The Qualification Exam

Y. Cem Sübakan

UIUC CS

September 18, 2014

Outline

Who is this kid?

- I am with Paris Smaragdis' group. This is my third semester here.

Who is this kid?

- I am with Paris Smaragdis' group. This is my third semester here.
- I do machine learning research (or so I claim).

Who is this kid?

- I am with Paris Smaragdis' group. This is my third semester here.
- I do machine learning research (or so I claim).
- Both my undergrad and masters are from electrical engineering, Bogazici Uni. - Istanbul.

Who is this kid?

- I am with Paris Smaragdis' group. This is my third semester here.
- I do machine learning research (or so I claim).
- Both my undergrad and masters are from electrical engineering, Bogazici Uni. - Istanbul.
- I started with Bayesian ML. My previous advisor was doing Bayesian Machine Learning.

Who is this kid?

- I am with Paris Smaragdis' group. This is my third semester here.
- I do machine learning research (or so I claim).
- Both my undergrad and masters are from electrical engineering, Bogazici Uni. - Istanbul.
- I started with Bayesian ML. My previous advisor was doing Bayesian Machine Learning.
- **My research in one sentence:**
  I like big algorithms for small data, and I like NIPS/ICML style machine learning.

Outline

1. Me
   - My background
   - My research
2. Paper 1: Estimating Latent Variable Graphical Models using Moments and Likelihoods
   - Introduction
   - Intro to method of moments for LVMs
   - The paper
3. Second Paper, The Visual Microphone: Passive Recovery of Sound from Video
   - Introduction, The problem setup
   - Processing step

My research overview

- I am interested in parameter estimation problem in latent variable models (mixture models/ HMMs/ MRFs etc.).

My research overview

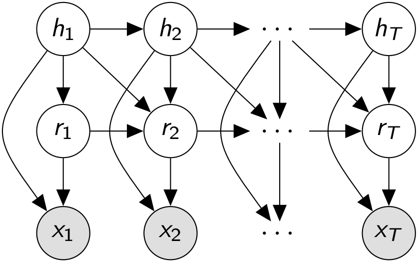
- I am interested in parameter estimation problem in latent variable models (mixture models/ HMMs/ MRFs etc.).
- In particular, I am working on Method of Moments (MoM) for parameter estimation in LVMs. (Also known as Spectral Learning). Score so far:
  - M.Sc. Thesis
  - 2 NIPS workshop papers
  - 1 journal paper
  - NIPS 2014 paper NEW!

My research overview

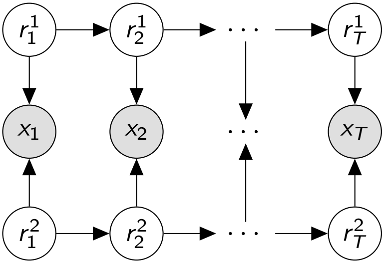
- I am interested in parameter estimation problem in latent variable models (mixture models/ HMMs/ MRFs etc.).
- In particular, I am working on Method of Moments (MoM) for parameter estimation in LVMs. (Also known as Spectral Learning). Score so far:
  - M.Sc. Thesis
  - 2 NIPS workshop papers
  - 1 journal paper
  - NIPS 2014 paper NEW!
- WHY MoM estimators?
  - They are cool, mathy and new (hip).
  - Avoid the everlasting local optima issue. (No initialization!)
  - Computationally much more efficient.
  - Learning guarantees.

Recent Research

- M.Sc. Thesis: Two new MoM algorithms for time series clustering.
- ICML 2014 submission: A Non-Negative Matrix Factorization (NMF) based framework for learning HMM variants with MoM.



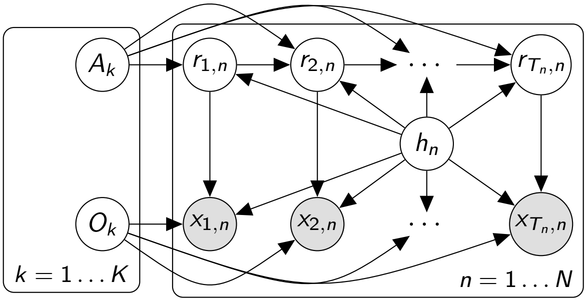
Switching HMM



Factorial HMM

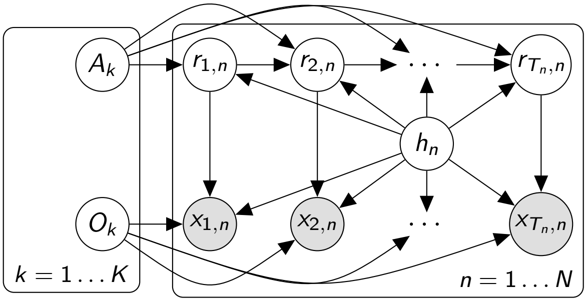
Recent Research: NIPS 2014 Paper

- Paper accepted to NIPS 2014! (acceptance rate: 414/1678) A Method of moments algorithm to learn mixture of HMMs.



Recent Research: NIPS 2014 Paper

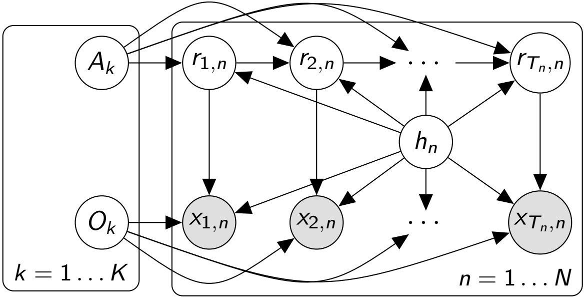
- Paper accepted to NIPS 2014! (acceptance rate: 414/1678) A Method of moments algorithm to learn mixture of HMMs.



- It is unclear how to use standard MoM algorithms for this model.

