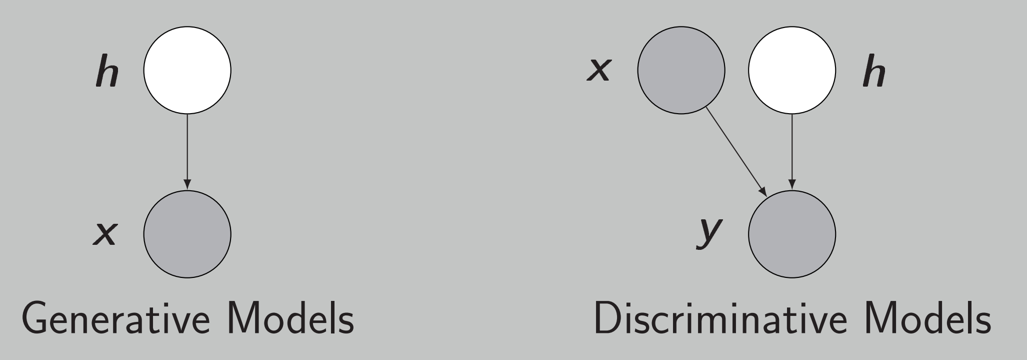
Spectral Experts for Estimating Mixtures of Linear Regressions

Arun Tejasvi Chaganty   Percy Liang

Department of Computer Science
Stanford University

Parameter Estimation in Latent Variable Models



- Latent variable models (LVMs) are hard to learn because latent variables introduce non-convexities in the log-likelihood function.
- In practice, local methods (EM, gradient descent, etc.) are employed, but these can stuck in local optima.
- **Can we develop efficient consistent estimators for discriminative latent variable models?**
  - Why discriminative LVMs? Easy to add features, often more accurate.
  - The method of moments has been used for consistent parameter estimation in several generative LVMs, e.g. HMMs[1], LDA[1], and stochastic block models[2].
  - Can we extend these techniques to discriminative LVMs?
- **Main result**: Consistent estimator for a simple discriminative model; the mixture of linear regressions.
  - **Key Idea:** Expose tensor factorization structure using regression.
  - **Theory:** We prove polynomial sample and computational complexity.

[1] Anandkumar, Hsu, Kakade, 2012; [2] Anandkumar, Ge, Hsu, Kakade, 2012

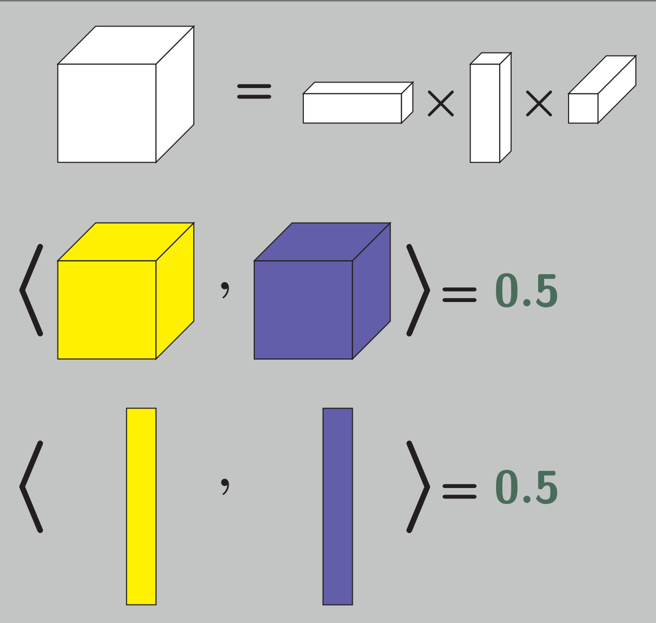
Aside: Tensor Operations

- Tensor Product

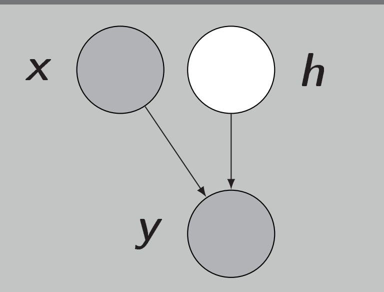
$$x^{\otimes 3} = x \otimes x \otimes x$$
$$x^{\otimes 3}_{ijk} = x_i x_j x_k$$

