Journal Pre-proof

Generalisations of stochastic supervision models

Xiaoou Lu, Yangqi Qiao, Rui Zhu, Guijin Wang, Zhanyu Ma,
Jing-Hao Xue

PII: S0031-3203(20)30378-2
DOI: https://doi.org/10.1016/j.patcog.2020.107575
Reference: PR 107575

To appear in: *Pattern Recognition*

Received date: 9 May 2019
Revised date: 15 June 2020
Accepted date: 4 August 2020

Please cite this article as: Xiaoou Lu, Yangqi Qiao, Rui Zhu, Guijin Wang, Zhanyu Ma, Jing-Hao Xue, Generalisations of stochastic supervision models, *Pattern Recognition* (2020), doi: https://doi.org/10.1016/j.patcog.2020.107575

Highlights

- We study the stochastic supervision problem where only probabilistic assessments are provided for classification.
- We propose four novel generalisations of stochastic supervision models.
- We also develop four new EM algorithms for the generalisations.

Generalisations of stochastic supervision models

Xiaoou Lu[a], Yangqi Qiao[a], Rui Zhu[b,c,*], Guijin Wang[d], Zhanyu Ma[e], Jing-Hao Xue[a]

[a] *Department of Statistical Science, University College London, London WC1E 6BT, UK*
[b] *Faculty of Actuarial Science and Insurance, Cass Business School, City, University of London, London EC1Y 8TZ, UK*
[c] *School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7FS, UK*
[d] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[e] *The Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Abstract

When the labelling information is not deterministic, traditional supervised learning algorithms cannot be applied. In this case, stochastic supervision models provide a valuable alternative to classification. However, these models are restricted in several aspects, which critically limits their applicability. In this paper, we provide four generalisations of stochastic supervision models, extending them to asymmetric assessments, multiple classes, feature-dependent assessments and multi-modal classes, respectively. Corresponding to these generalisations, we derive four new EM algorithms. We show the effectiveness of our generalisations through illustrative examples of simulated datasets, as well as real-world examples of three famous datasets, the MNIST

*Corresponding author. Tel.: +44 (0)20 7040 4707
Email addresses: xiaoou.lu.13@ucl.ac.uk (Xiaoou Lu), yangqi.qiao.15@alumni.ucl.ac.uk (Yangqi Qiao), rui.zhu@city.ac.uk (Rui Zhu), wangguijin@mail.tsinghua.edu.cn (Guijin Wang), mazhanyu@bupt.edu.cn (Zhanyu Ma), jinghao.xue@ucl.ac.uk (Jing-Hao Xue)

dataset, the CIFAR-10 dataset and the EMNIST dataset.

Keywords: EM algorithms, imperfect supervision, finite mixture model, stochastic supervision

1. Introduction

Generally speaking, the aim of various statistical learning methods is to infer the real label y of an input instance x. Classification and clustering are two extreme ends in the sense of amount of labelling information provided for the inference of y. In classification, the deterministic labels $\{y_n\}_{n=1}^N$ of N training instances $\{x_n\}_{n=1}^N$, represented by a binary or multilevel categorical random variable y, are usually provided in advance to train a classifier $f(y|x)$ on the information from both the input and output spaces via $(\{x_n\}_{n=1}^N, \{y_n\}_{n=1}^N)$. The trained (supervised) classifier is then used to infer the real label y of a test instance x. In contrast, in clustering, no labelling information is provided at all, hence a clustering method $f(y|x)$ is built on the information from only the input space via $\{x_n\}_{n=1}^N$.

In between classification and clustering, there exists partially-supervised classification [1–5] with various types of information provided to help inference. One example is called semi-supervised classification [6, 7], where only part of the deterministic labels $\{y_n\}_{n=1}^N$ are provided for classifier training. Another example is called imperfect supervision [8–12], where there are some wrong deterministic labels provided in $\{y_n\}_{n=1}^N$. Multiple instance learning [13] also deals with partially-supervised setting, where deterministic labels are provided for bags of multiple instances rather than for each specific instance. In this paper, we discuss another partially-supervised classification scheme called stochastic supervision, which, in contrast to all the cases aforementioned, provides no deterministic labels $\{y_n\}_{n=1}^N$ but only probabilistic assessments $\{z_n\}_{n=1}^N$ for inference of y. In other words, only some side information about the output is provided.

