

# **CONVEX:**

## **De novo Transcriptome Error Correction by Convexification**



**Meisam Razaviyayn**  
Stanford University



**Jeremy Guo**  
Stanford University

**David Tse**  
Stanford University - U.C.  
Berkeley

**Elizabeth Tseng**  
Pacific Biosciences

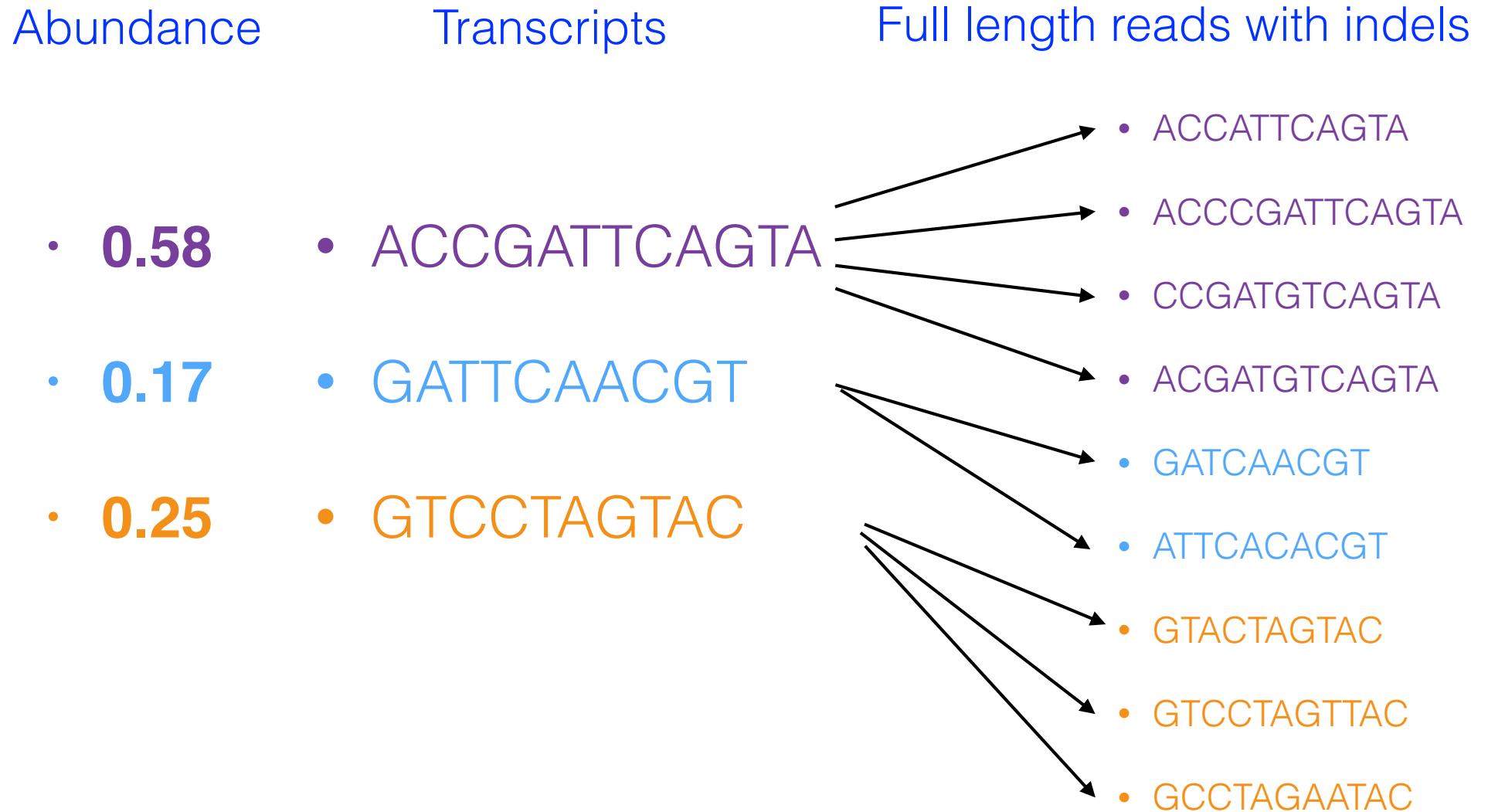
# Problem Statement

Abundance

Transcripts

- **0.58**     • ACCGATTCTAGTA
- **0.17**     • GATTCAACGT
- **0.25**     • GTCCTAGTAC

# Problem Statement



# Problem Statement

Abundance

- ?

