Improved Naïve Bayes with Mislabeled Data

Qianhan Zeng

Based on joint work with Yingqiu Zhu*, Xuening Zhu, Feifei Wang,

Weichen Zhao, Shuning Sun, Meng Su, and Hansheng Wang

Introduction

Figure 2. Front camera images collected from clear weather (col 1), nighttime (col 2), rain (col 3) and construction zones (col 4).

ImageNet Dataset

1.2 million training images

(Russakovsky et al., 2015)

nuScenes Dataset

1.4 million images

(Caesar et al., 2020)

Introduction

- **Incorrect labels have been found among many widely used datasets.**
  - ImageNet Dataset: 0.3% incorrect labels
  - QuickDraw Dataset: 10% incorrect labels
  - Amazon Reviews Dataset: 3.9% incorrect labels (Northcutt et al., 2021) https://labelerrors.com/

CIFAR-10 given label:
cat

Cleanlab guessed: **frog**

MTurk consensus: **frog**

ID: 2405

ImageNet given label:
red panda

Cleanlab guessed: **giant panda**

MTurk consensus: **giant panda**

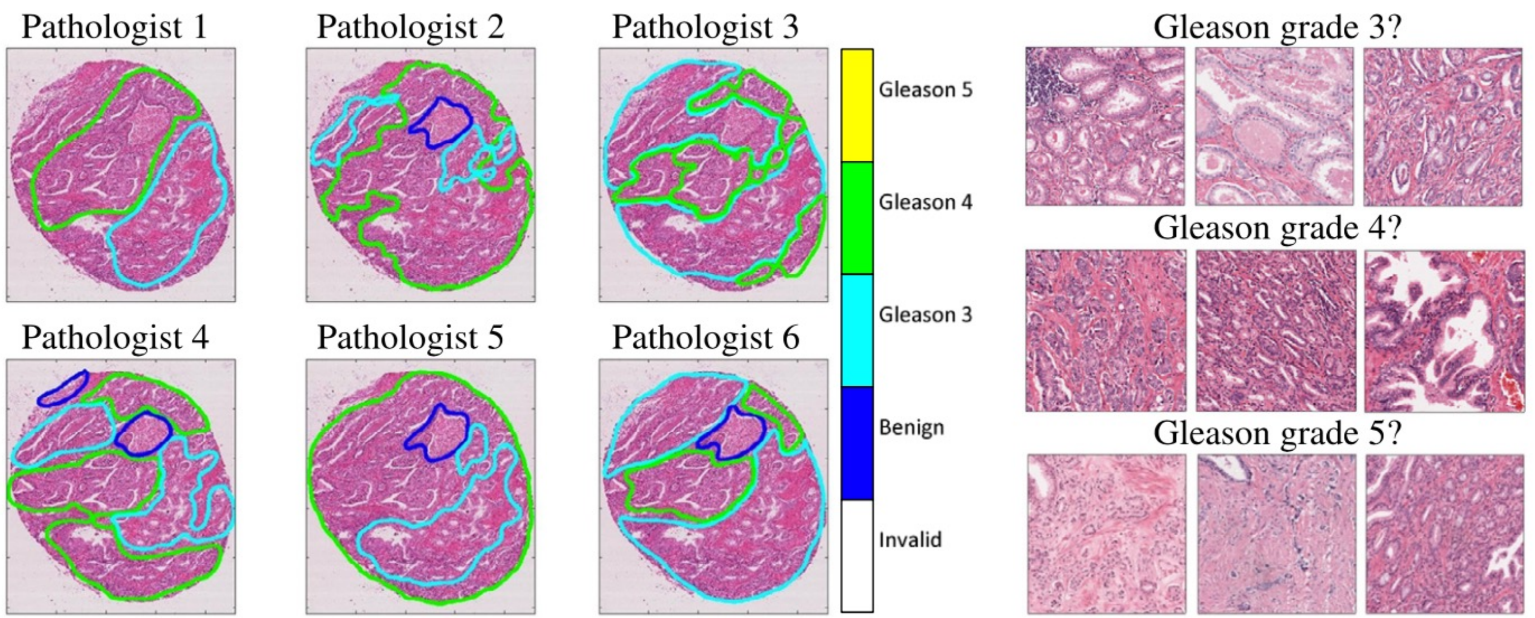
ID: 00031356

因思想而光华

Introduction

➢ **Possible causes of the incorrect labels:**

1. Subjective criteria (e.g., medical diagnosis)



(a) Inter-observer variability (pathologists are not in alphabetical order)