Recent Research: NIPS 2014 Paper

- Paper accepted to NIPS 2014! (acceptance rate: 414/1678) A Method of moments algorithm to learn mixture of HMMs.



- It is unclear how to use standard MoM algorithms for this model.
- However, we can learn an HMM with MoM.

Recent Research: NIPS 2014 Paper

- Key idea: Mixture of HMMs is an HMM with block diagonal transition matrix.
  - An MHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, \pi)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & A_K \end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix} \pi_1 \nu_1 \\ \pi_2 \nu_2 \\ \vdots \\ \pi_K \nu_K \end{bmatrix} .$$

Recent Research: NIPS 2014 Paper

- Key idea: Mixture of HMMs is an HMM with block diagonal transition matrix.
  - An MHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, \pi)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & A_K \end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix} \pi_1 \nu_1 \\ \pi_2 \nu_2 \\ \vdots \\ \pi_K \nu_K \end{bmatrix} .$$

- The problem: Arbitrary permutation on parameter estimates, Parameters of different clusters get mixed up.

Recent Research: NIPS 2014 Paper

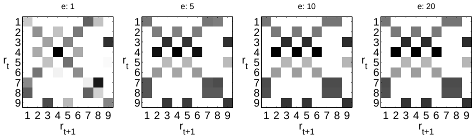
- Key idea: Mixture of HMMs is an HMM with block diagonal transition matrix.
  - An MHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, \pi)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & A_K \end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix} \pi_1 \nu_1 \\ \pi_2 \nu_2 \\ \vdots \\ \pi_K \nu_K \end{bmatrix} .$$

- The problem: Arbitrary permutation on parameter estimates, Parameters of different clusters get mixed up.
- Remedy: Block diagonal structure / spectral properties of the “global” transition matrix.

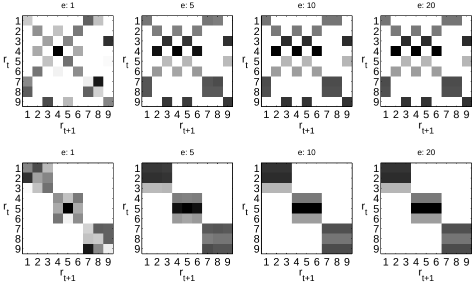
Recent Research: NIPS 2014 Paper

- Ideally, we have a "clean" block diagonal structure. $\lim_{e\to\infty} A^e$ reveals a 3 cluster structure.



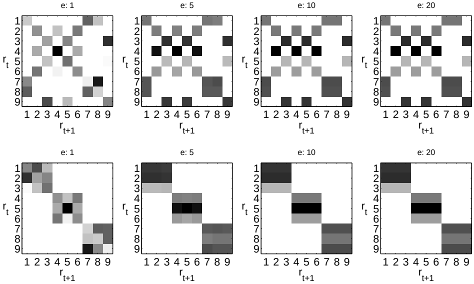
Recent Research: NIPS 2014 Paper

- Ideally, we have a "clean" block diagonal structure. $\lim_{e \to \infty} A^e$ reveals a 3 cluster structure.

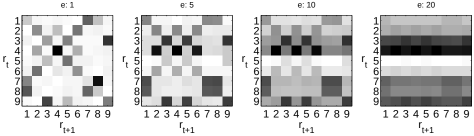


Recent Research: NIPS 2014 Paper

- Ideally, we have a "clean" block diagonal structure. $\lim_{e \to \infty} A^e$ reveals a 3 cluster structure.



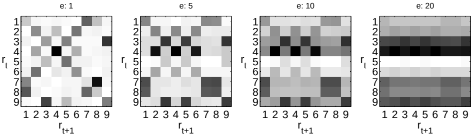
- In real world, we have noise on off-block diagonal elements. This results in a global stationary distribution.



Recent Research: NIPS 2014 Paper

- The key question:

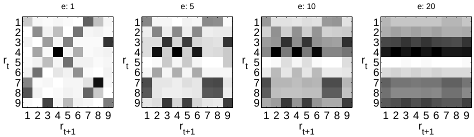
Can we recover a block diagonal structure despite the estimation noise?



Recent Research: NIPS 2014 Paper

- The key question:

  Can we recover a block diagonal structure despite the estimation noise?

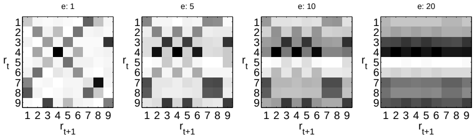


- If the noise is not too severe, then yes we can. (Experimental and theoretical justification)

Recent Research: NIPS 2014 Paper

- The key question:

  Can we recover a block diagonal structure despite the estimation noise?



- If the noise is not too severe, then yes we can. (Experimental and theoretical justification)
- Notice: Given the moments, computational burden does not depend on dataset size! (Unlike EM) COOL

Outline

Paper 1: Estimating Latent Variable Graphical Models using Moments and Likelihoods

- Standard MoM algorithms are not directly applicable to models beyond HMM, GMM, LDA.

Paper 1: Estimating Latent Variable Graphical Models using Moments and Likelihoods

- Standard MoM algorithms are not directly applicable to models beyond HMM, GMM, LDA.
- This work proposes a framework for learning general graphical models.

Paper 1: Estimating Latent Variable Graphical Models using Moments and Likelihoods

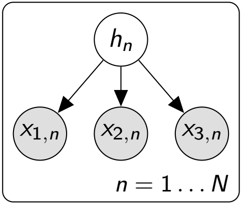
- Standard MoM algorithms are not directly applicable to models beyond HMM, GMM, LDA.
- This work proposes a framework for learning general graphical models.
- They divide the problem into two (three) stages, which helps generalizing.