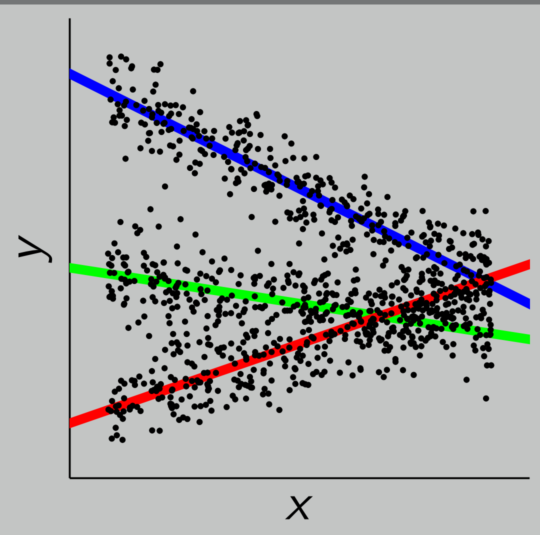
- Inner product

$$\langle A, B \rangle = \sum_{ijk} A_{ijk} B_{ijk} = \langle \mathbf{vec}\, A, \mathbf{vec}\, B \rangle$$

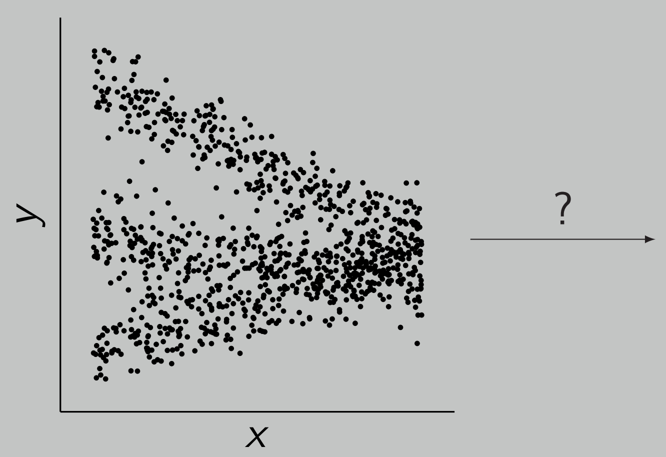


Mixture of Linear Regressions



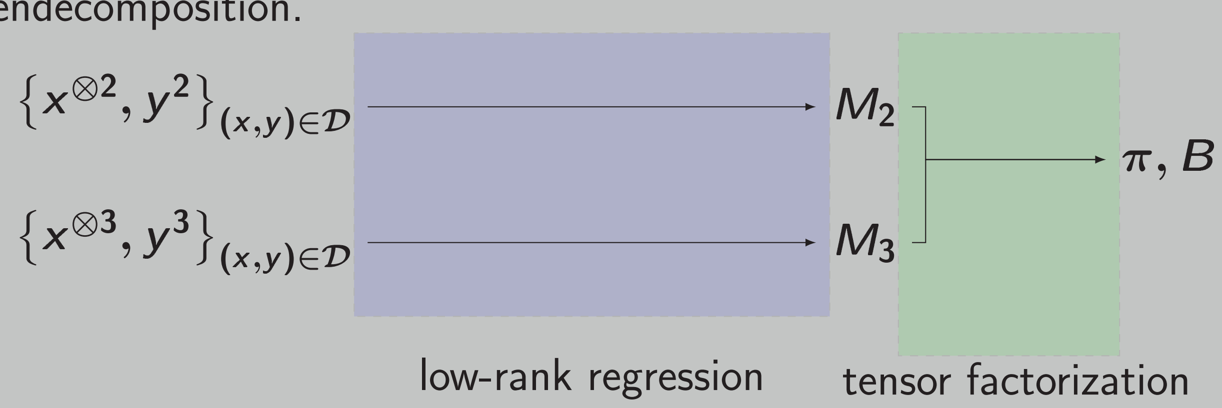


- For a particular x, we draw y as follows,
  - $h \sim \mathrm{Mult}([\pi_1, \pi_2, \cdots, \pi_k])$.
  - $y = \beta_h^T x + \epsilon$.
- Given $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we want to recover the parameters π and B.



$$\begin{bmatrix}\pi_1 \\ \pi_2 \\ \vdots \\ \pi_k\end{bmatrix} \quad \underbrace{\begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_k \end{bmatrix}}_{B}$$

- Our approach uses low-rank regression to reduce the problem to tensor eigendecomposition.