(b) Intra-class variability and inter-class similarity

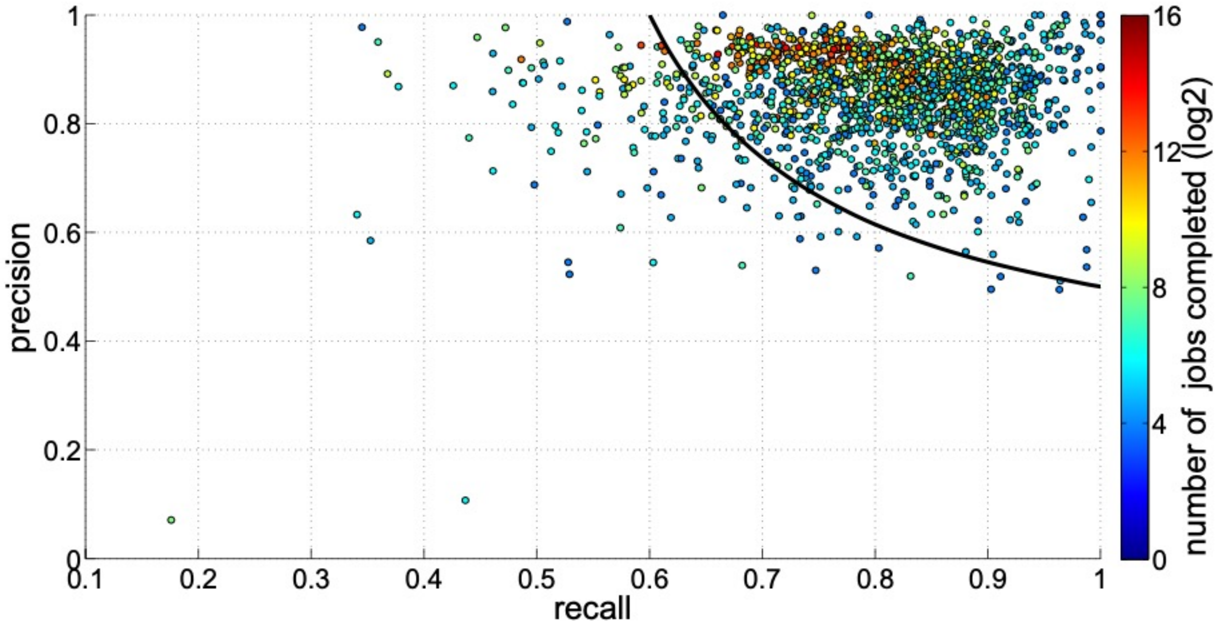
Fig. 1. Heterogeneity of PCa patterns and grading lead to classification challenges.

(Nir et al., 2018)

Introduction

➢ **Possible causes of the incorrect labels:**

1. Subjective criteria (e.g., medical diagnosis).
2. Practice makes perfect.



The precision and recall of workers on category labeling, with color indicating how many jobs they completed.

(Lin et al., 2014)

Introduction

- **Possible causes of the incorrect labels:**

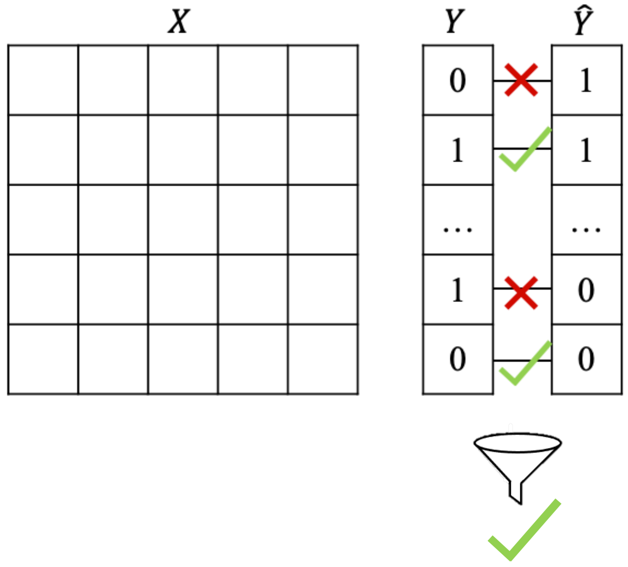
  1. Subjective criteria (e.g., medical diagnosis).
  2. Practice makes perfect.
  3. Professional knowledge.
  4. ……

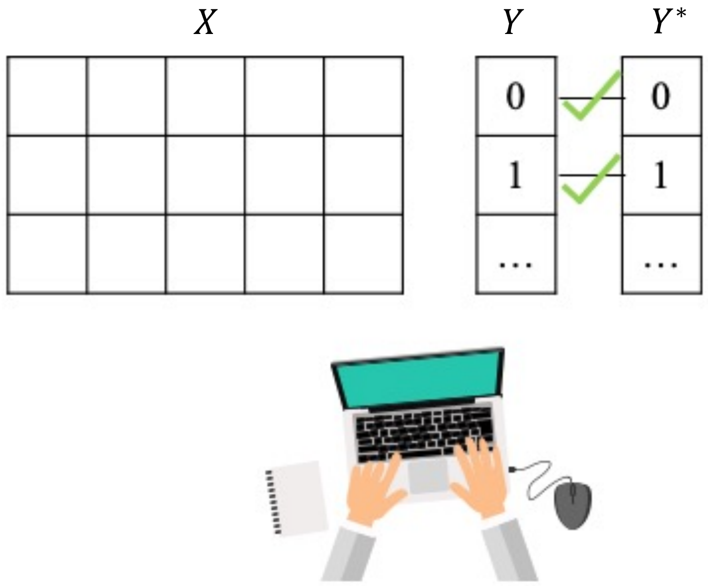
Similar birds
(Lin et al., 2014)

Literature Review

- Noise Filtering Method

1. Decision tree, k-nearest neighbor classifiers, and linear machines (Brodley & Friedl, 1999) – JAIR
2. Naïve Bayes (Farid et al., 2014) – Expert systems with applications
3. MentorNet (Jiang et al., 2018) – ICML