A motivation of stochastic supervision is that, in practice, data are often labelled by certain experts or say supervisors with subjective labelling to some extent, and in many situations an expert cannot provide deterministic labels. For example, in medical diagnostic, an expert may not be perfectly sure whether a patient has a certain disease, and they can only provide a subjective assessment, which is often expressed in a probabilistic manner. These probabilistic assessments can be represented by continuous random variables, from a space different from the discrete space of output label y. On the basis of these assessments (or say probabilistic labels), the statistical classification problem, of fitting a model to the training data and inferring the real labels of the test data, was studied under the nomenclature of stochastic supervision [14–19].

The research of stochastic supervision models for discriminant analysis was pioneered by Aitchison and Begg [14] and Krishnan and Nandy [15]. As with [15] we assume two classes, namely class 1 and class 2, with proportions π_1 and $\pi_2 = 1 - \pi_1$, respectively. In each class, the data available, including both the d-dimensional feature vector x of an instance and its supervisor's assessment z that the instance belongs to class j, follow a class-dependent distribution $f_j(x, z)$, for $j = 1, 2$. The task is to infer the real label y of the instance (x, z).

In [15], the class-dependent joint data-generating distribution $f_j(x, z)$ was further factorised as $f_j(x, z) = f_j(x)q_j(z)$, by assuming that the features x and the assessment z are independent of each other in each class. By supposing the features x are continuous random variables in the range of $(-\infty, \infty)$, it was assumed that $x|y = 1 \sim N(\mu_1, \Sigma)$ and $x|y = 2 \sim N(\mu_2, \Sigma)$, two class-dependent d-variate Gaussian distributions. We denote the pdfs of $x|y = 1$ and $x|y = 2$ as $f_1(x)$ and $f_2(x)$, respectively. In the meantime, as the probabilistic assessment z is a continuous random variable in the range of $[0, 1]$, it was assumed that $z|y = 1 \sim \text{Beta}(a, b)$ and $z|y = 2 \sim \text{Beta}(b, a)$, two Beta distributions symmetric between the two classes. We denote the pdfs of $z|y = 1$ and $z|y = 2$ as $q_1(z)$ and $q_2(z)$, respectively. That is to say, the model in [15] assumes that the data-generating process in class j follows a Gaussian distribution $f_j(x)$ for features x and a Beta distribution $q_j(z)$ for assessment z. Although the assessment z is given for each training instance x, the real label (denoted by y) is unknown, which leads the likelihood of the training instance, or say the joint distribution of x and z, as $p(x, z) = \pi_1 f_1(x, z) + \pi_2 f_2(x, z)$. Hence this is a latent variable (finite mixture) problem, and the model was fitted by an EM algorithm in [15].

However, there are two technical issues with Krishnan and Nandy's stochastic supervision model. Firstly, it cannot accept any assessment that $z > 1$ or $z < 0$, while in some real problems the assessment can be a random variable in the range of $(-\infty, \infty)$. Secondly, the EM algorithm for this model is complicated, because there is no exact solution in the M-step for the estimation of certain parameters due to the adoption of the Beta distributions for assessment z.

In order to overcome the two issues above, Titterington [16] introduced a new supervisor's assessment $w = \log \frac{z}{1-z}$ to replace the original z. This transformation is called additive logistic transformation [20], which extends the range of the assessment from $[0, 1]$ to the real line and thus the assessment can be modelled by Gaussian distributions. In Titterington's model, supervisor assessments $q_1(w)$ and $q_2(w)$ are assumed to follow two univariate Gaussian distributions $N(-\Delta, \Omega)$ and $N(\Delta, \Omega)$, respectively, where $\Delta > 0$ and $\Omega > 0$. In this model, the constraints of equal variances and symmetry in the assessment distributions between the two classes are preserved. Then Titterington [16] provided an EM algorithm to estimate parameters $\{\pi_1, \mu_1, \mu_2, \Sigma, \Omega, \Delta\}$.

In this paper, we aim to generalise Titterington's model in four aspects, to make it more flexible and generic to deal with more complicated real-world classification tasks. We note that the first three aspects have been suggested and discussed by Titterington in section 5.2 of [16], though no detailed derivation was provided as we shall present in this paper. Our four generalisations are briefly described as follows.

