Interpreting Latent Variables via Convex Optimization

Armeen Taeb
Electrical Engineering, Caltech

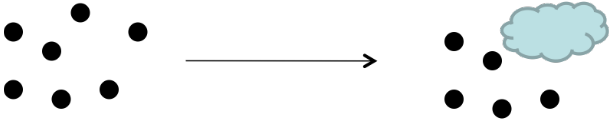
Joint work with Venkat Chandrasekaran

Statistical Model Selection

- Random variables:
  - Financial assets
  - Gene expressions
  - ...
- Describe the statistical behavior using concisely parameterized models
  - Manifold learning
  - Graphical models
  - Time series analysis
  - Principal components analysis
  - ...

Problem

- What if some of the variables are **not observed**?
- Don't know how many latent variables
- Don't know the effect of latent variables
- Confounding dependencies if latent variables are not taken into account



Many Approaches

- Modeling Frameworks
  - Factor Analysis [Spearman (1904)]
  - Mixture modeling
  - Graphical models with latent variables
- Algorithmic Techniques
  - EM Algorithm [Dempster, Laird, Rubin (1977)]
  - Convex relaxation
  - Greedy algorithms

Challenge

- The latent variables are purely mathematical constructs with no semantics!
- How do we obtain semantic information about latent variables?
- Example: Factor analysis
  - Attribute meaning to the latent variables
  - E.g., what are the factors influencing stock returns
- How do we think about this in a principled way?

Factor Modeling

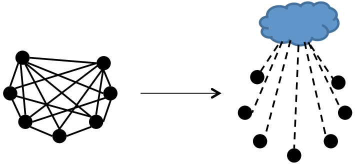
- A factor model over random variables $y \in \mathbb{R}^p$

$$y = \mathcal{A}\zeta + \epsilon$$

  - $\zeta \in \mathbb{R}^k \quad =$ latent variables, $\quad k \ll p$
  - $\mathcal{A} : \mathbb{R}^k \to \mathbb{R}^p \quad =$ linear map; $\mathcal{A}\zeta$: effect of latent variables
  - ζ and ϵ are independent; ϵ has independent components

A few latent variables explain the variability of y.

- Factor analysis: fit observations of y to identify:
  - Number of latent variables
  - The effect $\mathcal{A}\zeta$ (more about this later!)
  - Variance of ϵ



What are the Latent Variables

- Question: Can we assign semantics to these latent variables?

- Idea: obtain measurements of additional variables x that are related to y and associate these to ζ.

  - $\underbrace{\text{stock returns}}_{y} + \underbrace{\text{oil / currency / weather / GDP}}_{x}$

  - $\underbrace{\text{gene expressions}}_{y} + \underbrace{\text{physiological attributes}}_{x}$

  - number of covariates could be potentially very large!

- Challenge: There are infinitely many parameterizations of ζ.

  - Since $\mathcal{A}\zeta = \mathcal{A}\mathcal{W}[\mathcal{W}^{-1}\zeta]$ for any nonsingular $\mathcal{W}$, $\mathcal{W}^{-1}\zeta$ is an equally good parameterization.

> The key invariant is the column-space of $\mathcal{A}$ or the effect of ζ on y given by $\mathcal{A}\zeta$.

Decomposing the Effect of Latent Variables

- Decompose the column-space of $\mathcal{A}$ into:
  - A subspace captured by $x \in \mathbb{R}^q$ (observed variables)
  - A subspace captured by residual latent variables (unobserved phenomena)
- Identify linear maps $\mathcal{B} : \mathbb{R}^q \to \mathbb{R}^p, \mathcal{C} : \mathbb{R}^h \to \mathbb{R}^p$ such that:

$$\mathcal{A}\zeta \approx \mathcal{B}x + \mathcal{C}z$$

  - column-space $(\mathcal{A})$ = column-space $(\mathcal{B}) \oplus$ column-space (C)
  - z: residual latent variables
  - rank$(\mathcal{B}) \ll \{p, q\}$ since dim(col-space($\mathcal{B}$)) is small.
  - Column-space of $\mathcal{B}$ represents the interpretable component of the effect of latent variables

Composite Model (with covariates)