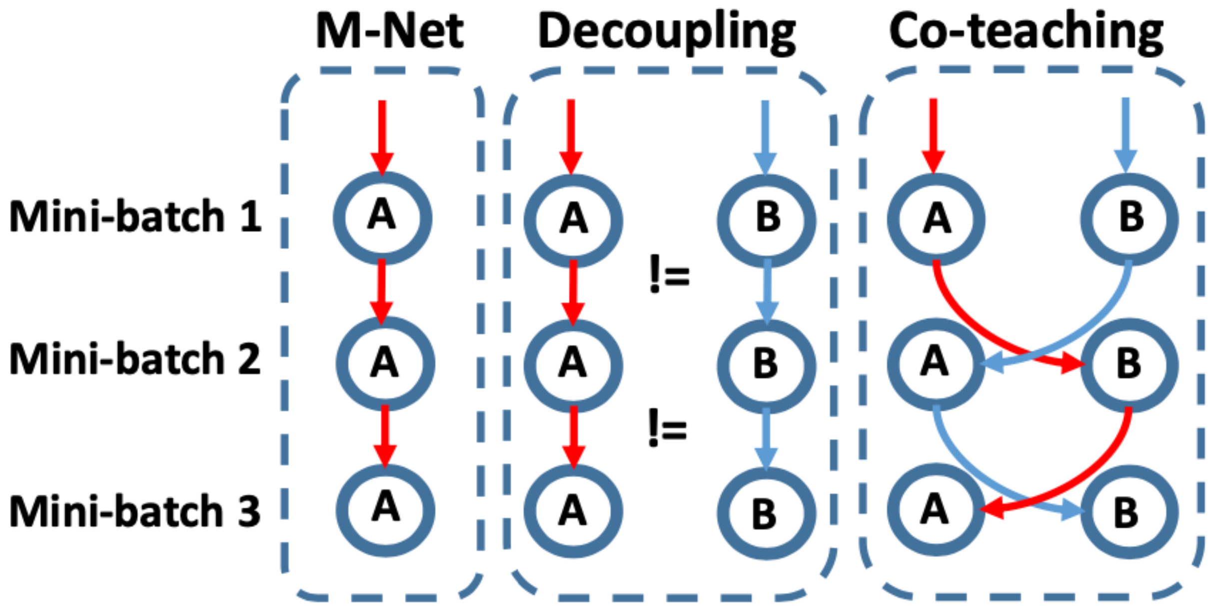




Literature Review

- Modified Model Architecture

  1. BayesANIL (Ramakrishnan et al., 2005) – ICML
  2. Decoupling (Malach & Shalev-Shwartz, 2017) – arXiv
  3. Co-teaching (Han et al., 2018) – NeurIPS



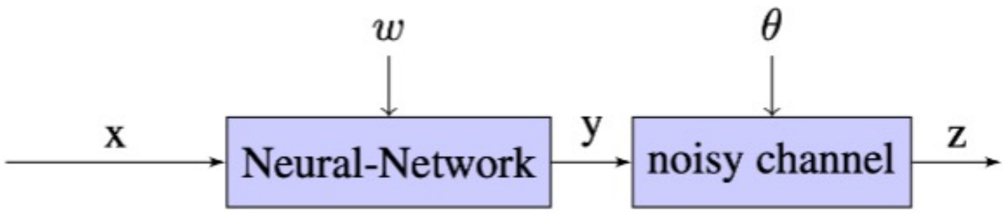
(Han et al., 2018) – NeurIPS

Literature Review

- Modified Model Architecture

4. Noisy Labels Neural-Network (NLNN) algorithm (Bekker & Goldberger, 2016) - ICASSP

   Noisy channel: $\theta(i,j) = p(z = j|y = i)$



feature vector x correct label y noisy label z

Literature Review

- Modified Model Architecture

  5. NLNN + a Noise Adaptation Layer (Goldberger & Ben-Reuven, 2017) - ICLR

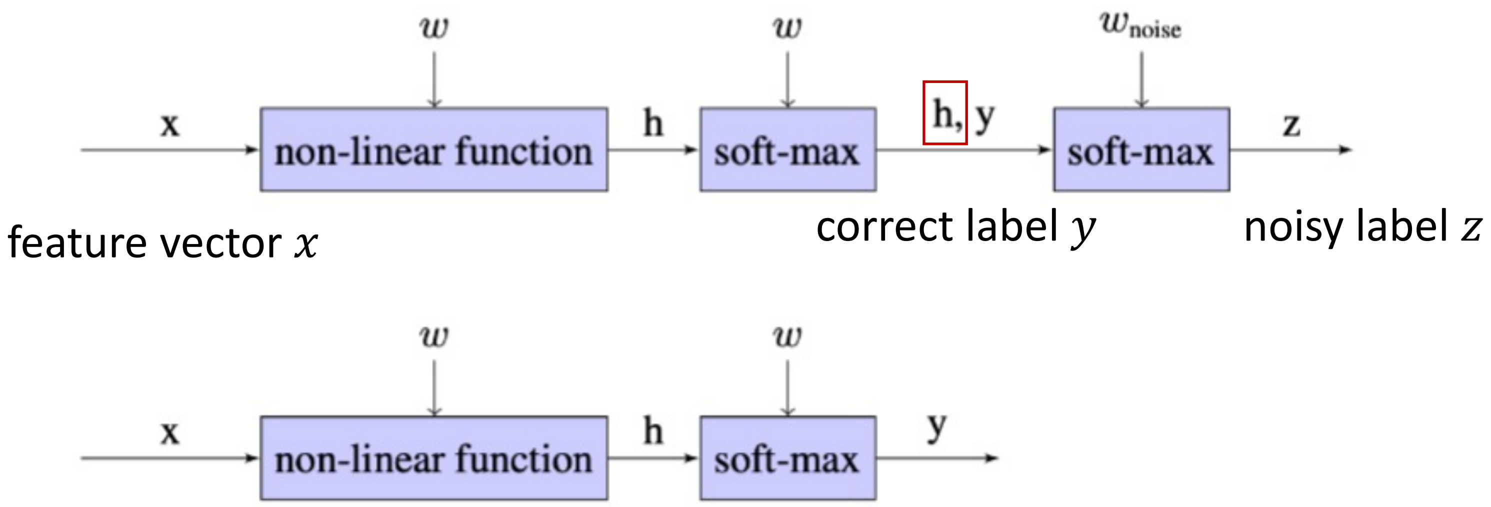
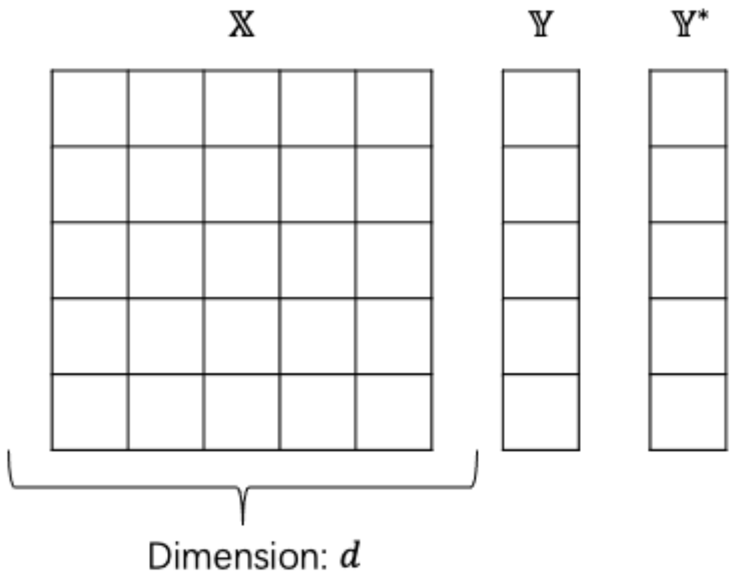