Outline

1. Me
   - My background
   - My research
2. Paper 1: Estimating Latent Variable Graphical Models using Moments and Likelihoods
   - Introduction
   - Intro to method of moments for LVMs
   - The paper
3. Second Paper, The Visual Microphone: Passive Recovery of Sound from Video
   - Introduction, The problem setup
   - Processing step

Problem Definition

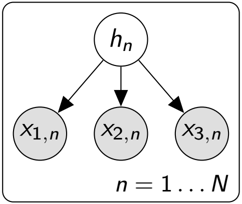
- Let's suppose we have the following graphical model:



$$
\begin{aligned}
h &\sim \text{Discrete}(\pi_{1:K}) \\
x_1 | h &\sim \mathcal{N}(\mu_{1,h}, \Sigma_1) \\
x_2 | h &\sim \mathcal{N}(\mu_{2,h}, \Sigma_2) \\
x_3 | h &\sim \mathcal{N}(\mu_{3,h}, \Sigma_3)
\end{aligned}
$$

Problem Definition

- Let's suppose we have the following graphical model:

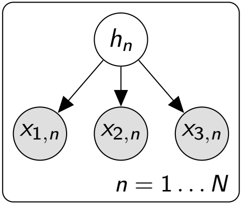


$$
\begin{aligned}
h &\sim \text{Discrete}(\pi_{1:K}) \\
x_1 | h &\sim \mathcal{N}(\mu_{1,h}, \Sigma_1) \\
x_2 | h &\sim \mathcal{N}(\mu_{2,h}, \Sigma_2) \\
x_3 | h &\sim \mathcal{N}(\mu_{3,h}, \Sigma_3)
\end{aligned}
$$

- Given $\{x_{1,n}, x_{2,n}, x_{3,n}\}_{n=1}^{N}$, can we estimate $\mu_{1,1:K}, \mu_{2,1:K}, \mu_{3,1:K}$?

Problem Definition

- Let's suppose we have the following graphical model:



$$\begin{aligned} h &\sim \text{Discrete}(\pi_{1:K}) \\ x_1 | h &\sim \mathcal{N}(\mu_{1,h}, \Sigma_1) \\ x_2 | h &\sim \mathcal{N}(\mu_{2,h}, \Sigma_2) \\ x_3 | h &\sim \mathcal{N}(\mu_{3,h}, \Sigma_3) \end{aligned}$$

- Given $\{x_{1,n}, x_{2,n}, x_{3,n}\}_{n=1}^{N}$, can we estimate $\mu_{1,1:K}, \mu_{2,1:K}, \mu_{3,1:K}$?

Yes we can!

The conventional way: EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_{\mu_{1:3}} p(x_{1:3,1:N}|\mu_{1:3}) = \max_{\mu_{1:3}} \sum_{h_{1:N}} p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})$$

The conventional way: EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_{\mu_{1:3}} p(x_{1:3,1:N}|\mu_{1:3}) = \max_{\mu_{1:3}} \sum_{h_{1:N}} p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h_{1:N})$:

$$\log \sum_{h_{1:N}} p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3}) \frac{q(h_{1:N})}{q(h_{1:N})} = \log \mathbb{E}_{q(h_{1:N})} \left[\frac{p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})}{q(h_{1:N})} \right]$$
$$\geq \mathbb{E}_{q(h_{1:N})} \left[\log p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3}) \right] + H_q$$

The conventional way: EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_{\mu_{1:3}} p(x_{1:3,1:N}|\mu_{1:3}) = \max_{\mu_{1:3}} \sum_{h_{1:N}} p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h_{1:N})$:

$$\log \sum_{h_{1:N}} p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3}) \frac{q(h_{1:N})}{q(h_{1:N})} = \log \mathbb{E}_{q(h_{1:N})} \left[\frac{p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})}{q(h_{1:N})} \right]$$

$$\geq \mathbb{E}_{q(h_{1:N})} [\log p(x_{1:3,1:N}, h_{1:N}|\mu_{1:3})] + H_q$$

- $q(h_{1:N}) = p(h_{1:N}|x_{1:N}, \mu_{1:3})$ in EM. In E step q is updated. In M step we maximize this lower bound. It is obviously prone to local optima.

The other way: Method of Moments

- The idea is to estimate the models parameters $\mu_{1:K}$ by solving a system of non-linear equations formed with moments $\mathbb{E}[g_k(x)]$, $k \in \{1, \ldots K\}$:

$$
\begin{aligned}
\mathbb{E}[g_1(x)] =& f_1(\mu_{1:K}) \\
& \vdots \\
\mathbb{E}[g_K(x)] =& f_K(\mu_{1:K})
\end{aligned}
$$

The other way: Method of Moments

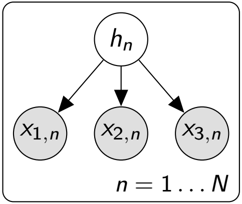
- The idea is to estimate the models parameters $\mu_{1:K}$ by solving a system of non-linear equations formed with moments $\mathbb{E}[g_k(x)]$, $k \in \{1, \ldots K\}$:

$$\begin{aligned}\mathbb{E}[g_1(x)] =& f_1(\mu_{1:K}) \\ &\vdots \\ \mathbb{E}[g_K(x)] =& f_K(\mu_{1:K})\end{aligned}$$

- Canonical Example: $x \sim \mathcal{G}(a, b)$:

$$\begin{aligned}\mathbb{E}[x] =& ab \\ \mathbb{E}[x^2] =& ab^2 + a^2 b^2\end{aligned} \quad \rightarrow \quad \begin{aligned}\widehat{b} =& (\mathbb{E}[x^2] - \mathbb{E}[x]^2)/\mathbb{E}[x] \\ \widehat{a} =& \mathbb{E}[x]^2/(\mathbb{E}[x^2] - \mathbb{E}[x]^2)\end{aligned}$$