- ?

- ?

Transcripts

- ????????????????

- ????????????

- ????????????????

Full length reads with indels

- ACCATTCA~~G~~T<sub>A</sub>
- ACCCGATTCA~~G~~T<sub>A</sub>
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATT~~C~~ACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

# Convexification

Abundance

- ?
- ?
- ?

Transcripts

- ????????????????
- ??????????????
- ????????????????

Full length reads with indels

- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Maximum Likelihood Estimation

**Nonconvex and NP-hard**

# Convexification

Abundance

Transcripts

Full length reads with indels

- ?
- ?
- ?

- ACCGATTTCAGTA
- GATTCAACGT
- GTCCTAGTAC

- ACCATTTCAGTA
- ACCCGATTTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTCACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Maximum Likelihood Estimation

**Convex and Easy**

# Convexification

Abundance

Transcripts

Full length reads with indels

- ?
- ?
- ?

- ACCGATTTCAGTA
- GATTCAACGT
- GTCCTAGTAC

- ACCATTTCAGTA
- ACCCGATTTCAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GTACTAGTAC
- GTCCTAGTTAC
- GCCTAGAATAC

Maximum Likelihood Estimation

**Convex and Easy**

Set of all possible sequences? **Convex but exponentially many**

### Prefixes

- AAA

- AAC

- •  
•

- ACC

- ACG

- •  
•

- GAG

- GAT

- GCA

- •  
•

- GTA

- GTC

- •  
•

# Greedy Algorithm

- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

## Abundance Prefixes

• 0.04	• AAA
• 0.05	• AAC
•	•
•	•
•	•
• 0.2	• ACC
• 0.01	• ACG
•	•
•	•
•	•
• 0.05	• GAG
• 0.13	• GAT
• 0.02	• GCA
•	•
•	•
•	•
• 0.05	• GTA
• 0.15	• GTC
•	•
•	•
•	•

# Greedy Algorithm

- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

## Abundance Prefixes

• 0.04 • AAA

• 0.05 • AAC

• 0.2 • ACC

• 0.01 • ACG

• 0.05 • GAG

• 0.13 • GAT

• 0.02 • GCA

• 0.05 • GTA

• 0.15 • GTC

# Greedy Algorithm

## Prefixes

ACCA

ACCC

ACCG

ACCT

GATA

GATC

GATG

GATT

GTCA

GTCC

GTCG

GTCT

- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

## Abundance Prefixes

- 0.04 • AAA

- 0.05 • AAC

- 0.2 • ACC

- 0.01 • ACG

- 0.05 • GAG

- 0.13 • GAT

- 0.02 • GCA

- 0.05 • GTA

- 0.15 • GTC

# Greedy Algorithm

## Prefixes      Abundance

ACCA 0.07

ACCC 0.04

ACCG 0.3

ACCT 0.08

GATA 0.02

GATC 0.04

GATG 0.04

GATT 0.15

GTCA 0.06

GTCC 0.17

GTCG 0.03

GTCT 0.02

- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

# Greedy Algorithm

Prefixes	Abundance	
• <u>ACCA</u>	<u>0.07</u>	• ACCATTTCAGTA
• <u>ACCC</u>	<u>0.04</u>	• ACCCGATTTCAGTA
• <u>ACCG</u>	<u>0.3</u>	• CCGATGTCAGTA
• <u>ACCT</u>	<u>0.08</u>	• ACGATGTCAGTA
• <u>GATA</u>	<u>0.02</u>	• GATCAACGT
• <u>GATC</u>	<u>0.04</u>	• ATTACACACGT
• <u>GATG</u>	<u>0.04</u>	• GACTAGTAC
• <u>GATT</u>	<u>0.15</u>	• GCCTAGTTAC
• <u>GTCA</u>	<u>0.06</u>	• GACTAGAATAC
• <u>GTCC</u>	<u>0.17</u>	
• <u>GTCG</u>	<u>0.03</u>	
• <u>GTCT</u>	<u>0.02</u>	

