Postprint of Teaching Evaluation Data Modeling Based on Discrete Poisson Mixture Model

Authors: Huang Hao, Yan Qian, Gan Ting, Shijun Li

Date: 2022-05-10T11:22:58Z

Abstract

Analyzing student evaluation data for teachers in teaching evaluation systems enables teachers to comprehend students' authentic attitudes toward instructors, summarize teaching experience, improve subsequent teaching methods, and enhance teaching quality. However, during teaching evaluations, issues such as random or malicious evaluations may arise among students, resulting in evaluation data containing substantial noise and leading to suboptimal feedback data. Therefore, a discrete Poisson mixture model is proposed to model the noisy student evaluation data, wherein each discrete Poisson component in the mixture model corresponds to a class of students with similar evaluation patterns, and the model parameters in the discrete Poisson distribution are used to represent the evaluation scores within the corresponding patterns. By constructing a log-likelihood function to measure the goodness-of-fit between the mixture model and the evaluation data, and employing a gradient descent method to solve for the model parameters with optimal fit, students' genuine evaluations of teachers can be identified, thereby ensuring effective communication between teachers and students in the teaching evaluation system. Extensive experimental results demonstrate that the model can rapidly and accurately identify students with different evaluation patterns from noisy evaluation data, and ascertain the authentic evaluation status of students toward teachers.

Full Text

Teaching Evaluation Data Modeling Based on Discrete Poisson Mixture Model

Huang Hao, Yan Qian, Gan Ting†, Li Shijun
(School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract

Analyzing student evaluation data of teachers in teaching evaluation systems helps educators understand students' genuine attitudes toward instruction, summarize teaching experiences, improve pedagogical methods, and enhance overall teaching quality. However, teaching evaluations often suffer from random or malicious ratings, introducing substantial noise that compromises feedback quality. This paper proposes a discrete Poisson mixture model to model noisy student evaluation data, where each discrete Poisson component corresponds to a group of students with similar evaluation patterns, and the model parameters represent the rating scores within each pattern. By constructing a log-likelihood function to measure the fit between the mixture model and evaluation data, we employ gradient descent to solve for the optimal model parameters, thereby identifying students' true evaluations of teachers and ensuring effective teacher-student communication. Extensive experimental results demonstrate that the model can quickly and accurately identify students with different evaluation patterns from noisy data and capture their genuine feedback.

Keywords: teaching evaluation system; crowdsourcing ideas; Poisson mixture model; parameter estimation method

0 Introduction

With the proliferation of online platforms, organizing student evaluations of teaching through web-based systems has become a widespread practice in universities, replacing manual statistical methods. As a critical component of teaching quality assessment, student evaluation of teaching has attracted increasing attention regarding how to manage instruction more effectively and scientifically. Teaching evaluation primarily involves students scoring teachers' instructional performance based on specific criteria, which helps educators summarize experiences, refine teaching methods, and ultimately promote teacher-student communication and improve teaching quality.

However, most current university teaching evaluation systems suffer from low participation and malicious evaluation. Students often lack motivation, believing their feedback will not change teaching practices. Even when evaluations are mandatory, students frequently complete them perfunctorily by assigning random scores to all instructors or deliberately giving low ratings. This undermines teaching enthusiasm, stagnates instructional effectiveness, and creates a vicious cycle where students perceive evaluations as meaningless.

Since collected evaluation data typically contains noise from random or malicious ratings that do not reflect genuine opinions, simple majority voting cannot reliably extract true student evaluations. Analyzing these noisy data requires more sophisticated approaches. Viewing students as crowdsourcing workers, this evaluation behavior resembles crowdsourcing services that aggregate extensive feedback. However, crowdsourcing data processing requires complex parametric models to model workers' labeling abilities, often employing EM-like algorithms that risk local optima and cannot efficiently obtain true task labels.

To address these limitations, this paper proposes modeling noisy student evaluation data using a discrete Poisson mixture model. Students with similar evaluation behaviors correspond to a single discrete Poisson component, with model parameters representing specific rating scores. We construct a maximum likelihood function to assess model-data fit and use gradient descent to find optimal parameters, thereby identifying students with different evaluation patterns and determining their true assessments. Extensive experiments show our model can quickly and accurately identify different student rating categories from noisy data while capturing accurate feedback on teaching performance.