1. *Asymmetric assessments.* In both Krishnan and Nandy's and Titterington's models, the two class-dependent distributions of assessments $q_j(z)$ (or $q_j(w)$) were symmetric and with equal variances. Our first generalisation aims to relax this restriction on the parameter setting of supervisor's assessments.
2. *Multiple classes.* The past models were for two-class discrimination. Our second generalisation is designed for classification of multiple classes.
3. *Feature-dependent assessments.* In Krishhan and Nandy's [15] and Tit-

terington's [16] work, the assessment and the features were modelled independent of each other. Our third generalisation aims to model their dependence.

4. *Multi-modal classes.* In the past research on stochastic supervision, each class was modelled by a Gaussian distribution, implying that there was only a single population for each class, which we call it a uni-modal class. In our fourth generalisation, we model the cases that each class contains multiple subclasses, making the class a multi-modal class.

We shall detail the four generalisations in four subsections of section 2 along with four EM algorithms and some numerical illustrations. In section 3, we present real-data examples to demonstrate the effectiveness of the generalisations.

2. Generalised models and their EM algorithms

2.1. Generalisation-1: asymmetric stochastic supervision

Let us first make the parameter setting of stochastic supervision models more flexible. In Titterington's model [16], the distributions of assessments in two classes are $w|y = 1 \sim N(-\Delta, \Omega)$ and $w|y = 2 \sim N(\Delta, \Omega)$. They are symmetric in the sense that their variances are the same and their means are the additive inverses of each other. Here as suggested by Titterington [16], we generalise them to $w|y = 1 \sim N(\Delta_1, \Omega_1)$ and $w|y = 2 \sim N(\Delta_2, \Omega_2)$. We denote the pdfs of $w|y = 1$ and $w|y = 2$ as $q_1(w)$ and $q_2(w)$, respectively.

2.1.1. Formulation of generalisation-1

Our notation is established as follows. The observable dataset is denoted by $\mathcal{X} = \{X, W\}$, the latent variable set by $\mathcal{Y} = \{Y\}$, and the parameter set by $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \Sigma, \Omega_1, \Delta_1, \Omega_2, \Delta_2\}$, where $X = \{x_n\}$, $W = \{w_n\}$ and $Y = \{y_n\}$, for $n = 1, \ldots, N$, are N instances, assessments and real labels of the instances, respectively. For each instance, $y_n = (y_{n1}, y_{n2})$ is a latent variable vector (representing its real label) such that for class j we have $y_{nj} \in \{0, 1\}$ and for two classes together we have $\sum_{j=1}^{2} y_{nj} = 1$. That is, y_n is a latent indicator vector with only one element being true.

Hence, for complete data $(\mathcal{Y}, \mathcal{X}) = \{(y_n, x_n, w_n), n = 1, \ldots, N\}$, the complete-data likelihood is

$$p(\mathcal{Y}, \mathcal{X}) = \prod_{n=1}^{N} \{[\pi_1 f_1(x_n) q_1(w_n)]^{y_{n1}} + [\pi_2 f_2(x_n) q_2(w_n)]^{y_{n2}}\} \ .$$

Since this model contains latent variables y_n, we can estimate the model parameters by deriving an EM algorithm. In general, an EM algorithm [21] is an iterative algorithm providing a maximum likelihood solution for incomplete data. We can also use the EM algorithm for models with latent variables. In each of its iterations, the EM algorithm has two alternating steps, the expectation (E-)step and the maximisation (M-)step.

In the E-step, we fix current parameters and compute expectation of the complete-data log-likelihood function with respect to the conditional distributions of latent variables given observed data $\mathcal{X}$: $Q(\theta, \theta^{old}) = \mathbb{E}_{\mathcal{Y}|\mathcal{X},\theta^{old}}(\log p(\mathcal{Y}, \mathcal{X}|\theta))$.

In the M-step, we find new parameters by maximising the expectation obtained in the E-step: $\theta^{new} = \arg\max_\theta Q(\theta, \theta^{old})$.

2.1.2. EM algorithm of generalisation-1

E-step. For the generalisation-1, in the E-step, we compute the posterior probabilities of latent variables $\gamma(y_{nj}) = p(y_{nj} = 1|\mathcal{X}, \theta)$. By the Bayes rule, we have

$$\gamma(y_{nj}) = \frac{p(x_n, w_n, y_{nj}|\theta)}{p(x_n, w_n|\theta)} = \frac{\pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)}{\sum_{j=1}^{2} \pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)} ,$$

which are called responsibilities that class j takes for explaining x_n [22].