Sample Complexity: $O\left(k \; \|x\|^{12} \|\beta\|^6 \| \mathbb{E}[\epsilon^2] \|^6\right) \quad O\left(\frac{k \pi_{\max}^2}{\sigma_k(M_2)^5}\right)$

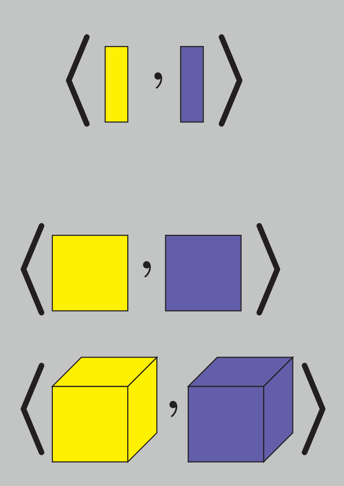
Step 1: Finding Tensor Structure via Regression

- **Key Observation:** Regression on the powers of (y, x) gives us the expected powers of the regression coefficients β.

$$y = \langle \beta_h, x \rangle + \epsilon = \underbrace{\langle \mathbb{E}[\beta_h], x \rangle}_{\text{linear measurement}} + \underbrace{(\beta_h - \mathbb{E}[\beta_h])^T x + \epsilon}_{\text{noise}}$$

$$y^2 = (\langle \beta_h, x \rangle + \epsilon)^2 = \underbrace{\langle \mathbb{E}[\beta_h^{\otimes 2}], x^{\otimes 2} \rangle}_{M_2} + \mathrm{bias}_2 + \mathrm{noise}_2$$

$$y^3 = \underbrace{\langle \mathbb{E}[\beta_h^{\otimes 3}], x^{\otimes 3} \rangle}_{M_3} + \mathrm{bias}_3 + \mathrm{noise}_3$$



- M_2 and M_3 are both of rank k, so we can use low rank regression[3,4]!

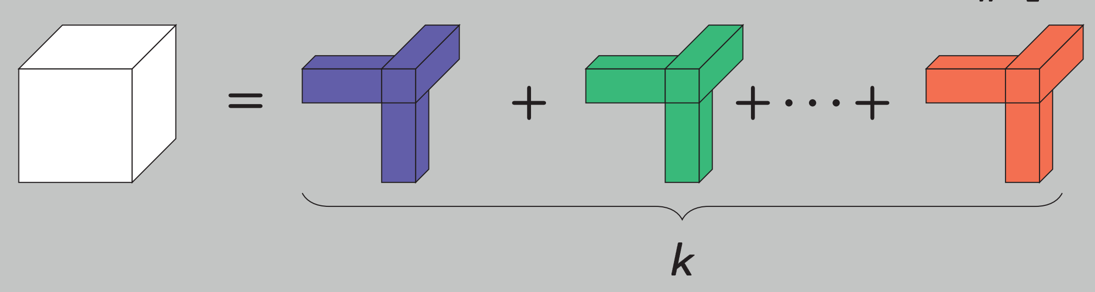
$$\hat{M}_2 = \arg\min_M \sum_{(x,y)\in\mathcal{D}} \left(y^2 - \langle M, x^{\otimes 2} \rangle - \mathrm{bias}_2\right)^2 + \lambda_2 \underbrace{\|M\|_*}_{\sum_i \sigma_i(M)}$$

$$\hat{M}_3 = \arg\min_M \sum_{(x,y)\in\mathcal{D}} \left(y^3 - \langle M, x^{\otimes 3} \rangle - \mathrm{bias}_3\right)^2 + \lambda_3 \|M\|_*$$

[3] Fazel, 2002; [4] Tomoika, Hayashi and Kashima, 2010

Step 2: Parameter Recovery via Tensor Factorization

- M_3 has a low-rank tensor decomposition: $M_3 = \sum_{h=1}^k \pi_h \beta_h^{\otimes 3}$