Figure 1: An illustration of the noisy-label neural network architecture for the training phase (above) and test phase (below).

Notations

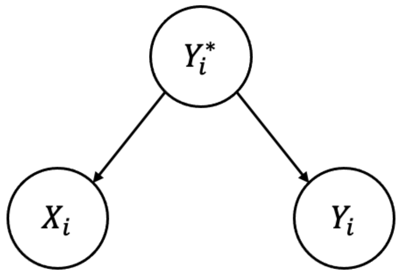
- Instances: $\mathbb{X} = \{X_1, \dots, X_N\}$ with $X_i = (X_{i1}, \dots, X_{id})^{\mathrm{T}}$. Each $X_{ij} \in \{0,1\}$.
- Observed labels: $\mathbb{Y} = \{Y_1, \dots, Y_N\}$. Each $Y_i \in \{1, \dots, K\}$.
- True labels: $\mathbb{Y}^* = \{Y_1^*, \dots, Y_N^*\}$.
- The probability of true class being class k: $\pi_k = P(Y_i^* = k)$.
- The probability of the jth feature being 1 in class k: $p_{jk} = P\left(X_{ij} = 1 \middle| Y_i^* = k\right)$.
- Total parameter set: θ.



Mislabeling Mechanism

- Data Generating Assumption
  - Class-conditional noise (Patrini et al., 2017; Zhang et al., 2021)

$$P(Y_i|Y_i^*, X_i) = P(Y_i|Y_i^*)$$



  - Mislabeling probability matrix: $P(Y_i = k_1|Y_i^* = k_2) = \rho_{k_1k_2}$

$$\begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1} & \rho_{K2} & \cdots & \rho_{KK} \end{pmatrix},$$

with $\sum_{k_1=1}^{K} \rho_{k_1k_2} = 1$ for $1 \leq k_2 \leq K$.

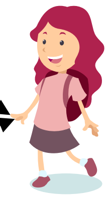
Mislabeling Impact

- Uniform label noise (Frenay et al., 2014)

$$\begin{pmatrix} \rho & \frac{1-\rho}{K-1} & \cdots & \frac{1-\rho}{K-1} \\ \frac{1-\rho}{K-1} & \rho & \cdots & \frac{1-\rho}{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-\rho}{K-1} & \frac{1-\rho}{K-1} & \cdots & \rho \end{pmatrix},$$

  - $K = 2$

$$\boxed{p_{j1} > p_{j2}} \rightarrow \begin{cases} \boxed{P\left(Y_i^* = 1 \middle| X_{ij} = 1\right) > P(Y_i^* = 2|X_{ij} = 1)} \\ \boxed{P\left(Y_i = 1 \middle| X_{ij} = 1\right) > P(Y_i = 2|X_{ij} = 1)} \end{cases}$$



  - $K > 2$: similar results

Mislabeling Impact

- Varying mislabeling probability: $\rho_{k_1 k_2} = P(Y_i = k_1 | Y_i^* = k_2)$

$$\begin{array}{ll} \text{Class 1} & \\ & \end{array} \begin{pmatrix} \rho & \boxed{1-\rho} & \boxed{1-\rho} & \cdots & \boxed{1-\rho} \\ 1-\rho & \rho & 0 & \cdots & 0 \\ 0 & 0 & \rho & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \rho \end{pmatrix},$$