1 Related Work

Teaching evaluation primarily collects student feedback on courses to analyze their true attitudes, involving data collection and processing.

Current evaluation data collection relies mainly on web surveys, but survey fatigue and students' satisfaction with course grades often result in low response rates or poor data quality. Some research focuses on obtaining high-quality evaluation data by using AI-powered conversational agents for personalized student interviews to generate richer feedback. Other researchers supplement evaluation data with additional platforms containing instructor information for more comprehensive assessment.

Open-ended textual feedback is processed through topic detection or sentiment classification algorithms. Many researchers employ LDA topic models to extract themes from written feedback, which identifies more relevant topics than clustering models. Classification techniques then categorize comments as positive or negative to better capture student sentiment. Others directly apply natural language processing techniques to analyze student comments, capturing meaningful emotional information. To mitigate the impact of anomalous information on analysis accuracy, some methods incorporate neural network-based anomaly detection algorithms.

Processing objective rating data resembles crowdsourcing tasks and represents our primary focus. We review relevant crowdsourcing research below.

Crowdsourcing is not a new phenomenon, but recent success of internet business models has renewed interest. Crowdsourcing leverages distributed human computation via the internet to solve specific problem sets, where humans participate actively or passively in computational processes, particularly for tasks inherently easier for humans than computers. Two primary development models exist: integrative crowdsourcing, where individual contributions are negligible but aggregated results yield significant value, and selective crowdsourcing, where only one optimal solution among many is adopted.

This has spurred research on modeling data from multiple unreliable sources. For crowdsourcing label tasks, the primary goal is label prediction—obtaining reliable instance labels. Mainstream methods assume a ground truth label exists for each instance and predict it based on worker annotations.

Karger et al. proposed an algorithm iteratively passing messages between instances and workers. Liu et al. extended this by incorporating worker priors and using graphical model variants to infer generative models. Whitehill et al. proposed modeling varying worker abilities and instance difficulties alongside ground truth labels through probabilistic models. Additional techniques like noise correction and imbalanced learning improve label quality, especially with noisy labels.

When instances can be represented in vector space, a related topic is learning classifiers from labels. This can be done by first inferring true labels using prediction techniques, then applying traditional classification. More sophisticated methods learn directly from worker labels while inferring hidden worker abilities, treat workers as individual classifiers related to the final classifier, or model worker abilities as functions of instance space to infer parameters jointly with the final classifier. These approaches model worker abilities differently but do not explicitly address instance difficulty when treating instance space holistically. A drawback is that vector representations are not always readily available for real-world tasks.

While most existing work aims to predict reliable per-instance labels, some address the problem differently. Wang and Zhou proposed a theoretical framework to identify high-quality workers. Welinder et al. introduced the concept of "schools of thought" to extract different perspective groups. Tian and Zhu extended this by clustering worker labels to estimate abilities and instance difficulty. Ertekin et al. studied approximating crowd opinions by querying only a subset of workers.

Despite numerous techniques for handling crowdsourcing imprecision, randomness, and uncertainty, most existing work covers only specific aspects. In contrast, our model predicts annotator data generation comprehensively, capturing behavior patterns and diverse opinions while remaining simple and flexible.

---

2 Related Concepts

2.1 Mixture Models

Mixture models address the limitations of single-distribution models by combining several single distributions to create more complex models capable of generating more sophisticated samples. Suppose random variable x follows a mixture distribution G composed of M populations $G_1, \dots, G_m$ with proportions $\lambda_1, \dots, \lambda_m$. The density function can be expressed as:

$$f(x|\Lambda, \Theta) = \sum_{m=1}^{M} \lambda_m f_m(x|\theta_m)$$

where $f_m(x|\theta_m)$ is the probability density function of the m-th component distribution, representing the probability of generating x given the m-th model; λ_m is the weight of the m-th component, interpretable as its prior probability. Adjusting weights λ_m substantially affects the mixture model' s probability density curve, enabling it to fit more complex and variable samples.