M-step. In the M-step, we take partial differential of $l(\theta) = Q(\theta, \theta^{old})$ with respect to $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \Sigma, \Omega_1, \Delta_1, \Omega_2, \Delta_2\}$ and set it equal to zero to obtain updated parameters θ^{new}. It follows that

$$\mu_1^{new} = \frac{\sum\limits_{n=1}^{N} \gamma(y_{n1}) x_n}{\sum\limits_{n=1}^{N} \gamma(y_{n1})} , \ \mu_2^{new} = \frac{\sum\limits_{n=1}^{N} \gamma(y_{n2}) x_n}{\sum\limits_{n=1}^{N} \gamma(y_{n2})} ,$$

indicating that the updated mean μ_j^{new} of the features in class j becomes a weighted average of all data points from the two classes, weighted by the responsibilities; and similarly

$$\Delta_1^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{n1}) w_n}{\sum_{n=1}^{N} \gamma(y_{n1})} , \ \Delta_2^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{n2}) w_n}{\sum_{n=1}^{N} \gamma(y_{n2})} ,$$

i.e., the updated mean Δ_j^{new} of assessments in class j becomes a weighted average of all assessments over the two classes.

Also, the updated covariance matrix of the features is

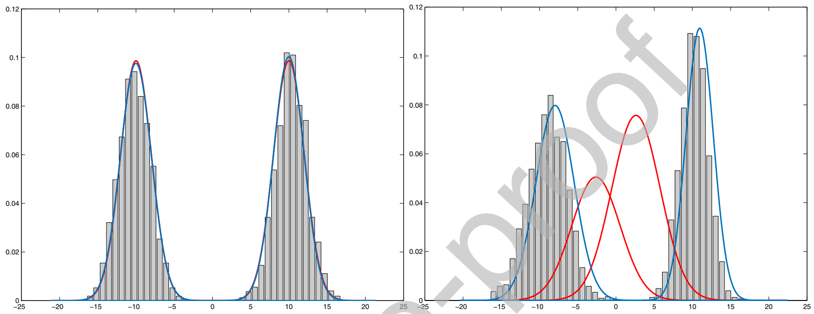
$$\Sigma^{new} = \frac{\sum\limits_{n=1}^{N} \sum\limits_{j=1}^{2} \gamma(y_{nj})(x_n - \mu_j)(x_n - \mu_j)^T}{\sum\limits_{n=1}^{N} \sum\limits_{j=1}^{2} \gamma(y_{nj})} ,$$

a weighted pooled covariance matrix; and similarly the updated variances of class-specific assessments are

$$\Omega_1^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{n1})(w_n - \Delta_1)^2}{\sum_{n=1}^{N} \gamma(y_{n1})} , \ \Omega_2^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{n2})(w_n - \Delta_2)^2}{\sum_{n=1}^{N} \gamma(y_{n2})} .$$

Since the two mixing weights have to satisfy $\pi_0 + \pi_1 = 1$, we can set $\partial l(\theta)/\partial \pi_j + \lambda = 0$, where λ is a Lagrange multiplier. It then follows that $\pi_1^{new} = \frac{1}{N}\sum_{n=1}^{N} \gamma(y_{n1})$, $\pi_2^{new} = 1 - \pi_1^{new}$, indicating that each of the updated mixing weights is an average of the responsibilities.