The new way: Method of Moments

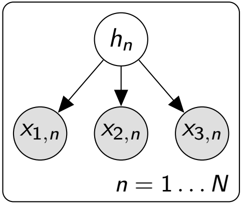


$$
\begin{aligned}
h_n &\sim \text{Discrete}(\pi) \\
x_1|h &\sim \mathcal{N}(\mu_{1,h}, \Sigma_1) \\
x_2|h &\sim \mathcal{N}(\mu_{2,h}, \Sigma_2) \\
x_3|h &\sim \mathcal{N}(\mu_{3,h}, \Sigma_3)
\end{aligned}
$$

- Let's write down some moments:

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] = \sum_{h=1}^{K} \pi_h \, \mu_{1,h} \otimes \mu_{2,h}$$

The new way: Method of Moments



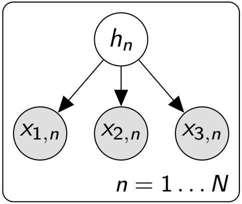
$$h_n \sim \text{Discrete}(\pi)$$
$$x_1|h \sim \mathcal{N}(\mu_{1,h}, \Sigma_1)$$
$$x_2|h \sim \mathcal{N}(\mu_{2,h}, \Sigma_2)$$
$$x_3|h \sim \mathcal{N}(\mu_{3,h}, \Sigma_3)$$

- Let's write down some moments:

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h}$$

$$P_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h} \otimes \mu_{3,h}$$

The new way: Method of Moments



$$
\begin{aligned}
h_n &\sim \text{Discrete}(\pi) \\
x_1 | h &\sim \mathcal{N}(\mu_{1,h}, \Sigma_1) \\
x_2 | h &\sim \mathcal{N}(\mu_{2,h}, \Sigma_2) \\
x_3 | h &\sim \mathcal{N}(\mu_{3,h}, \Sigma_3)
\end{aligned}
$$

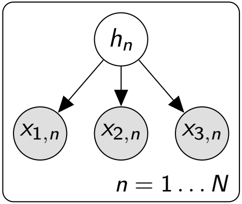
- Let's write down some moments:

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] = \sum_{h=1}^{K} \pi_h \, \mu_{1,h} \otimes \mu_{2,h}$$

$$P_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h=1}^{K} \pi_h \, \mu_{1,h} \otimes \mu_{2,h} \otimes \mu_{3,h}$$

- So, $P_2 = M_1 \text{diag}(\pi) M_2$ and $P_{3,i} = M_1 \text{diag}(M_3(i,:)) \text{diag}(\pi) M_2$.

The new way: Method of Moments



$$h_n \sim \text{Discrete}(\pi)$$
$$x_1 | h \sim \mathcal{N}(\mu_{1,h}, \Sigma_1)$$
$$x_2 | h \sim \mathcal{N}(\mu_{2,h}, \Sigma_2)$$
$$x_3 | h \sim \mathcal{N}(\mu_{3,h}, \Sigma_3)$$

- Let's write down some moments:

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1 | h] \otimes \mathbb{E}[x_2 | h] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h}$$

$$P_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h} \otimes \mu_{3,h}$$

- So, $P_2 = M_1 \text{diag}(\pi) M_2$ and $P_{3,i} = M_1 \text{diag}(M_3(i,:)) \text{diag}(\pi) M_2$.
- And, $P_{3,i} P_2^{-1} = M_1 \text{diag}(M_3(i,:)) M_1^{-1}$, which is an eigenvalue decomposition (assuming invertibility).

The new way: Method of Moments

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h}$$

$$P_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h} \otimes \mu_{3,h}$$

- So, $P_2 = M_1 \mathrm{diag}(\pi) M_2$ and $P_{3,i} = M_1 \mathrm{diag}(M_3(i,:))\mathrm{diag}(\pi) M_2$.
- And, $P_{3,i} P_2^{-1} = M_1 \mathrm{diag}(M_3(i,:)) M_1^{-1}$, which is an eigenvalue decomposition (assuming invertibility).

The new way: Method of Moments

$$P_2 := \mathbb{E}[x_1 \otimes x_2] = \sum_{h=1}^{K} \pi_h \mathbb{E}[x_1|h] \otimes \mathbb{E}[x_2|h] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h}$$

$$P_3 := \mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h=1}^{K} \pi_h \; \mu_{1,h} \otimes \mu_{2,h} \otimes \mu_{3,h}$$

- So, $P_2 = M_1 \text{diag}(\pi) M_2$ and $P_{3,i} = M_1 \text{diag}(M_3(i,:)) \text{diag}(\pi) M_2$.
- And, $P_{3,i} P_2^{-1} = M_1 \text{diag}(M_3(i,:)) M_1^{-1}$, which is an eigenvalue decomposition (assuming invertibility).
- This is from Anandkumar et al. 2012, COLT paper. There are statistically more efficient ways now. (Using all three slices instead of one. Anandkumar et al. 2014, to appear in JMLR)

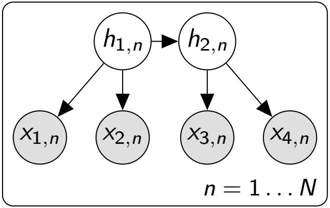
Outline

Good, but what about your qual paper?

- The paper is about generalizing method of moments idea to general graph structures.

Good, but what about your qual paper?

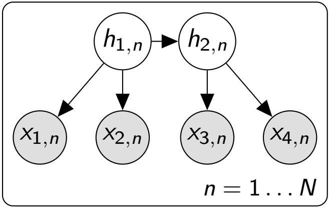
- The paper is about generalizing method of moments idea to general graph structures.
- For example,



$$
\begin{aligned}
h_1 &\sim \text{Discrete}(\pi) \\
h_2 | h_1 &\sim \text{Discrete}(A(:, h_1)) \\
x_1 | h_1 &\sim \mathcal{N}(\mu_{1,h_1}, \Sigma_1) \\
x_2 | h_1 &\sim \mathcal{N}(\mu_{2,h_1}, \Sigma_2) \\
x_3 | h_2 &\sim \mathcal{N}(\mu_{3,h_2}, \Sigma_3) \\
x_4 | h_2 &\sim \mathcal{N}(\mu_{4,h_2}, \Sigma_3)
\end{aligned}
$$

Good, but what about your qual paper?