- Assume $0.9 \le \rho < 1, K \gg 11.$

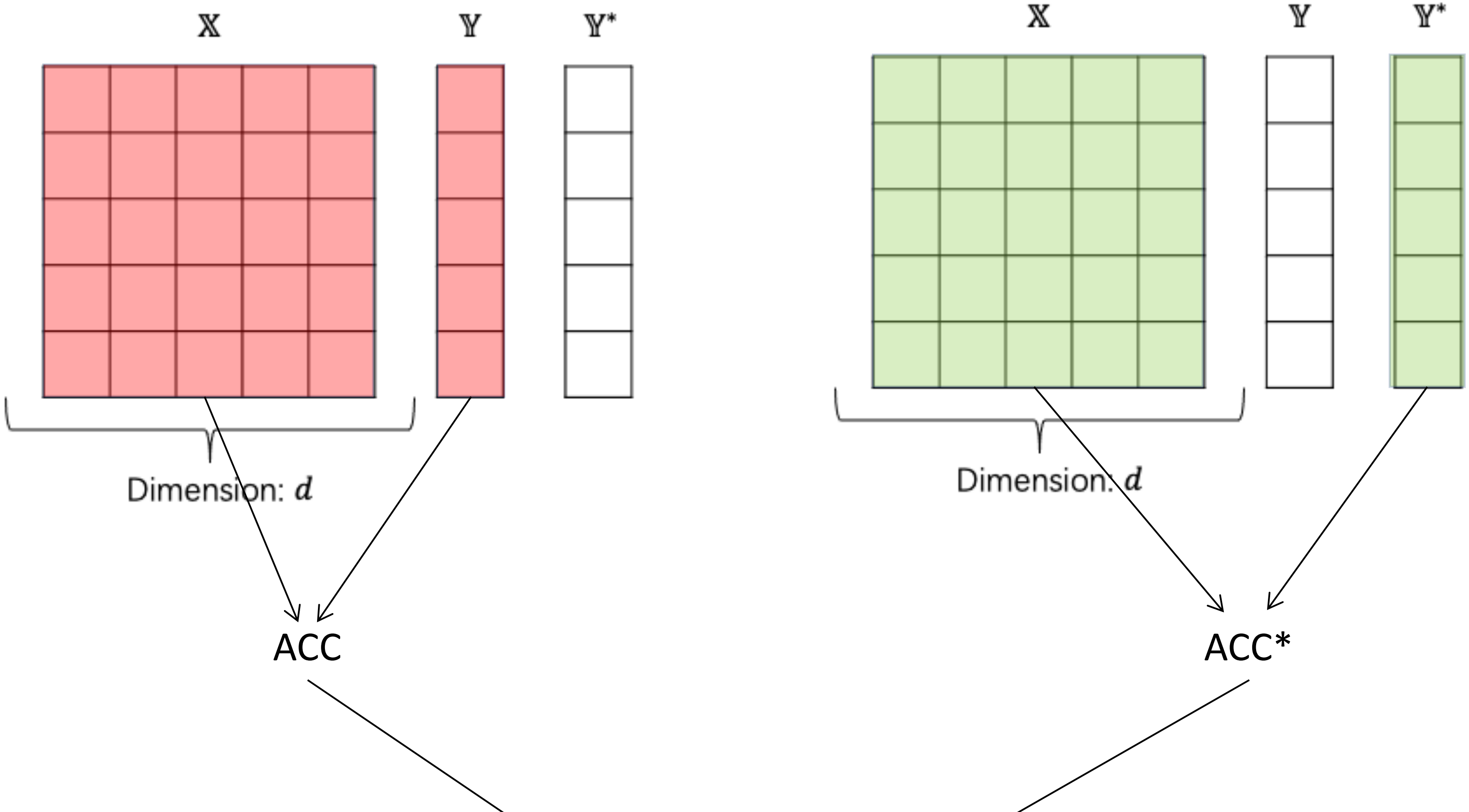
$\boxed{p_{j1} \ll p_{jk} \text{ for } k \ge 2}$ → $\boxed{P\left(Y_i^* = 1 \middle| X_{ij} = 1\right) < P(Y_i^* = k | X_{ij} = 1)}$

$\boxed{p_{j1} \ll p_{jk} \text{ for } k \ge 2}$ → $\boxed{P\left(Y_i = 1 \middle| X_{ij} = 1\right) > P(Y_i = k | X_{ij} = 1)}$

for $k \ge 2$

Mislabeling Impact

- Evaluation of the mislabeling impact:



$$\Delta\mathrm{ACC} = \mathrm{ACC} - \mathrm{ACC}^*$$

Log-Likelihood Function

➢ Log-likelihood function:

$$\ell(\theta) = \ln P(\mathbb{X}, \mathbb{Y}, \mathbb{Y}^* | \theta)$$

$$= \sum_{i=1}^{N} \left\{ \ln P(Y_i^* | \theta) + \ln P(Y_i | Y_i^*, \theta) + \sum_{j=1}^{d} \ln P(X_{ij} | Y_i^*, \theta) \right\}$$

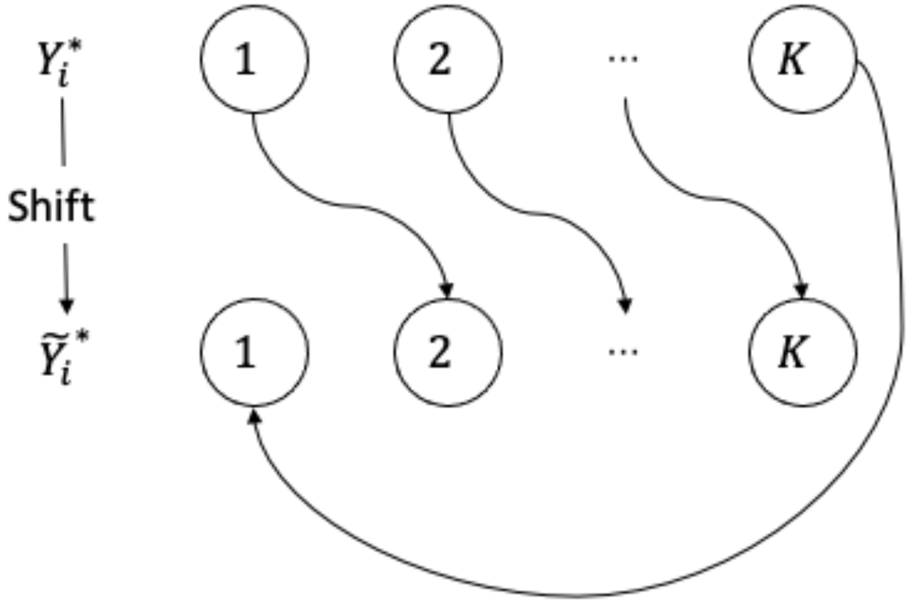
$$= \sum_{i=1}^{N} \ln \boxed{\pi_{Y_i^*}} + \sum_{i=1}^{N} \ln \boxed{\rho_{Y_i Y_i^*}} + \sum_{i=1}^{N} \sum_{j=1}^{d} X_{ij} \ln \boxed{p_{jY_i^*}}$$

$$+ \sum_{i=1}^{N} \sum_{j=1}^{d} (1 - X_{ij}) \ln (1 - \boxed{p_{jY_i^*}}). \quad (2)$$

Y_i^* is latent!