2.1.3. Illustrative example for generalisation-1

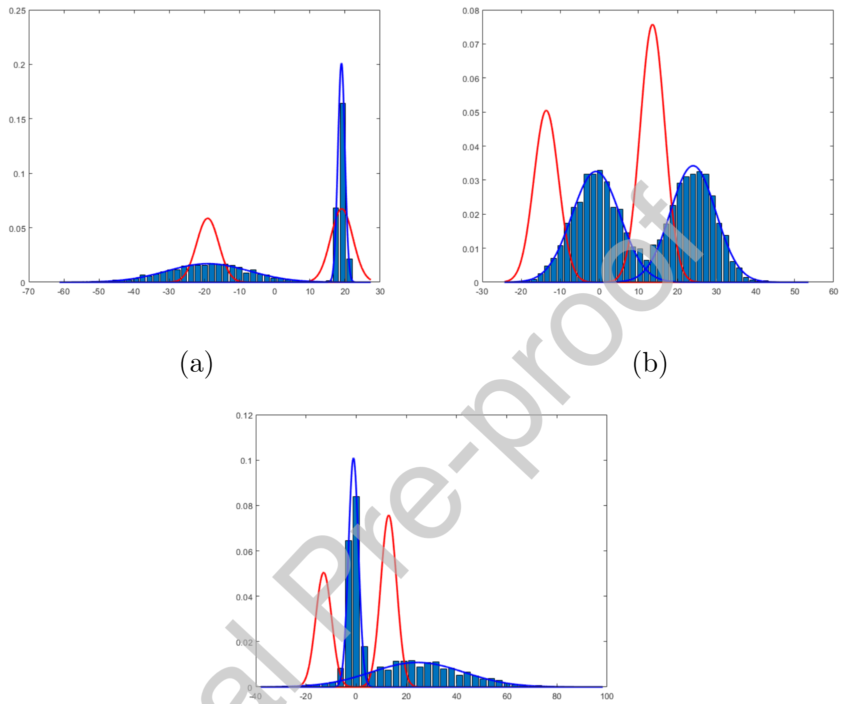


(a) (b)

Figure 1: (a) Supervisor assessments with *equal* variances and *symmetrical* means between the two classes. Red curve: assessments density estimated by Titterington's model. Blue curve: assessments density estimated by the generalisation-1. (b) Supervisor assessments with *unequal* variances and *asymmetrical* means between the two classes. The rest caption is as for Figure 1(a).

As shown in Figure 1(a) and Figure 1(b), compared with Titterington's original model, the generalisation-1 is more flexible in accommodating the distributions of supervisor's assessments of various shapes. Let us appreciate it from two aspects.

Firstly, we simulate the supervisor's assessments from two Gaussian distributions with *equal* variances and *symmetrical* means; this setting satisfies the assumption underlying Titterington's model. In this case, as shown in Figure 1(a), the generalisation-1 performs similarly to Titterington's model.



(a) (b)

(c)

Figure 2: Three extreme cases of supervisor assessments. (a) Supervisor assessments with large *unequal* variances and *symmetrical* means between the two classes. Red curve: assessments density estimated by Titterington's model. Blue curve: assessments density estimated by the generalisation-1. (b) Supervisor assessments with large *equal* variances and *asymmetrical* means between the two classes. The rest caption is as for Figure 2(a). (c) Supervisor assessments with large *unequal* variances and *asymmetrical* means between the two classes. The rest caption is as for Figure 2(a).

Secondly, we simulate the supervisor's assessments from two Gaussian distributions with *unequal* variances and *asymmetrical* means; this setting does not satisfy the assumption underlying Titterington's model. In this case, as shown in Figure 1(b), the generalisation-1 has much better fitting performance than Titterington's model.

Besides the moderate unequal variances and asymmetrical case shown in Figure 1(b), we also present the superior fitting performances of the generalisation-1 in three extreme cases in Figure 2: supervisor's assessments simulated from two Gaussian distributions with large *unequal* variances and *symmetrical* means in Figure 2(a), large *equal* variances and *asymmetrical* means in Figure 2(b) and large *unequal* variances and *asymmetrical* means in Figure 2(c). Obviously, the generalisation-1 can provide better fittings than Titterington's model under these extreme unequal variances and asymmetrical cases.

2.2. Generalisation-2: multi-class stochastic supervision

Original stochastic supervision models were only for two-class discrimination. In practice multi-class classification problems are also prevailing. Hence here we extend Titterington's model to multi-class cases, as suggested by Titterington [16].

2.2.1. Formulation of generalisation-2

Suppose there are J classes. As with [16], the supervisor's assessment of an instance x is now a J-variate vector of 'probabilities', $z = (z_1, \ldots, z_J)$, and we can define a new assessment vector $w_j = \log \frac{z_j}{z_J}$ for $j = 1, \ldots, J-1$, which extends the supervisor's assessments from $(0, 1)$ to $(-\infty, \infty)$. Then we can assume that, for each class j, the assessments $w = (w_1, \ldots, w_{J-1})$ follow $(J-1)$-variate Gaussian distributions: $q_j(w) \sim N(\Delta_j, \Omega_j)$, where $q_j(w)$ is the pdf of $w|y = j$.