- The paper is about generalizing method of moments idea to general graph structures.
- For example,

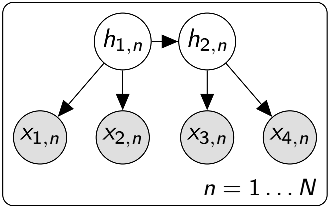


$$
\begin{aligned}
h_1 &\sim \text{Discrete}(\pi) \\
h_2 | h_1 &\sim \text{Discrete}(A(:, h_1)) \\
x_1 | h_1 &\sim \mathcal{N}(\mu_{1,h_1}, \Sigma_1) \\
x_2 | h_1 &\sim \mathcal{N}(\mu_{2,h_1}, \Sigma_2) \\
x_3 | h_2 &\sim \mathcal{N}(\mu_{3,h_2}, \Sigma_3) \\
x_4 | h_2 &\sim \mathcal{N}(\mu_{4,h_2}, \Sigma_3)
\end{aligned}
$$

- Now, can we learn A, $\mu_1, \mu_2, \mu_3, \mu_4$ using moments?

Good, but what about your qual paper?

- The paper is about generalizing method of moments idea to general graph structures.
- For example,



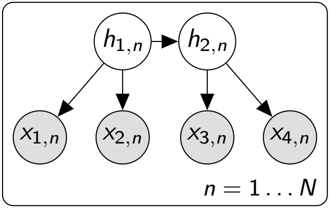
$$h_1 \sim \text{Discrete}(\pi)$$
$$h_2|h_1 \sim \text{Discrete}(A(:, h_1))$$
$$x_1|h_1 \sim \mathcal{N}(\mu_{1,h_1}, \Sigma_1)$$
$$x_2|h_1 \sim \mathcal{N}(\mu_{2,h_1}, \Sigma_2)$$
$$x_3|h_2 \sim \mathcal{N}(\mu_{3,h_2}, \Sigma_3)$$
$$x_4|h_2 \sim \mathcal{N}(\mu_{4,h_2}, \Sigma_3)$$

- Now, can we learn A, $\mu_1, \mu_2, \mu_3, \mu_4$ using moments?
- Not straightforwardly with original work. But this paper says,

Yes, we can!

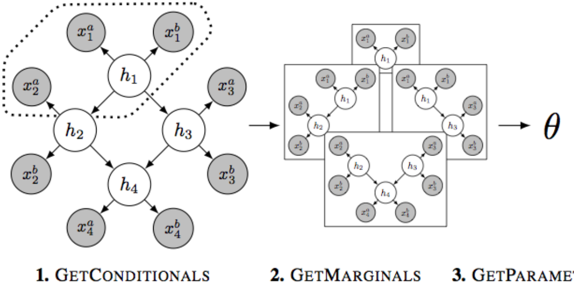
Paper 1: Key Idea

- Learning conditional moments and hidden marginals separately

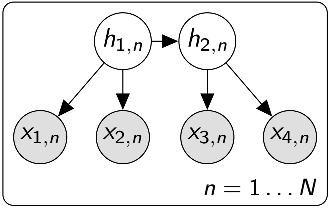


- First estimate the conditional moments $\mathbb{E}[x_i|h_k]$.
- Then obtaining the hidden potential $p(h_2|h_1)$ is easy.

- The pipeline:

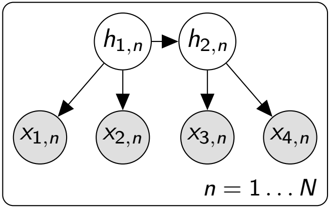


Part 1: Estimating the conditional moments



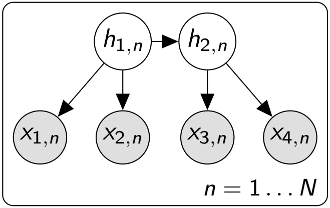
- Notice, h_1 has three conditionally independent "views". Thus, we can estimate $\mathbb{E}[x_1|h_1]$, $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_1]$.

Part 1: Estimating the conditional moments



- Notice, h_1 has three conditionally independent "views". Thus, we can estimate $\mathbb{E}[x_1|h_1]$, $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_1]$.
- h_2 has x_2, x_3, x_4. So, $\mathbb{E}[x_2|h_2]$, $\mathbb{E}[x_3|h_2]$ and $\mathbb{E}[x_4|h_2]$ are available.

Part 1: Estimating the conditional moments



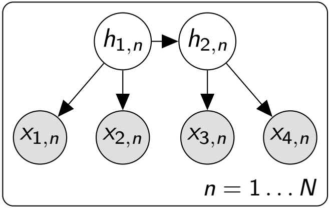
- Notice, h_1 has three conditionally independent "views". Thus, we can estimate $\mathbb{E}[x_1|h_1]$, $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_1]$.
- h_2 has x_2, x_3, x_4. So, $\mathbb{E}[x_2|h_2]$, $\mathbb{E}[x_3|h_2]$ and $\mathbb{E}[x_4|h_2]$ are available.

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{h_1} \sum_{h_2} p(h_1)p(h_2|h_1)\ \mathbb{E}[x_1|h_1]\mathbb{E}[x_2|h_1]\mathbb{E}[x_3|h_2]$$

$$= \sum_{h_1} p(h_1)\ \mathbb{E}[x_1|h_1]\mathbb{E}[x_2|h_1] \left(\sum_{h_2} p(h_2|h_1)\mathbb{E}[x_3|h_2, h_1] \right)$$

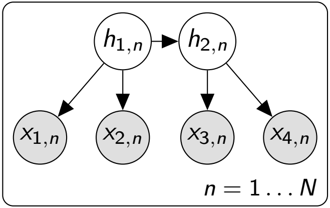
$$= \sum_{h_1} p(h_1)\ \mathbb{E}[x_1|h_1]\mathbb{E}[x_2|h_1]\mathbb{E}[x_3|h_1] \rightarrow \text{Right form for MoM!}$$

Part 2: Estimating the hidden potentials



- Given $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_2]$, estimating $p(h_2, h_1)$ is child's play.