Identifiability Issue

- A shift case:



$$\begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1K} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1} & \rho_{K2} & \cdots & \rho_{KK} \end{pmatrix},$$

- $\ell(\theta) = \tilde{\ell}(\tilde{\theta})$

- Assumption: ρ_{kk} is larger than the off-diagonal elements.

EM Algorithm → INB Algorithm

- **E step:**
  - Denote $\hat{\gamma}_{ik}^{(t)} = P\left(Y_i^* = k \middle| X_i, Y_i, \hat{\theta}^{(t)}\right)$.

$$\hat{\gamma}_{ik}^{(t)} = \frac{\hat{\pi}_k^{(t)} \hat{\rho}_{Y_i k}^{(t)} \prod_{j=1}^{d} \hat{p}_{jk}^{(t) X_{ij}} \left\{1 - \hat{p}_{jk}^{(t)}\right\}^{1-X_{ij}}}{\sum_{k=1}^{K} \hat{\pi}_k^{(t)} \hat{\rho}_{Y_i k}^{(t)} \prod_{j=1}^{d} \hat{p}_{jk}^{(t) X_{ij}} \left\{1 - \hat{p}_{jk}^{(t)}\right\}^{1-X_{ij}}}$$

Update $\hat{\gamma}_{ik}^{(t)}$.

- **M step:**

$$\widehat{\pi}_k^{(t+1)} = \sum_{i=1}^{N} \widehat{\gamma}_{ik}^{(t)} / N, \ 1 \le k \le K,$$

$$\widehat{p}_{jk}^{(t+1)} = \left(\sum_{i=1}^{N} X_{ij} \widehat{\gamma}_{ik}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N} \widehat{\gamma}_{ik}^{(t)}\right), \ 1 \le j \le d, \ 1 \le k \le K,$$

$$\widehat{\rho}_{k_1 k_2}^{(t+1)} = \left(\sum_{i=1}^{N} I(Y_i = k_1) \widehat{\gamma}_{ik_2}^{(t)}\right) \Big/ \left(\sum_{i=1}^{N} \widehat{\gamma}_{ik_2}^{(t)}\right), \ 1 \le k_1, k_2 \le K.$$

Update estimators.

Simulation Experiments

- **Setups:**
  - X_i's dimension: $d = 500$.
  - Number of classes: $K = 5$.
  - Size of data: $n = 500, 1000$, and 5000. 80% in the training set and 20% in the testing set.
  - Prior probability: $\pi_k = 1/K$.
  - The probability of $X_{ij} = 1$: $p_{jk} \sim [0, 0.1) + \mathcal{N}(0.65,\ 0.06^2)$
  - Mislabeling Probability matrix ρ_{kk}: uniformly generated from an interval
  - $B = 100$.
- Baseline methods:

  1. Naïve Bayes (NB) model

  2. NLNN method of (Bekker et al., 2016); 3. NAL method of (Goldberger et al., 2017)

  4. NB-T

Simulation Performances

Table 1. Finite sample performances of different methods with different ρ_{kk} intervals and sample sizes.

ρ_{kk} Intervals	n	MSE$\times 10^{-3}$		ACC (%)					AUC (%)					Mislabeling Impact
		NB	INB	NB	INB	NB-T	NLNN	NAL	NB	INB	NB-T	NLNN	NAL	ΔACC (%)
[0.55, 0.65)	500	3.3	2.9	66.0	83.2	96.6	20.7	28.3	90.7	97.2	99.7	52.7	63.7	-22.3
	1000	2.2	1.2	75.9	92.6	96.9	21.1	39.3	94.7	99.4	99.8	55.4	74.8	-17.6
	5000	1.0	0.2	90.9	95.0	95.7	30.8	80.6	99.1	99.7	99.8	64.4	96.5	-4.2
[0.65, 0.75)	500	3.0	2.8	73.8	84.0	96.6	20.7	30.5	94.2	97.7	99.7	53.8	67.8	-14.5
	1000	1.7	1.2	85.3	92.7	96.9	21.1	46.8	97.9	99.4	99.8	56.4	80.6	-8.1
	5000	0.7	0.2	93.3	95.1	95.7	35.3	86.1	99.5	99.7	99.8	66.7	98.0	-1.8
[0.75, 0.85)	500	2.7	2.7	78.7	85.6	96.6	20.8	32.9	96.0	98.1	99.7	55.0	70.0	-9.6
	1000	1.4	1.2	89.2	93.0	96.9	22.0	54.6	98.8	99.4	99.8	58.6	85.2	-4.2
	5000	0.4	0.2	94.2	95.1	95.7	34.9	88.6	99.6	99.7	99.8	65.6	98.6	-0.9
[0.85, 0.95)	500	2.4	2.5	84.3	86.8	96.6	21.7	36.8	97.9	98.4	99.7	56.7	73.3	-4.0
	1000	1.2	1.2	91.8	93.2	96.9	22.3	60.0	99.3	99.5	99.8	58.7	88.1	-1.6
	5000	0.3	0.2	94.7	95.1	95.7	35.6	90.1	99.7	99.7	99.8	65.2	99.0	-0.4
[1.0, 1.0]	500	2.3	2.4	88.2	88.0	96.6	22.1	40.0	98.8	98.7	99.7	56.9	76.2	0.0
	1000	1.1	1.1	93.4	93.3	96.9	24.0	64.5	99.5	99.5	99.8	61.7	90.4	0.0
	5000	0.2	0.2	95.1	95.1	95.7	31.2	92.1	99.7	99.7	99.8	62.4	99.3	0.0