- Composite model

$$y = \mathcal{B}x + \mathcal{C}z + \epsilon$$

  - $\mathcal{B} : \mathbb{R}^q \to \mathbb{R}^p$; rank$(\mathcal{B}) \ll \{p, q\}$
  - $\mathcal{C} : \mathbb{R}^h \to \mathbb{R}^p$; $h \ll \{p, q\}$
  - x, z, ϵ independent random variables; ϵ has independent components, $z \in \mathbb{R}^h$

- The quantity $\mathcal{B}x$ is the interpretable component of the effect of latent variables

- Take away: interpreting latent variables means learning an accurate composite factor model

Composite Model (with covariates)

- Composite model

$$y = \mathcal{B}x + \mathcal{C}z + \epsilon$$

- Fit joint observations of (y, x) to composite model and identify:
  - The map $\mathcal{B}$
  - Number of residual latent variables
  - The effect of residual latent variables: $\mathcal{C}z$
  - Variance of ϵ
- How do we learn parameters of this model?

$$y = \mathcal{B}x + \mathcal{C}z + \epsilon$$

- Inverse covariance matrix of (y, x) has rich structure:

$$\Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta_{xy} & \Theta_x \end{pmatrix} \in \mathbb{S}^{(p+q)\times(p+q)}$$

  - rank(Θ_{yx}) = rank$(\mathcal{B}) \ll \min\{p, q\}$ since $\mathcal{B} = \Theta_y^{-1}\Theta_{yx}$
  - $\Sigma_{y \mid x} = \underbrace{\Sigma_\epsilon}_{\text{diagonal}} + \underbrace{\mathcal{C}zz'C'}_{\text{low rank}}$
  - $\Theta_y = \Sigma_{y \mid x}^{-1} =$ Diagonal - Low rank

Learning a Composite Model (with covariates)

- Given observations $\mathcal{D}_n = \left\{(y^{(1)}, x^{(1)}), (y^{(2)}, x^{(2)}), \ldots (y^{(n)}, x^{(n)})\right\}$

- A natural approach for $\lambda_n, \gamma > 0$

$$\arg\min_{\Theta, D, L} \quad -\text{log.lik}\ (\Theta, \mathcal{D}_n) + \lambda_n[\gamma\ \text{rank}(L) + \text{rank}(\Theta_{yx})]$$

$$\text{s.t.} \quad \Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta_{xy} & \Theta_x \end{pmatrix}; \Theta_y = D - L,\ L \succeq 0, D \text{ is diagonal}$$

- Rank penalty is nonconvex!

Convex relaxation

- Computationally tractable relaxation for inducing low rank structure [Fazel (2002), Boyd, Recht, Parrilo, ...]

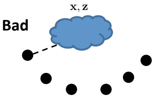
$$\mathsf{rank}(M) \quad \longrightarrow \quad \|M\|_{\mathsf{nuc}} = \sum_i \sigma_i(M)$$

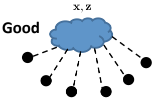
- A convex relaxation for $\lambda_n, \gamma > 0$

$$\begin{aligned} \arg\min_{\Theta, D, L} \qquad & -\mathsf{log.lik}\ (\Theta, \mathcal{D}_n) + \lambda_n[\gamma\ \|L\|_{\mathsf{nuc}} + \|\Theta_{yx}\|_{\mathsf{nuc}}] \\ \text{s.t.} \quad & \Theta = \begin{pmatrix} \Theta_y & \Theta_{yx} \\ \Theta_{xy} & \Theta_x \end{pmatrix}; \Theta_y = D - L,\ L \succeq 0, D \text{ is diagonal} \end{aligned}$$

- When does the estimator identify the underlying model?

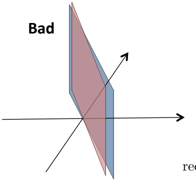
$$y = \mathcal{B}x + \mathcal{C}z + \epsilon$$



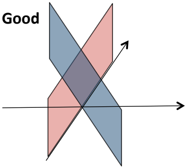


Assumption 1: The effect of x, z must not "concentrate" on any y

- Otherwise, the effect of x, z can be absorbed into ϵ.



red = col-space($\mathcal{B}$)
blue = col-space($\mathcal{C}$)



Assumption 2: The column-spaces of $\mathcal{B}$ and $\mathcal{C}$ must be sufficiently transverse

- Otherwise, the effect of x and z cannot be distinguished.