Part 2: Estimating the hidden potentials

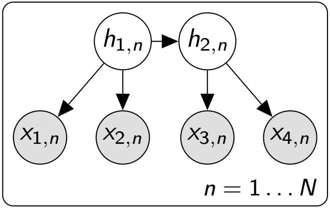


- Given $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_2]$, estimating $p(h_2, h_1)$ is child's play.

- For example,

$$\mathbb{E}[x_2 \otimes x_3] = \sum_{h_1,h_2} \mathbb{E}[x_2|h_1]p(h_2, h_1)\mathbb{E}[x_3|h_2]$$
$$= M_2 S M_3$$

Part 2: Estimating the hidden potentials



- Given $\mathbb{E}[x_2|h_1]$ and $\mathbb{E}[x_3|h_2]$, estimating $p(h_2, h_1)$ is child's play.

- For example,

$$\mathbb{E}[x_2 \otimes x_3] = \sum_{h_1, h_2} \mathbb{E}[x_2|h_1] p(h_2, h_1) \mathbb{E}[x_3|h_2]$$
$$= M_2 S M_3$$

- One way to do it is convex optimization.

Part 2: Estimating Hidden Potentials

- If we choose to, $\min_S \|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\|_F$, then the solution is $\widehat{S} = M_2^{\dagger}\mathbb{E}[x_2 \otimes x_3]M_3^{\dagger}$. (This is the first thing they do in the paper)

Part 2: Estimating Hidden Potentials

- If we choose to, $\min_S \|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\|_F$, then the solution is $\widehat{S} = M_2^\dagger \mathbb{E}[x_2 \otimes x_3] M_3^\dagger$. (This is the first thing they do in the paper)
- Or better we can do,

$$\min_S \left\|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\right\|_F$$
$$S \geq 0$$
$$1^T S 1 = 1$$

Part 2: Estimating Hidden Potentials

- If we choose to, $\min_S \ \|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\|_F$, then the solution is $\widehat{S} = M_2^\dagger \mathbb{E}[x_2 \otimes x_3] M_3^\dagger$. (This is the first thing they do in the paper)
- Or better we can do,

$$\min_S \left\|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\right\|_F$$
$$S \geq 0$$
$$1^T S 1 = 1$$

(I use CVX! haha!)

Part 2: Estimating Hidden Potentials

- If we choose to, $\min_S \ \|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\|_F$, then the solution is $\widehat{S} = M_2^{\dagger}\mathbb{E}[x_2 \otimes x_3]M_3^{\dagger}$. (This is the first thing they do in the paper)
- Or better we can do,

$$\min_S \left\|\mathbb{E}[x_2 \otimes x_3] - M_2 S M_3\right\|_F$$
$$S \geq 0$$
$$1^T S 1 = 1$$

(I use CVX! haha!)

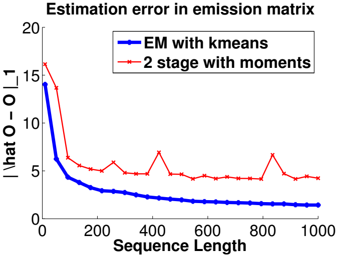
- Or even better (so they claim),

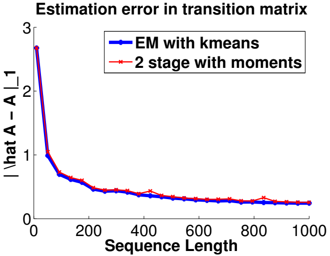
$$\max_S \mathbb{E}[\log p(x_2, x_3)]$$

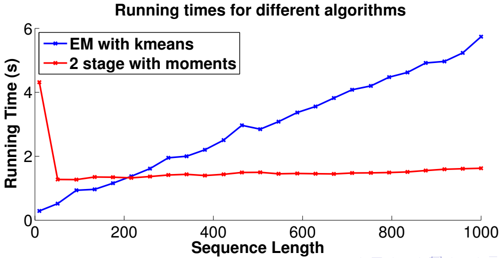
This is called the "Composite Likelihood"

A simulation for computational and statistical efficiency

- Statistical and computational efficiencies of the two stage estimation and EM for HMM with Gaussian observations. $(K = 5)$





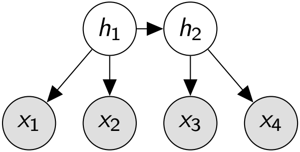


Conditions for recoverability of a Directed Graphical Model

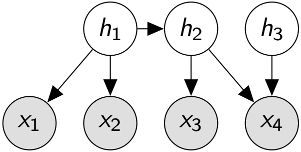
- We need to be able to recover all conditional expectations:
  - Every hidden node must be a "bottleneck" in the worst case.
  - There must be at least three cond. indep. variables for a node to be a bottleneck.
  - The conditional expectation matrices have to have full column rank.

Conditions for recoverability of a Directed Graphical Model

- We need to be able to recover all conditional expectations:
  - Every hidden node must be a “bottleneck” in the worst case.
  - There must be at least three cond. indep. variables for a node to be a bottleneck.
  - The conditional expectation matrices have to have full column rank.
- Examples:



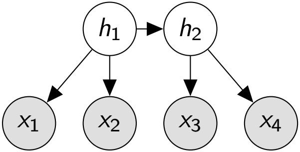
PASS



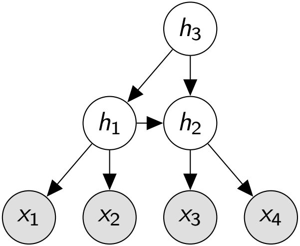
FAIL

Conditions for recoverability of a Directed Graphical Model

- Hidden nodes must possess the "Exclusive Views" property.
  - A hidden node has to have at least one conditionally independent observation on its own to have this property.
- If we want to estimate all hidden potentials:



PASS



FAIL

Part 3: Undirected Graphs (MRFs)

- The joint distribution is defined with clique "potentials".