Real Data Experiments

- 20 Newsgroups Benchmark Dataset:
  - 18,864 documents with 15,076 in the training set and 3,770 in the testing set.
  - Top 7,302 words with the highest TF-IDF values are maintained.
  - Mislabeled instances are artificially generated. (20%)

- Models:
  1. NB (wrong)
  2. INB method
  3. NLNN method
  4. NAL method
  5. NB (correct)

因思想而光华

Real Data Experiments

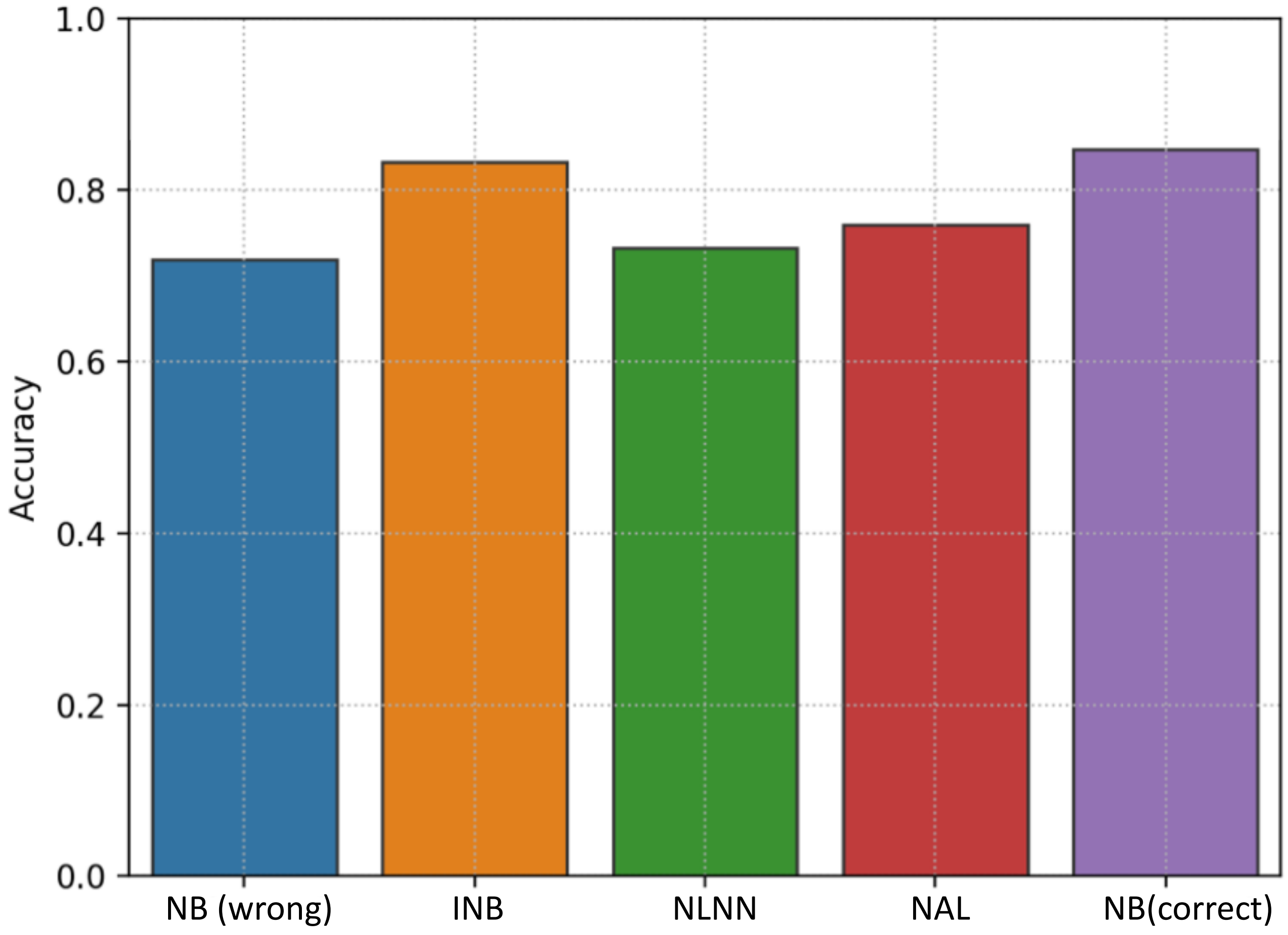


Figure 2. Classification accuracy results on the 20 Newsgroups Dataset.

Real Data Experiments

- Live Streaming Dialog Dataset:
  - $N = 1416$
  - Y: $K = 13$

Table 3. Thirteen Categories that the messages are classified into and their corresponding responses.

Category Number	Category Description	Response Strategy
1	Questions related to loans	Ask for the consumer's contact information.
2	Questions related to discounts	Directly reply "The discount is XX%".
3	Questions related to car prices	Directly reply "The car price is XX RMB".
4	Questions related to total cost	Directly reply "The total cost is XX RMB".
5	Questions related to availability	Directly reply "The car is available/unavailable".
6	Questions related to license plate	Answer "Yes" for the same province/"No" otherwise.
7	Questions related to store address	Directly reply the store address.
8	Questions totally irrelevant	Ignore the message and do not reply.
9	Leaving contact information	Directly reply "Message received".
10	Asking for contact information	A salesman/saleswoman will be automatically assigned.
11	Greeting message without car information	Ask for the consumer's car preference.
12	Messages without configuration information	Ask for the consumer's configuration preference.
13	Unclear message about new or second-hand cars	Directly ask "Do you mean a new or second-hand car".

Real Data Experiments

- Live Streaming Dialog Dataset:
  - $N = 1416$
  - Y: $K = 13$
  - X: $d = 22$

Table 4. Descriptions for the independent variables.