Then, given the real label $y_n = (y_{n1}, \ldots, y_{nJ})$ is unknown, the joint distribution of the observed features x_n and assessment w_n of the nth instance becomes $p(x_n, w_n) = \sum_{j=1}^{J} \pi_j f_j(x_n, w_n)$, where $f_j(x_n, w_n) = f_j(x_n) q_j(w_n)$ and $\pi_j = p(y_{nj} = 1)$ is the mixing weight of class j.

Before going further, we recall some notation to be used for the generalisation-2:

- set of the latent labels $Y = \{y_n\}$, for $n = 1, \ldots, N$, where y_n is a J-variate latent vector of real labels, and we have $y_{nj} \in \{0, 1\}$ and $\sum_{j=1}^{J} y_{nj} = 1$;
- set of the class mixing weights $\Pi = \{\pi_j\}$, for $j = 1, \ldots, J$, where π_j is a scalar;
- set of the class means $U = \{\mu_j\}$, for $j = 1, \ldots, J$, where μ_j is a d-variate vector;
- set of the class covariances $\Sigma = \{\Sigma_j\}$, for $j = 1, \ldots, J$, where Σ_j is a $d \times d$ matrix;
- set of the assessment means $\Delta = \{\Delta_j\}$, for $j = 1, \ldots, J$, where Δ_j is a $(J-1)$-variate vector; and
- set of the assessment covariances $\Omega = \{\Omega_j\}$, for $j = 1, \ldots, J$, where Ω_j is a $(J-1) \times (J-1)$ matrix.

In this notation, the parameter set for the generalisation-2 is $\theta = \{\Pi, U, \Sigma, \Delta, \Omega\}$; the complete-data likelihood of observed data $\mathcal{X}$ and latent data $\mathcal{Y}$ is $p(\mathcal{Y}, \mathcal{X}|\theta) = \prod_{n=1}^{N} \sum_{j=1}^{J} [\pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)]^{y_{nj}}$, and the marginal likelihood of observed data $\mathcal{X}$ is $p(\mathcal{X}|\theta) = \prod_{n=1}^{N} \sum_{j=1}^{J} \pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)$.

2.2.2. EM algorithm of generalisation-2

E-step. In the E-step we can update posterior distribution of latent variables by setting $q^{new}(\mathcal{Y}) = p(\mathcal{Y}|\mathcal{X}, \theta^{old})$. Since

$$p(\mathcal{Y}|\mathcal{X}, \theta^{old}) = \prod_{n=1}^{N} \frac{\sum_{j=1}^{J} y_{nj} [\pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)]}{\sum_{j=1}^{J} \pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)} ,$$

we have the class responsibilities as

$$\gamma(y_{nj}) = \frac{\pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)}{\sum_{j=1}^{J} \pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)} .$$

M-step. In the M-step, we update θ by $\theta^{new} = \arg\max_\theta \sum_{\mathcal{Y}} q^{new}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}|\theta)$. Since the mixing weights π_j satisfy the sum-to-one constraint, as in section 2.1 we introduce a Lagrange multiplier λ and set $\partial l(\theta)/\partial \pi_j + \lambda(\sum_{j=1}^{J} \pi_j - 1) = 0$, which results in the updated mixing weights as $\pi_j^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma(y_{nj})$, which is again an average of the responsibilities over all the data points. Similarly to the M-step in section 2.1, we can obtain the updated means and covariance matrices as

$$\mu_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj}) x_n}{\sum_{n=1}^{N} \gamma(y_{nj})} , \; \Sigma_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj})(x_n - \mu_{jk})(x_n - \mu_{jk})^T}{\sum_{n=1}^{N} \gamma(y_{nj})} ,$$

$$\Delta_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj}) w_n}{\sum_{n=1}^{N} \gamma(y_{nj})}, \ \Omega_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj})(w_n - \Delta_j)(w_n - \Delta_j)^T}{\sum_{n=1}^{N} \gamma(y_{nj})}.$$