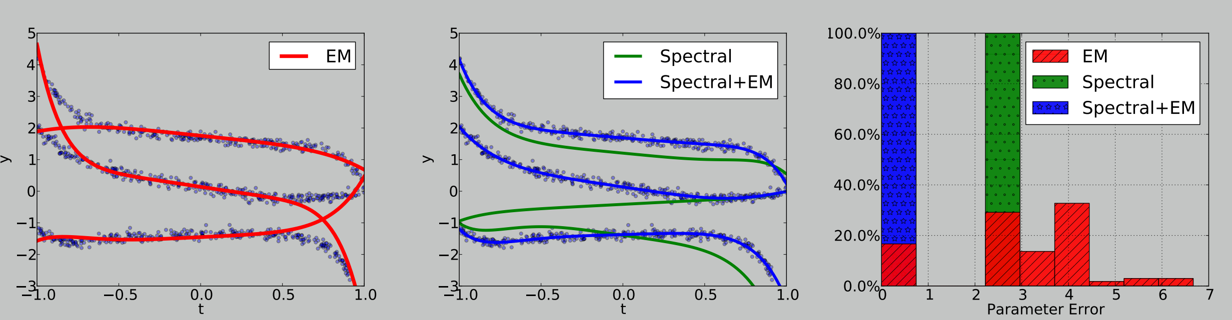


- **Key Observation:** If β_h are orthogonal, they are eigenvectors[5]; $M_3(\beta_h, \beta_h) = \pi_h \beta_h$.
- In general, we can whiten M_3 first.

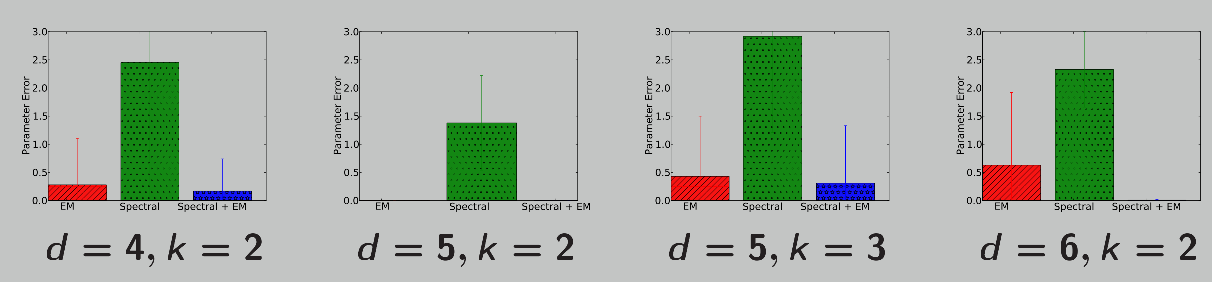
[5]: Anandkumar, Ge, Hsu, Kakade, Telgarksy, 2012.

Experiments

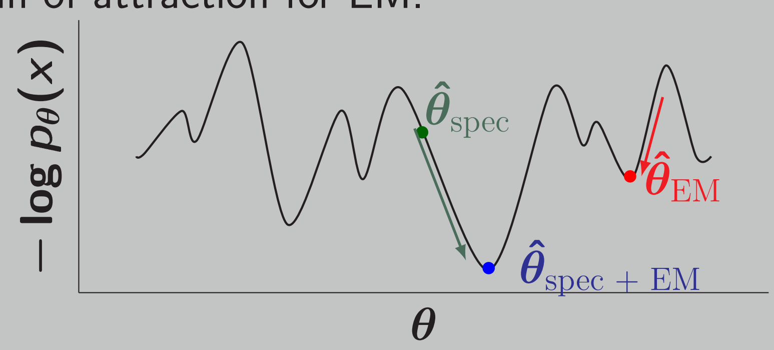
- With finite samples, Spectral Experts seems to find parameters that sufficiently separate components that EM initialized with these parameters recovers true parameters more often than EM with random initializations.
- In this example, $y = \beta^T [1, t, t^4, t^7]^T + \epsilon$. $k = 3, d = 4, n = 10^5$,



- Below are parameter errors averaged over 10 initializations on 10 different simulated datasets with the specified parameter configurations,



- **Log-likelihood cartoon:** It seems that our parameter estimates fall in the right basin of attraction for EM.



Future Work

- How can we handle other discriminative models?
  - Non-linear link functions (hidden variable logistic regression).
  - Dependencies between h and x (mixture of experts).

ℙλ at Stanford University   chaganty@cs.stanford.edu, pliang@cs.stanford.edu