Variable Name	The Practical Meaning
X_1	Whether the message contains car information only?
X_2	Is the message a question?
X_3	Is the message the first message sent by the consumer?
X_4	Whether this message is about one specific car?
X_5	Whether detailed car configuration information is provided in the message?
X_6	Is configuration information included in the message?
X_7	Is this a message about the car store address?
X_8	Whether the consumer's contact information is given in the message?
X_9	Whether this message is about a new car?
X_{10}	Whether the message is about a second-hand car?
X_{11}	Does the consumer ask for contact information in this message?
X_{12}	Is the message a statement about one specific car?
X_{13}	Has the consumer left his contact information in the previous messages?
X_{14}	Is the message a question on license plates?
X_{15}	Is the message a question on total cost?
X_{16}	Is the message a question on car prices?
X_{17}	Is the message a question on discounts?
X_{18}	Is the message a question on whether the car is available or needs reservations?
X_{19}	Is the message a question on loans?
X_{20}	Is the message not about the loan, total cost, car price, discount, asking for contact information, leaving contact information, availability, car store address, or license plate?
X_{21}	Is the message not about the loan, total cost, car price, discount, asking for contact information, leaving contact information, availability, car store address, or license plate? Is the message the first message sent by the consumer?
X_{22}	Is the message not about the loan, total cost, car price, discount, asking for contact information, leaving contact information, availability, car store address, or license plate? Is the message the first message sent by the consumer? Can we tell which car the consumer refers to in this message?

Real Data Experiments

- Live Streaming Dialog Dataset:
  - $N = 1416$
  - Y: $K = 13$
  - X: $d = 22$
  - Mislabeling rate: about 19.49%
  - Train/Test split: 80%/20%
  - $B = 100$
- Models:
  - NB(wrong)
  - INB
  - NLNN
  - NAL
  - NB(correct)

因思想 而光华

Real Data Experiments

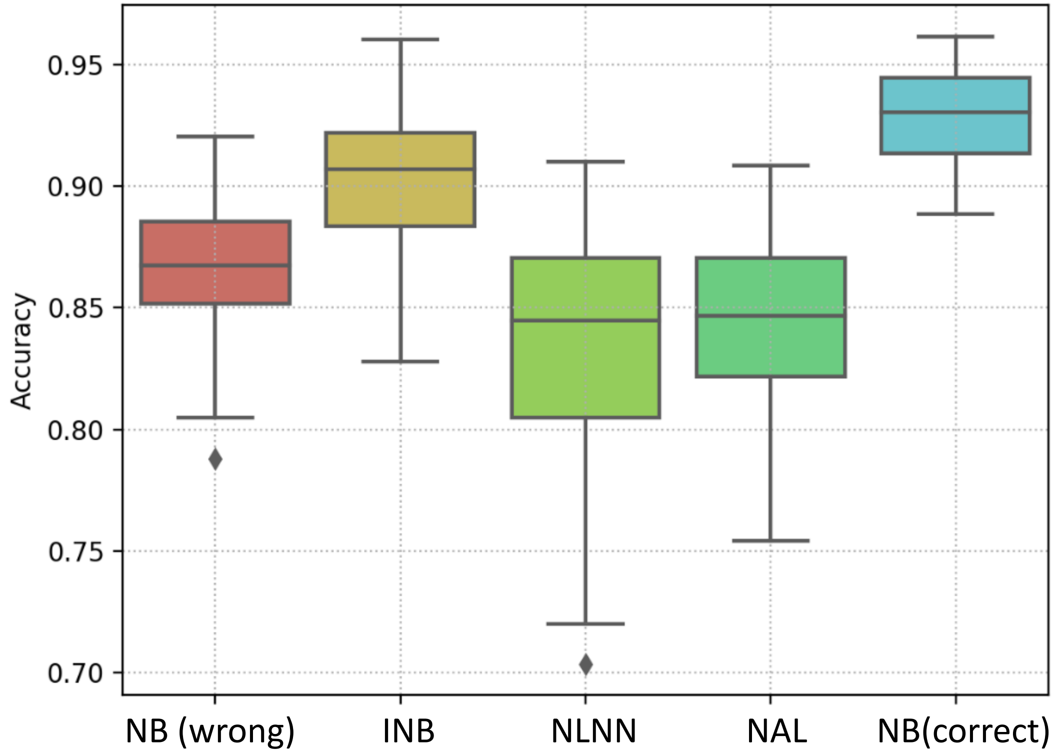


Figure 3. Classification accuracy results on the Live Streaming Dialog Dataset.

Future Work

- How to accommodate continuous features?

$$p(z_{ij}|Y_i^* = k) = \phi_{jk}(z_{ij})$$

$$\begin{aligned}
\ell_c(\theta) &= \ln P(\mathbb{X}, \mathbb{Z}, \mathbb{Y}, \mathbb{Y}^*|\theta) \\
&= \sum_{i=1}^{N} \ln P(Y_i^*|\theta) + \sum_{i=1}^{N} \ln P(Y_i|Y_i^*, \theta) \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{d_1} \ln P(X_{ij}|Y_i^*, \theta) + \sum_{i=1}^{N}\sum_{j=1}^{d_2} \ln P(Z_{ij}|Y_i^*, \theta) \\
&= \sum_{i=1}^{N} \ln \pi_{Y_i^*} + \sum_{i=1}^{N} \ln \rho_{Y_i Y_i^*} + \sum_{i=1}^{N}\sum_{j=1}^{d_1} X_{ij} \ln p_{jY_i^*} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{d_1} (1 - X_{ij}) \ln (1 - p_{jY_i^*}) + \sum_{i=1}^{N}\sum_{j=1}^{d_2} \ln \phi_{jY_i^*}(Z_{ij}).
\end{aligned}$$

Thanks!