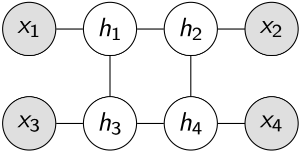
$$p(h_{1:K}, x_{1:J}|\theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{G}} \exp(\theta^T \phi(x_C, h_C))$$

Part 3: Undirected Graphs (MRFs)

- The joint distribution is defined with clique "potentials".

$$p(h_{1:K}, x_{1:J}|\theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{G}} \exp(\theta^T \phi(x_C, h_C))$$

- Example: (An image segmentation model)



$$\begin{aligned}\phi(x_C, h_C) &= \phi_1(h_i, h_{\mathcal{N}(i)}) + \phi_2(x_i, h_i) \\ &= \theta_1 \mathbf{1}_{[h_i = h_{\mathcal{N}(i)}]} + \theta_2 \mathbf{1}_{[h_i \neq h_{\mathcal{N}(i)}]} \\ &\quad + \sum_{l,k} \theta_{3,i,k} \mathbf{1}_{[x_i = l][h_i = k]}\end{aligned}$$

$$Z(\theta) = \int \prod_{C \in \mathcal{G}} \exp(\theta^T \phi(x_C, h_C)) dx_{1:J} dh_{1:K}$$

The notorious partition function!

Part 3: Learning Undirected Graphs

- The lower bound on likelihood is:

$$\log p(x_{1:K}|\theta) \geq \mathbb{E}_{p(x_{1:J}, h_{1:K}|\theta)}[\log p(x_{1:J}, h_{1:K}|\theta)] = \mathcal{L}(\theta)$$

Part 3: Learning Undirected Graphs

- The lower bound on likelihood is:

$$\log p(x_{1:K}|\theta) \geq \mathbb{E}_{p(x_{1:J},h_{1:K}|\theta)}[\log p(x_{1:J}, h_{1:K}|\theta)] = \mathcal{L}(\theta)$$

- Computing $p(x_{1:J}, h_{1:K}|\theta)$ is not trivial in general graphs. But approximations are made in practice. (e.g. Loopy Belief Propagation)

Part 3: Learning Undirected Graphs

- The lower bound on likelihood is:

$$\log p(x_{1:K}|\theta) \geq \mathbb{E}_{p(x_{1:J},h_{1:K}|\theta)}[\log p(x_{1:J}, h_{1:K}|\theta)] = \mathcal{L}(\theta)$$

- Computing $p(x_{1:J}, h_{1:K}|\theta)$ is not trivial in general graphs. But approximations are made in practice. (e.g. Loopy Belief Propagation)
- With MoM, we can estimate $p(x_{1:J}, h_{1:K}|\theta)$ from data.

$$\mathcal{L}(\theta) = \theta^T \left(\sum_{C \in \mathcal{G}} \mathbb{E}[\phi(x_{1:J}, h_{1:K})] \right) - A(\theta)$$

$$\textit{where}, \quad \mathbb{E}[\phi(x_{1:J}, h_{1:K})] = \sum_{x_{1:J}, h_{1:K}} p(x_{1:J}, h_{1:K})\phi(x_{1:J}, h_{1:K})$$

Part 3: Learning Undirected Graphs

- The lower bound on likelihood is:

$$\log p(x_{1:K}|\theta) \geq \mathbb{E}_{p(x_{1:J},h_{1:K}|\theta)}[\log p(x_{1:J}, h_{1:K}|\theta)] = \mathcal{L}(\theta)$$

- Computing $p(x_{1:J}, h_{1:K}|\theta)$ is not trivial in general graphs. But approximations are made in practice. (e.g. Loopy Belief Propagation)
- With MoM, we can estimate $p(x_{1:J}, h_{1:K}|\theta)$ from data.

$$\mathcal{L}(\theta) = \theta^T \left(\sum_{C \in \mathcal{G}} \mathbb{E}[\phi(x_{1:J}, h_{1:K})] \right) - A(\theta)$$

$$where, \quad \mathbb{E}[\phi(x_{1:J}, h_{1:K})] = \sum_{x_{1:J},h_{1:K}} p(x_{1:J}, h_{1:K})\phi(x_{1:J}, h_{1:K})$$

- So, the MoM lower bound is concave w.r.t. θ.

Conclusions

- It's a good paper, that opens new possibilities for MoM learning.
- The moral of the story: MoM and likelihood maximization can be used synergistically to learn a variety of models.
- The story isn't finished yet: Models like MHMM is not covered. (where not all variables are bottlenecks.)
- Experimental verification is necessary as follow-up work.

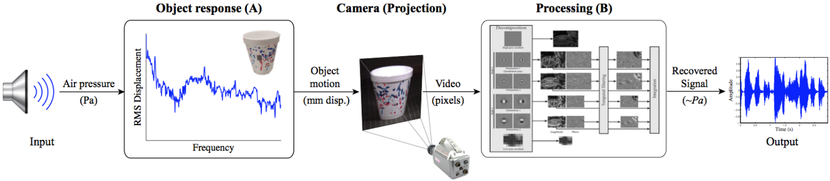
Outline

Introduction

- The goal: Recovering sound from video.