Mixture models are flexible and powerful probabilistic tools with extensive theoretical and practical applications, offering two key advantages: (1) They provide an effective framework for approximating complex distributions with simple structures. For instance, the normal distribution is the most common and important distribution in practice, and many random phenomena can be approximated by normal distributions with sufficiently large sample sizes. Theory proves that mixture normal distributions (also called mixture Gaussian distributions) can approximate any smooth distribution—given enough components M and properly set weights, mixture models can describe complex phenomena, helping solve many real-world problems. (2) The simulation provided by mixture models is natural. When $M = 1$, the model reduces to a single distribution, indicating homogeneous data properties; when $M > 1$, data come from different distributions with varying properties, making mixture models widely applicable in clustering and discriminant analysis.

2.2 Latent Variables

Mixture models contain unknown data called latent variables. The observed random variable x is generated by first selecting the m-th distribution G_m, then sampling from its probability distribution $f_m(x|\theta_m)$. Among N observations $x_1, \dots, x_N$, multiple may originate from the same component. Here, the observed data x_n are known, but the component membership γ_{nm} indicating which population distribution generated x_n is unknown:

$$\gamma_{nm} = \begin{cases} 1 & \text{if the } n\text{-th observation comes from the } m\text{-th component} \\ 0 & \text{otherwise} \end{cases}$$

for $n = 1, \dots, N$ and $m = 1, \dots, M$.

3 Model Framework

Student evaluation is the most direct, authentic, and reliable source in teaching evaluation systems, as students are the direct beneficiaries of teaching effectiveness. Students evaluate teaching performance, with students as raters and teaching work as the rated object, jointly promoting teaching implementation and improvement through the evaluation system. We first describe the problem, then introduce the discrete Poisson mixture model for probabilistic fitting, and finally present the parameter estimation method.

3.1 Problem Description

Assume S students rate N teachers' teaching performance, with each student' s rating drawn from $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, where higher scores indicate greater satisfaction. All student ratings are represented by the set $Y = \{Y_s\}_{s=1}^{S}$, where $Y_s = \{y_{s1}, \dots, y_{sN}\}$ denotes student s' s ratings for N teachers. The teaching evaluation problem is to identify different categories of rating patterns among students and determine the true evaluation of each teacher' s teaching work given rating set Y.

To create a simple yet flexible and practical model, we directly model the generation process of rating set Y. Experienced teachers exhibit relatively stable knowledge delivery and classroom performance, making the evaluation target (teaching work) have a relatively fixed pattern. While different student categories may respond differently to this pattern, students within the same category likely produce similar evaluations. This allows using generative machine learning models to discover how this teaching pattern affects different student categories. We treat each student as a unit, simulate the probability distribution of their ratings across all tasks, and assume each student' s rating results follow this distribution. By selecting an appropriate model form and deriving model parameters θ from observed labels in Y, we can capture these patterns.

3.2 Discrete Poisson Mixture Model

We select an N-dimensional discrete Poisson mixture model to fit the probability distribution of variable y for two reasons: (1) Each rating label value is influenced by a discrete Poisson distribution. (2) Variable y is N-dimensional as it reflects each student' s ratings across N tasks.

The discrete Poisson mixture model can be described as:

$$p(y|\mu, \alpha) = \sum_{k=1}^{K} \alpha_k g_k(y|\mu_k)$$

where $g_k(y|\mu_k)$ is the normalization factor for the discrete Poisson distribution, defined as:

$$g_k(y|\mu_k) = \frac{\mu_k^y}{\sum_{v=0}^{9} \frac{\mu_k^v}{v!}}$$

α_k is the coefficient (or weight) of the k-th Poisson component, and μ_k is the Poisson parameter of the k-th component.

In this model, each student s’s rating result is considered generated by a Poisson component. If a group of students behaves similarly (e.g., they all rate casually, or maliciously disrupt results), their ratings likely originate from the same Poisson component. Moreover, if this student group constitutes a large proportion, its corresponding Poisson component coefficient α_k will exceed others. We do not explicitly model how influencing factors jointly cause uncertainty and errors in rating set Y; instead, we directly simulate Y’ s generation process, incorporating the effects of influencing factors into Poisson parameter μ_k.

