)

Let $V \in \mathbb{R}^{t \times t}, V = V^T$ and its EVD, $P^T V P = \Lambda(\langle\lambda_1, \ldots, \lambda_t\rangle)$, then

$$\max_{M: \mathrm{tr}(M) = q, I \succcurlyeq M \succcurlyeq 0} \mathrm{tr}\, MV^T = \sum_{i=1}^{q} \lambda_i \quad \text{and} \quad \operatorname*{argmax}_{M: \mathrm{tr}(M) = q, I \succcurlyeq M \succcurlyeq 0} \mathrm{tr}\, MV^T \ni P_q P_q^T.$$

To use the "hammer", Problem (2) is reformulated to

$$\max_{c} \left[-\frac{1}{2}\sigma^2 c^T c - \frac{t}{2\alpha q} \max_{\bar{M}: \mathrm{tr}\, \bar{M} = q, I \succcurlyeq \bar{M} \succcurlyeq 0} \mathrm{tr}(\bar{M}(XX^T \odot (\tilde{y} - c)(\tilde{y} - c)^T))\right].$$

(Sketchy) Steps

- Simply by interchanging $\min_M$ and $\max_c$, invoking distributivity property, rearranging the terms and defining $\tilde{y} := \frac{y}{\sigma^2}$, problem (2) can be rewritten as

$$\max_{c} \left[-\frac{1}{2}\sigma^2 c^T c - \frac{1}{2\alpha} \max_{M: \mathrm{tr}\, M = t, \gamma t I \succcurlyeq M \succcurlyeq 0} \mathrm{tr}(M(XX^T \odot (\tilde{y} - c)(\tilde{y} - c)^T))\right].$$

- Let $q = \{u : u = \max\{1, \ldots, t\}, u \leq \gamma^{-1}\}$ and define $\bar{M} := (q/t)M$ to transform the constraint set from $\{M : \mathrm{tr}\, M = t, \gamma t I \succcurlyeq M \succcurlyeq 0\}$ to $\{M : \mathrm{tr}\, M = q, I \succcurlyeq M \succcurlyeq 0\}$.

Algorithm 2: Min-Min Reformulation

Problem (2) is also equivalent to

$$\min_{\{M: I \geq M \geq 0, \mathrm{tr}\, M = 1/\gamma\}} \min_{A} \left[\frac{1}{\sigma^2} y^T \mathrm{diag}(XA^T) + \frac{1}{2\sigma^2} \mathrm{diag}(XA^T)^T \mathrm{diag}(XA^T) + \frac{\alpha}{2\gamma t} \mathrm{tr}(A^T M^{-1} A)\right].$$

Steps; $$(2) = \min_{\{M: I \geq M \geq 0, \mathrm{tr}\, M = 1/\gamma\}} \max_{\{c, C: C = \Lambda(c - \tilde{y})X\}} -\frac{\sigma^2}{2} c^T c - \frac{\gamma t}{2\alpha} \mathrm{tr}(C^T MC)$$

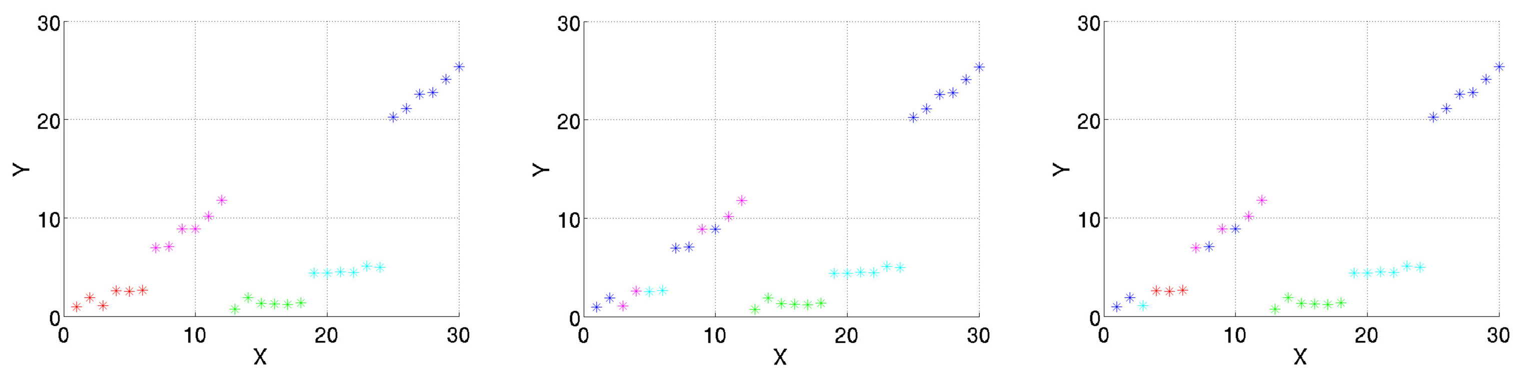
$$= \min_{\{M: I \geq M \geq 0, \mathrm{tr}\, M = 1/\gamma\}} \min_{A} \max_{c, C} -\frac{\sigma^2}{2} c^T c - \frac{\gamma t}{2\alpha} \mathrm{tr}(C^T MC) + \mathrm{tr}(A^T C) - \mathrm{tr}(A^T \Lambda(c - \tilde{y})X).$$

Lastly, c and C can be solved as $c = -\frac{1}{\sigma^2} \mathrm{diag}(XA^T)$ and $C = \frac{\alpha}{\gamma t} M^{-1} A$.