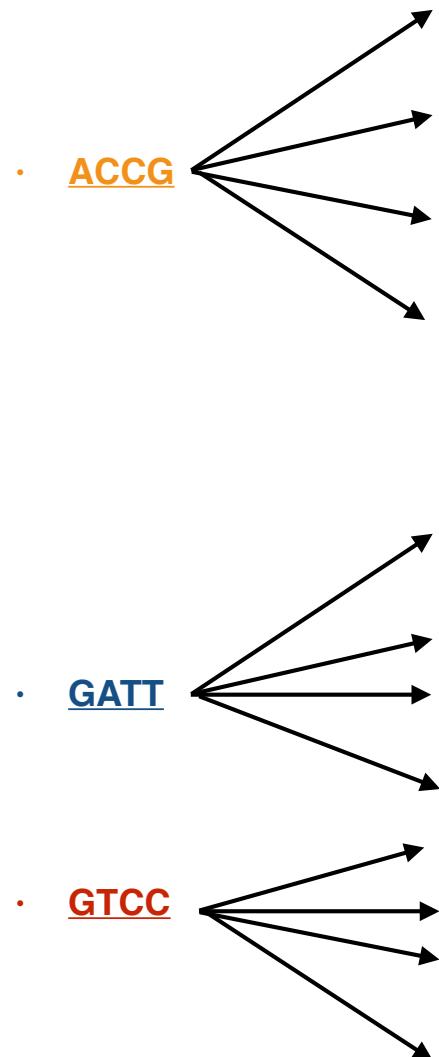
# Greedy Algorithm

Prefixes	Abundance	
· <u>ACCA</u>	<u>0.07</u>	· ACCATTTCAGTA
· <u>ACCC</u>	<u>0.04</u>	· ACCCGATTTCAGTA
· <b><u>ACCG</u></b>	<b><u>0.3</u></b>	· CCGATGTCAGTA
· <u>ACCT</u>	<u>0.08</u>	· ACGATGTCAGTA
· <u>GATA</u>	<u>0.02</u>	· GATCAACGT
· <u>GATC</u>	<u>0.04</u>	· ATTACACACGT
· <u>GATG</u>	<u>0.04</u>	· GACTAGTAC
· <b><u>GATT</u></b>	<b><u>0.15</u></b>	· GCCTAGTTAC
· <u>GTCA</u>	<u>0.06</u>	· GACTAGAATAC
· <b><u>GTCC</u></b>	<b><u>0.17</u></b>	
· <u>GTCG</u>	<u>0.03</u>	
· <u>GTCT</u>	<u>0.02</u>	

# Greedy Algorithm

- ACCG 0.3
  - ACCATTAGTA
  - ACCCGATTAGTA
  - CCGATGTCAGTA
  - ACGATGTCAGTA
  - GATCAACGT
  - ATTACACACGT
  - GACTAGTAC
  - GCCTAGTTAC
  - GACTAGAATAC
- GATT 0.15
- GTCC 0.17

# Greedy Algorithm



- ACCATTCAAGTA
- ACCCGATTCAAGTA
- CCGATGTCAGTA
- ACGATGTCAGTA
- GATCAACGT
- ATTACACACGT
- GACTAGTAC
- GCCTAGTTAC
- GACTAGAATAC