3.3 Model Parameter Estimation

Based on student rating set $Y = \{Y_s\}_{s=1}^{S}$, we construct the likelihood function for discrete Poisson mixture model parameters μ and α:

$$p(Y|\mu,\alpha) = \prod_{s=1}^{S} \left(\sum_{k=1}^{K} \alpha_k \prod_{i=1}^{N} g_k(y_{si}|\mu_k) \right)$$

The log-likelihood function is:

$$\ln p(Y|\mu,\alpha) = \sum_{s=1}^{S} \ln \left(\sum_{k=1}^{K} \alpha_k \prod_{i=1}^{N} g_k(y_{si}|\mu_k) \right)$$

We maximize this log-likelihood to obtain optimal parameters μ^* and α^*:

$$\mu^*, \alpha^* = \arg\max_{\mu,\alpha} \ln p(Y|\mu,\alpha)$$

Taking partial derivatives of the log-likelihood with respect to each component of μ and α:

$$\frac{\partial \ln p(Y|\mu,\alpha)}{\partial \mu_k} = \sum_{s=1}^{S} \frac{\alpha_k \prod_{i=1}^{N} g_k(y_{si}|\mu_k)}{\sum_{m=1}^{K} \alpha_m \prod_{i=1}^{N} g_m(y_{si}|\mu_m)} \cdot \sum_{i=1}^{N} \frac{g_k'(y_{si}|\mu_k)}{g_k(y_{si}|\mu_k)}$$

where $g_k'(y|\mu_k)$ is the derivative of $g_k(y|\mu_k)$ with respect to its argument:

$$g_k'(y|\mu_k) = \frac{\mu_k^{y-1}}{(y-1)!} \cdot \frac{1}{\sum_{v=0}^{9} \frac{\mu_k^v}{v!}} - \frac{\mu_k^y}{y!} \cdot \frac{\sum_{v=0}^{9} \frac{v \cdot \mu_k^{v-1}}{v!}}{\left(\sum_{v=0}^{9} \frac{\mu_k^v}{v!} \right)^2}$$

Using gradient descent, we initialize μ and α as $\mu^{(0)}$ and $\alpha^{(0)}$, then update them iteratively until convergence:

$$\mu^{(t+1)} = \mu^{(t)} + \theta^{(t)} \frac{\partial \ln p(Y|\mu^{(t)}, \alpha^{(t)})}{\partial \mu^{(t)}}$$

$$\alpha^{(t+1)} = \alpha^{(t)} + \theta^{(t)} \frac{\partial \ln p(Y|\mu^{(t)}, \alpha^{(t)})}{\partial \alpha^{(t)}}$$

where $\theta^{(t)}$ is the step size at iteration t. To ensure termination, we require:

$$\sum_{t=1}^{\infty} \theta^{(t)} = \infty \quad \text{and} \quad \lim_{t \to \infty} \theta^{(t)} = 0$$

3.4 Poisson Component Analysis

Section 3.3 estimates parameters for the discrete Poisson mixture model proposed in Section 3.2 using observed student ratings. For any student s with ratings y_s, the association between their ratings and the k-th component of the discrete Poisson mixture model can be estimated via the likelihood function:

$$p(y_s|\mu, \alpha) = \sum_{k=1}^{K} \alpha_k \prod_{i=1}^{N} g_k(y_{si}|\mu_k)$$

3.5 Teacher Teaching Quality Analysis

After obtaining parameter estimates for the teaching rating discrete Poisson mixture model, we can analyze teaching quality for each task. Consider the score y_i for the i-th teaching task. We estimate the scores of the remaining $N-1$ tasks without variable y_i, denoted as y_{-i}. The estimated score for the i-th teaching task is:

$$p(y_i|y_{-i}, \mu, \alpha) = \frac{\sum_{k=1}^{K} \alpha_k g_k(y_i|\mu_k) \prod_{j \neq i} g_k(y_j|\mu_k)}{\sum_{l=0}^{9} \sum_{k=1}^{K} \alpha_k g_k(l|\mu_k) \prod_{j \neq i} g_k(y_j|\mu_k)}$$

We can use the score m with maximum probability as the task score or take the expectation:

$$y_i^* = \arg \max_{m \in \{1,\dots,10\}} p(y_i = m|y_{-i}, \mu, \alpha)$$

4 Experimental Results and Analysis

This section introduces the dataset and evaluation metrics, then validates our model' s effectiveness in (1) classifying students by behavior patterns and (2) accurately predicting true labels.