Applications

Toy Data

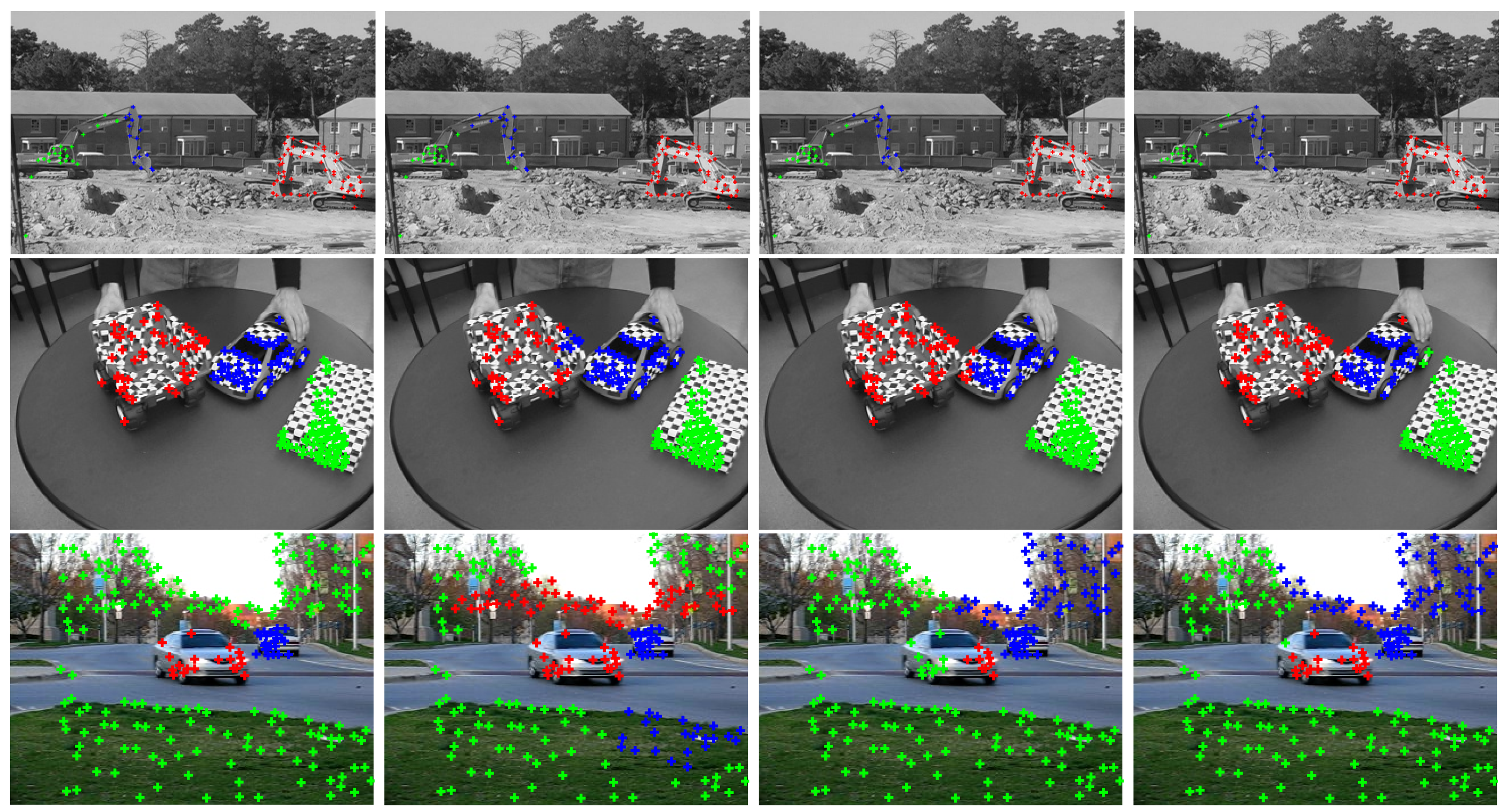
- Dataset: 30 synthetic data points is generated according to $y_i = (\pi_i \otimes x_i)w + \epsilon_i$, with $x_i \in \mathbb{R}$, $\epsilon_i \sim N(0, 1)$ and $w \in U(0, 1)$. The response variable y_i is assumed to be generated from a mixture of 5 components .
- Performance comparison with EM (100 random restarts was used to avoid poor local optima).
- Error rates are 0.347 ± 0.086 (EM) and 0.280 ± 0.063 (Max-Min) across 10 different runs.



Ground Truth EM Max-Min

Motion Segmentation from Video Data

- Dataset: Hopkins 155 (`http://www.vision.jhu.edu/data/hopkins155/`).
- Given a pair of corresponding points p_i and q_i from two frames and k motion groups, we have the following epipolar equation (Li 2007), $q_i^T (\sum_{j=1}^{k} \pi_{ij} F_j) p_i = 0$.
- Redefine $[q_i^x p_i^x \; q_i^x p_i^y \; q_i^x p_i^\pi \; \ldots \; q_i^\pi p_i^\pi]^T := x_i$ and $\mathrm{vec}(F_j^T) := w$, our problem is now $\sum_{j=1}^{k} \pi_{ij} x_i^T w_j = 0$.



Ground Truth EM Max-Min Min-Min

References

- S. Gaffney and P. Smyth. Trajectory clustering with mixtures of regression models. In *ACM SIGKDD*, volume 62, pages 63–72, 1999.
- M. Overton and R. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62:321–357, 1993.
- H. Li. Two-view motion segmentation from linear programming relaxation. In *CVPR*, 2007.