Assumptions: Identifiability

$$y = \mathcal{B}x + \mathcal{C}\zeta + \epsilon$$

- Assumption 1: column-space($\mathcal{B}$) and column-space($\mathcal{C}$) should not contain elements from the standard basis
    - $\max_i \|\mathcal{P}_{\text{col-space}(\mathcal{B})} e_i\|_2$ must be small
    - $\max_i \|\mathcal{P}_{\text{col-space}(\mathcal{C})} e_i\|_2$ must be small
    - related to the coherence parameter in Candes, Recht [2010]
- Assumption 2: column-space($\mathcal{B}$) and column-space($\mathcal{C}$) should be sufficiently transverse (i.e. have large angle)
- **Thm**: Estimator identifies underlying model w.h.p provided
    - n is larger than the combined dimension of (y, x)
    - Identifiability conditions are satisfied

Experiment 1: Financial Asset Problem

- Responses: Monthly stock return of 45 companies (1982-2016); $y \in \mathbb{R}^{45}$
- 13 Covariates, $x \in \mathbb{R}^{13}$:
  - EUR to USD exchange rate
  - Government expenditures
  - Federal debt
  - Federal reserve rate
  - GDP growth rate
  - Industrial production rate
  - Mortgage Rate
  - Oil import
  - Saving rate
  - Consumer price index
  - Producer price index
  - Home ownership rate
  - Inflation rate

Experiment 1: Financial Asset Problem

- Pure factor model (only on responses): $y = \mathcal{A}\zeta + \epsilon$
  - 10 latent factors
- Composite model (responses and covariates) $y = \mathcal{B}x + \mathcal{C}z + \epsilon$
  - $\dim(z) = 8$, $\mathrm{rank}(\mathcal{B}) = 2$
- How good are the covariates at giving interpretation to latent variables of the factor model?
  - Principal angles between the 2-dimensional column-space of $\mathcal{B}$ and 10-dimensional column-space of $\mathcal{A}$ are :
    - $6, 16$ degrees
- Projection of x onto 2-dimensional row-space of $\mathcal{B}$ represents relevant component
  - EUR to USD exchange and Government spending are most relevant for capturing the latent phenomena.

Experiment 2: California Reservoir Modeling

- Collaborators: Michael Turmon and JT Reager (Jet Propulsion Laboratory)
- Responses: Monthly average levels of 55 reservoirs in California (2003-2014); $y \in \mathbb{R}^{55}$
- 8 Covariates; $x \in \mathbb{R}^8$:
  - Palmer Drought Index
  - Hydroelectric Power
  - Unemployment rate
  - Temperature in Sacramento region
  - Temperature in San Joaquin region
  - Precipitation in Sacramento region
  - Precipitation in San Joaquin region
  - Consumer price index

Experiment 2: California Reservoir Modeling

- Pure factor model (only on responses) $y = \mathcal{A}\zeta + \epsilon$
  - 14 latent factors
- Composite model (responses and covariates) $y = \mathcal{B}x + \mathcal{C}z + \epsilon$
  - dim$(z) = 2$, rank$(\mathcal{B}) = 2$
- How good are the covariates at giving interpretation to latent variables of the factor model?
  - Principal angles between the 4-dimensional column-space of $\mathcal{B}$ and 14-dimensional column-space of $\mathcal{A}$ are :
    - $0.347, 1.6, 2.45, 4.55$ degrees
- Projection of x onto 4-dimensional row-space of $\mathcal{B}$ represents relevant component
  - Drought index and hydroelectric power are most relevant for capturing the latent phenomena.

Summary

- **Semantic information** about latent variables of a factor model
  - Measure additional variables, and link these to latent variables ; A convex approach with statistical guarantees
- Several extensions
  - Graphical models with latent variables
- Future work
  - Extension to Generalized linear models or non-Gaussian models

Paper

- Interpreting latent variables in factor models via convex optimization, preprint 2016.
  - **T.**, and Chandrasekaran.
- California reservoir drought sensitivity and exhaustion risk using statistical graphical models, preprint 2016
  - **T.**, Reager, Turmon, and Chandrasekaran.

http://www.its.caltech.edu/ ataeb/index.html