4.1 Experimental Setup

The teaching evaluation survey administered to each student includes course time, class number, course name, course ID, instructor, and rating score (Table 1). Students rate courses based on their experience using a 1-10 scale. Our experimental rating data was simulated from 100 students rating 50 teachers. We first assigned each teacher a true rating label, then divided students into three categories: normal raters who scored around the given teacher labels with minor variations, random raters who scored uniformly from 0-10, and malicious raters who scored maliciously from 0-5.

Table 1. Teaching Assessment Questionnaire

Content	Description
Semester	2020-2021 Academic Year, First Semester
Course Info	Time, Class Number, Name, ID
Instructor	Teacher Name
Rating Score	1-10 scale based on course experience

Evaluation Metrics:

1. **Student Classification**: Calculate each student' s membership degree to the dominant Poisson component based on their category during data generation. For normal students, the mean membership is computed as:

$$\text{membership}_{\text{normal}} = \frac{1}{|S_{\text{normal}}|} \sum_{s \in S_{\text{normal}}} r_{s,k}$$

where $r_{s,k}$ denotes student s' s membership to the dominant Poisson component k, and $I(\cdot)$ is an indicator function returning 1 when the condition holds and 0 otherwise.

2. **Label Prediction**: Compute the difference between predicted teacher labels (derived from Poisson parameters) and given true labels. For N teachers with true labels y_i and predicted labels p_i, the difference score is:

$$\text{score} = \frac{1}{N} \sum_{i=1}^{N} |y_i - p_i|$$

4.2.1 Student Classification

To validate our model' s ability to classify students with different behavior patterns, we first generate random rating labels for each teacher, then create simulated rating data. We test three scenarios: (1) normal + random ratings, (2) normal + malicious ratings, and (3) normal + random + malicious ratings. Let α be the proportion of normal students (varying from 0.6 to 0.8). In scenarios with only random or malicious raters, their proportion is $1-\alpha$. When both are present, each constitutes $(1-\alpha)/2$. For reliability, we generate multiple experimental datasets, record average expectations per execution, and report means and variances across runs (shown as shaded regions in figures).

Figure 1 shows membership degrees in the dominant Poisson component for different student categories. Normal students exhibit significantly higher average membership than random or malicious raters, indicating accurate classification. The model successfully distinguishes genuine evaluators from noisy contributors.

4.2.2 Label Prediction

To verify that our model can extract true teacher evaluations from rating data, we compare predicted scores against true labels under the three scenarios described above. We also compare against majority voting and Raykar et al.'s label prediction algorithm. **Figure 2** demonstrates that our model achieves higher accuracy than both baselines. This is because random and malicious raters are assigned to non-dominant Poisson components, leaving genuine raters in the dominant component, which yields more accurate teacher score predictions.

5 Conclusion

This paper proposes a discrete Poisson mixture model to simulate student rating results for teaching performance, along with a gradient descent method for parameter estimation. The model directly simulates the rating generation process without requiring additional assumptions about student rating abilities or instance difficulty. Experimental results validate its effectiveness in label prediction and classifying students with different behavior patterns. Compared to previous teaching evaluation methods, our model demonstrates higher fault tolerance, producing reliable assessment results even with random and malicious raters, thereby providing accurate feedback that reflects actual teaching conditions.

References

[8] Mao Y, Zhu Y, Zhang S, et al. Detecting interest-factor influenced abnormal evaluation of teaching via multimodal embedding and priori knowledge based

neural network [C]// ISPA/BDCloud/SocialCom/SustainCom. Piscataway, NJ: IEEE Press, 2019: 1201-1209.

[9] Bhatti S S, Gao X, Chen G. General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey [J]. Journal of Systems and Software, 2020, 167: 110611.