# Remarks

- Consistency
- 
-

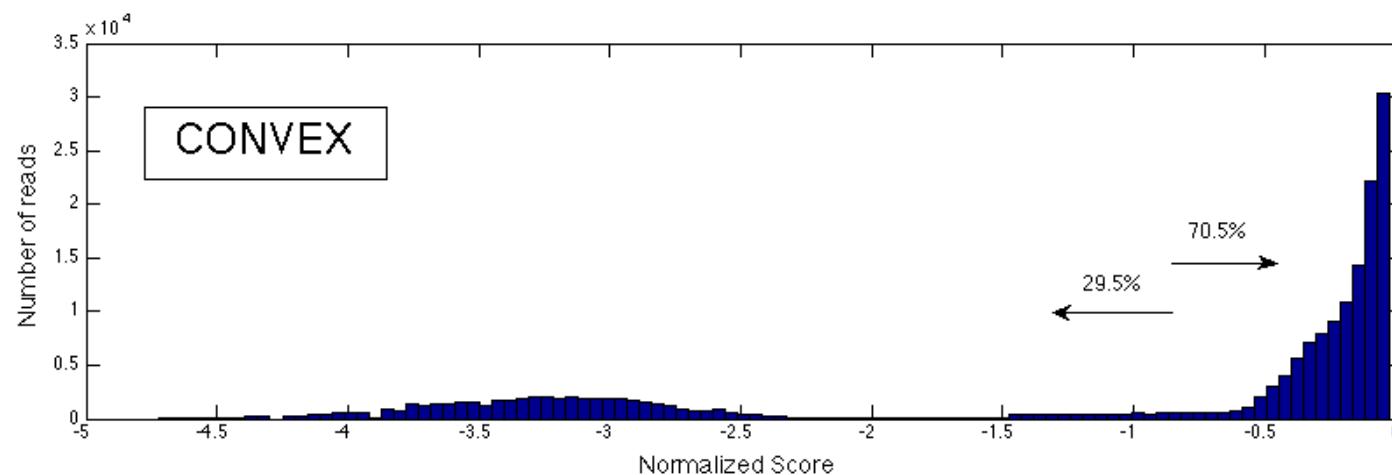
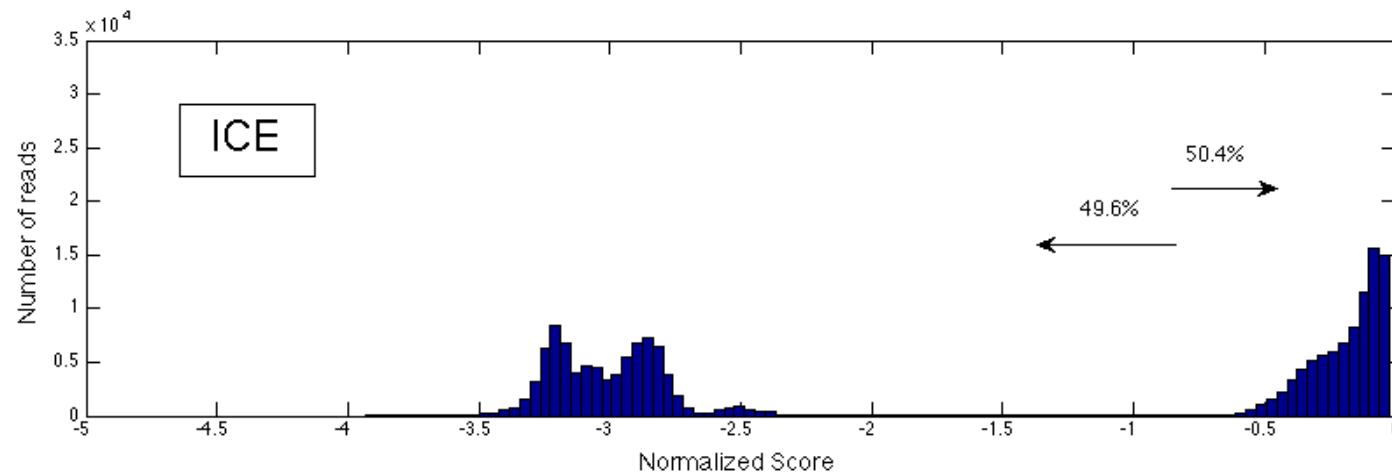
# Remarks