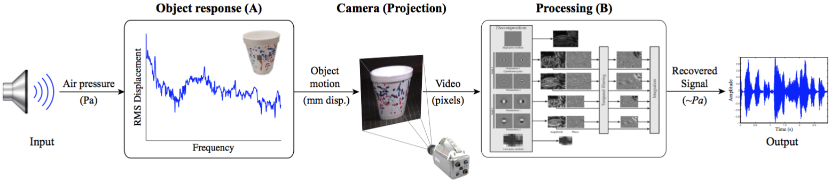
Introduction

- The goal: Recovering sound from video.
- Sound waves cause minute vibrations on objects. High speed camera footage of these vibrations are used to reconstruct the sound.



Introduction

- The goal: Recovering sound from video.
- Sound waves cause minute vibrations on objects. High speed camera footage of these vibrations are used to reconstruct the sound.



- We'll mostly be interested in "Processing" step, which is somewhat involved in signal processing/vision.

Outline

Recovering Sound from Video : Local Motion Signal

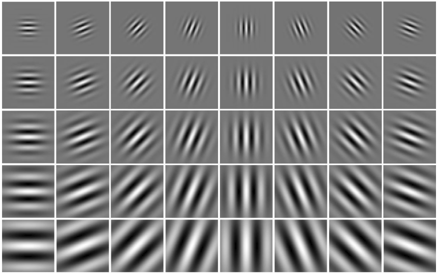
- The first step: A wavelet transform (steerable pyramid representation) of the video for every frame:

Recovering Sound from Video : Local Motion Signal

- The first step: A wavelet transform (steerable pyramid representation) of the video for every frame:
  - It is a filter bank consisting of sombrero type of filters with different orientations and scales.

Recovering Sound from Video : Local Motion Signal

- The first step: A wavelet transform (steerable pyramid representation) of the video for every frame:
  - It is a filter bank consisting of sombrero type of filters with different orientations and scales.
- A similar filter bank, Gabor Wavelets (real parts) for several scales (r): and orientations θ:



- In 1D, it's of form $f(x; \sigma^2, \omega) = \mathcal{N}(x, 0, \sigma^2)e^{j2\pi\omega x}$

Recovering Sound from Video : Local Motion Signal

- After wavelet transform, we have:

$$\mathcal{W}(V) = \underbrace{A(r,\theta,x,y,t)}_{amplitude} e^{j\overbrace{\psi(r,\theta,x,y,t)}^{phase}}$$

- This is a phasor representation, $A(.)$ is the amplitude and $\psi(.)$ is the phase.

Recovering Sound from Video : Local Motion Signal

- After wavelet transform, we have:

$$\mathcal{W}(V) = \underbrace{A(r, \theta, x, y, t)}_{amplitude} e^{j \overbrace{\psi(r, \theta, x, y, t)}^{phase}}$$

- This is a phasor representation, $A(.)$ is the amplitude and $\psi(.)$ is the phase.
- Then phase variations wrt. to a reference frame t_0 is computed $\psi_v(., t) = \psi(., t) - \psi(., t_0)$.
  - For small motions these variations

This is the local motion signal.

Recovering Sound from Video : Global Motion Signal

- The output of this stage is the reconstruction!.
- First average over the spatial coordinates:

$$\Phi(r, \theta, t) = \sum_{x,y} A(r, \theta, x, y, t)^2 \psi_v(r, \theta, x, y, t)$$

Recovering Sound from Video : Global Motion Signal

- The output of this stage is the reconstruction!.
- First average over the spatial coordinates:

$$\Phi(r,\theta,t) = \sum_{x,y} A(r,\theta,x,y,t)^2 \psi_v(r,\theta,x,y,t)$$

- Then align the signals:

$$t_i = \arg\max_{t_i} \Phi(r_0,\theta_0,t)^T \Phi(r_i,\theta_i,t-t_i)$$

Recovering Sound from Video : Global Motion Signal

- The output of this stage is the reconstruction!.
- First average over the spatial coordinates:

$$\Phi(r,\theta,t) = \sum_{x,y} A(r,\theta,x,y,t)^2 \psi_v(r,\theta,x,y,t)$$

- Then align the signals:

$$t_i = \arg\max_{t_i} \Phi(r_0,\theta_0,t)^T \Phi(r_i,\theta_i,t-t_i)$$

- The reconstructed signal is:

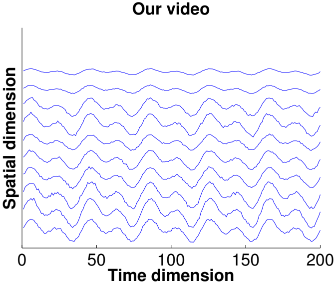
$$\widehat{s}(t) = \sum_i \Phi(r_i,\theta_i,t-t_i)$$

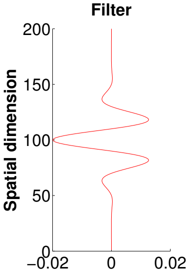
Say we have the following video..

(Loading Video...)

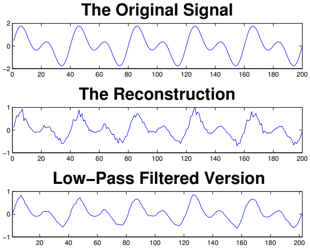
- Can we reconstruct a sound?

Yes we can!





- Original signal is $x(t) = \sin(2\pi\omega_0 t) + \sin(2\pi 2\omega_0 t)$
- I corrupt the original signal in each dimension with $a\, x(t+\theta) + \epsilon$, $a \sim \mathcal{U}([0 \;\; 1])$, $\theta, \epsilon \sim \mathcal{N}(0, 0, 1)$



A glance at their experiments

- Objects behave like low-pass filters. It's harder to obtain high frequencies, as one would expect.
- For speech, their method generally works worse than an active method.
- They claim that unintelligible sound may also be useful for surveillance type applications.
- They have the vibration mode estimation application also.
- Limitations: Sampling rate / Magnification

Conclusions

- A (very) good paper with lots of experiments.
- I would have liked to see some theoretical justification for the processing step.
- Experiments are really good, and they provide several applications, and some analysis. It's definitely a well studied, exciting (even for me) paper.