[10] Tong Y, Zhou Z, Zeng Y, et al. Spatial crowdsourcing: a survey [J]. The VLDB Journal, 2020, 29 (1): 217-250.

[11] Huo Xuyan. Research on the status quo of crowdsourcing development in China [J]. ShangQing, 2011 (47): 117-117.

[1] Wambsganss T, Winkler R, Söllner M, et al. A conversational agent to improve response quality in course evaluations [C]// Proc of ACM CHI. New York: ACM Press, 2020: 1-9.

[12] Raykar V C, Yu S, Zhao L H, et al. Learning From Crowds [J]. Journal of Machine Learning Research, 2010, 11 (2): 1297-1322.

[2] Kavalchuk A, Goldenberg A, Hussain I. An empirical study of teaching qualities of popular computer science and software engineering instructors using ratemyprofessor.com data [C]// Proc of the 42th International Conference on Software Engineering: Software Engineering Education and Training. Piscataway, NJ: IEEE Press, 2020: 12-20.

[3] Lin Q, Zhu Y, Lu H, et al. Improving university faculty evaluations via multi-view knowledge graph [J]. Future Generation Computer Systems, 2021, 117: 181-192.

[4] Gottipati S, Shankararaman V, Lin J. Latent Dirichlet Allocation for textual student feedback analysis [C]// Proc of the 26th International Conference on Computers in Education. 2018: 220-227.

[5] Unankard S, Nadee W. Topic detection for online course feedback using LDA [C]// Proc of the 4th International Symposium on Emerging Technologies for Education. Berlin: Springer, 2019: 133-142.

[6] Andersson E, Dryden C, Variawa C. Methods of applying machine learning to student feedback through clustering and sentiment analysis [C]/ Proc of CEEA. 2018. https://doi.org/10.24908/pceea.v0i0.13059

[7] Onan A. Mining opinions from instructor evaluation reviews: a deep learning approach [J]. Computer Applications in Engineering Education, 2020, 28 (1): 117-138.

[13] Karger D R, Oh S, Shah D. Iterative learning for reliable crowdsourcing systems [C]// Proc of the 24th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2011: 1953-1961.

[14] Liu Q, Peng J, Ihler A. Variational inference for crowdsourcing [C]// Proc of the 25th International Conference on Neural Information Processing Systems.

Cambridge, MA: MIT Press, 2012: 692-700.

[15] Whitehill J, Ruvolo P, Wu T, et al. Whose vote should count more: optimal integration of labels from labelers of unknown expertise [C]// Proc of the 22th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 2035-2043.

[16] Zhang J, Sheng V S, Wu J, et al. Improving label quality in crowdsourcing using noise correction [C]// Proc of the 24th ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1931-1934.

[17] Zhang J, Wu X, Sheng V S. Imbalanced Multiple Noisy Labeling [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27 (2): 489-503.

[18] Yan Y, Rosales R, Fung G, et al. Modeling Multiple Annotator Expertise in the Semi-Supervised Learning Scenario [J]. Computer Science, 2012: 1-9.

[19] Kajino H, Tsuboi Y, Kashima H. Clustering crowds [C]// Proc of the 27th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2013, 27 (1): 1120-1127.

[20] Zhao Y, Zhu Q. Evaluation on crowdsourcing research: Current status and future direction [J]. Information Systems Frontiers, 2014, 16 (3): 417-434.

[21] Wang W, Zhou Z H. Crowdsourcing label quality: a theoretical analysis [J]. Science China Information Sciences, 2015, 58 (11): 1-12.

[22] Welinder P, Branson S, Belongie S, et al. The multidimensional wisdom of crowds [C]// Proc of the 23th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2010: 2424-2432.

[23] Tian Y, Zhu J. Learning from crowds in the presence of schools of thought [C]// Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 226-234.

[24] Ertekin S, Rudin C, Hirsh H. Approximating the crowd [J]. Data Mining and Knowledge Discovery, 2014, 28 (5-6): 1189-1221.

[25] Sheng V S, Zhang J. Machine learning with crowdsourcing: A brief summary of the past research and future directions [C]// Proc of the 33th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019, 33 (01): 9837-9843.

Note: Figure translations are in progress. See original paper for figures.