- Consistency
- No pairwise alignment between reads
-

# Remarks

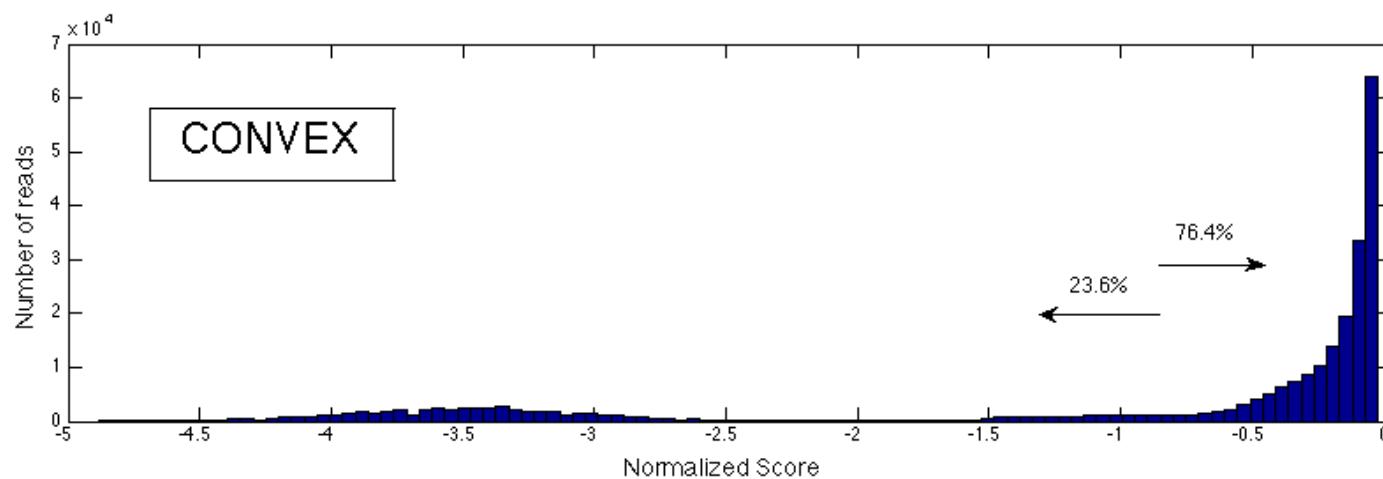
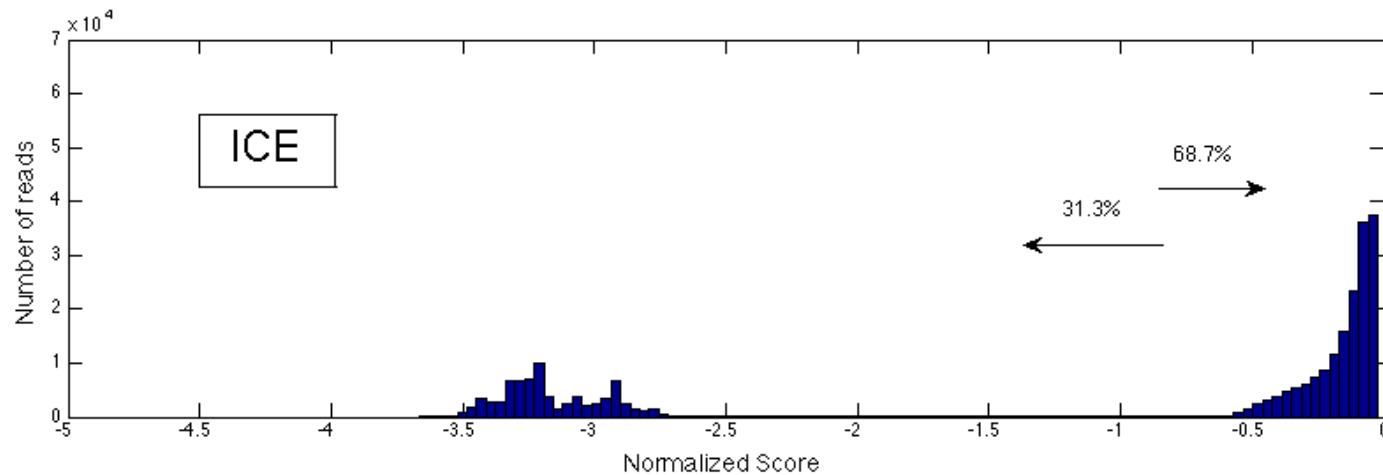
- Consistency
- No pairwise alignment between reads
- Linear computational complexity

# Numerical Experiments: Heart Tissue



Number of CCS FL Reads	173,566
Number of Transcripts	6,896
Average Length	2,761

# Numerical Experiments: Liver Tissue



Number of CCS FL Reads	245,381
Number of Transcripts	6,124
Average Length	2,198

Questions?

# Estimating Abundances: Mixture Model

- Isoforms/centroids  $s_1, s_2, \dots, s_M$
- Abundances  $\rho_1, \rho_2, \dots, \rho_M$
- Full length reads  $r_1, r_2, \dots, r_N$
- Mixture model:  
$$\mathbb{P}(r_n; \mathbf{s}, \boldsymbol{\rho}) = \sum_{m=1}^M \rho_m \mathbb{P}(r_n \mid r_n \text{ from } s_m)$$
- Maximum likelihood estimation:  
$$\arg \max_{\boldsymbol{\rho}} \sum_{n=1}^N \log \left( \sum_{m=1}^M \rho_m \mathbb{P}(r_n \mid r_n \text{ from } s_m) \right)$$