2.2.3. Illustrative example for generalisation-2

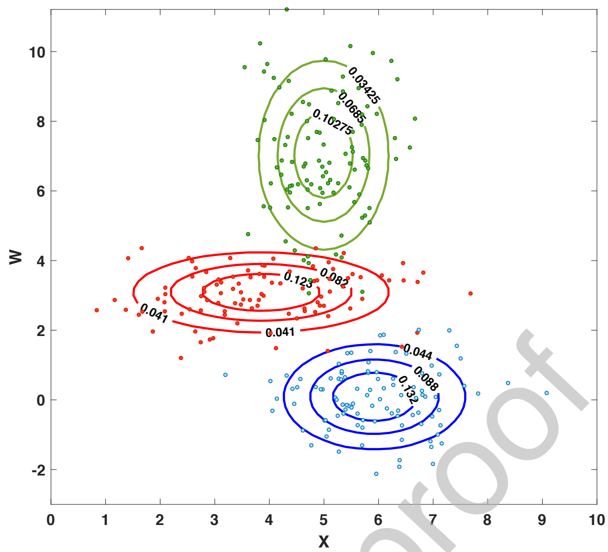
In Figure 3(a), we depict a simple example of three classes with a one-dimensional feature x (in the horizontal axis) and one dimension of the assessment w (in the vertical axis). The joint distribution of the feature and the assessment is thus a three-component mixture of Gaussian distributions. Figure 3(a) shows that the generalisation-2 works in this case. From Figure 3(b), we can observe that the feature's distributions of the three classes seriously overlap. However, with the assessments information added, we can see that the three classes are much more separable, as shown in Figure 3(a).

2.3. Generalisation-3: feature-dependent stochastic supervision

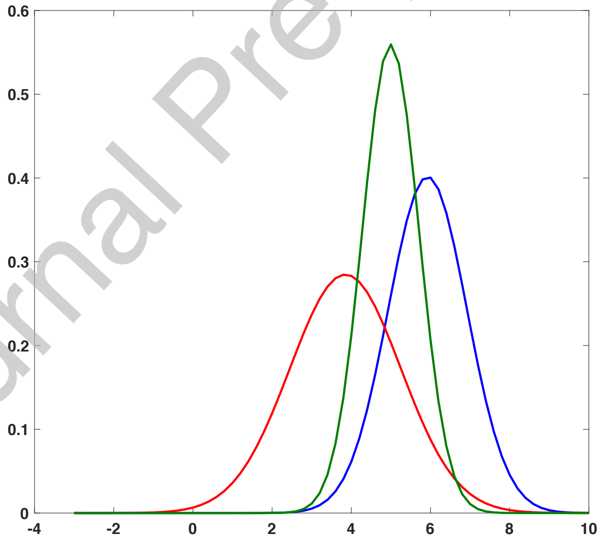
Titterington [16] suggested to generalise the stochastic supervision model to the scenarios that the supervisor's assessment w is dependent on the features x. In the generalisation-3, we assume that there is a linear relationship between the assessment and the features. To check the validity of this assumption, we can calculate the Pearson correlation coefficient between x and w if there is one feature or the adjusted R^2 [23] when regressing w against x for multiple features.

2.3.1. Formulation of generalisation-3

The formulation of this generalisation is quite similar to that of the original stochastic supervision model, except that the distribution of assessment is



(a)



(b)

Figure 3: (a) Joint distribution of feature and (one dimension of) assessment for three classes in red, blue and green, respectively. The contour plots were estimated by the generalisation-2. Each contour is labelled by its corresponding density. (b) Distributions of the feature for three classes in red, blue and green, respectively.

now conditional on the features by replacing $q_j(w)$ with $q_j(w|x)$. This makes the joint distribution of (x_n, w_n) as $p(x_n, w_n) = \sum_{j=1}^{J} \pi_j f_j(x_n) q_j(w_n|x_n)$.

As suggested in [16], a simple way to model $q_j(w_n|x_n)$ is to use the Gaussian distribution $N(\alpha_j + \beta_j^T x_n, \Omega_j)$, and in this case the joint distribution $f_j(x_n, w_n)$ is simply another Gaussian distribution $N(\nu_j, \Psi_j)$, where

$$\nu_j = \begin{pmatrix} \mu_j \\ \alpha_j + \beta_j^T \mu_j \end{pmatrix}, \ \Psi_j = \begin{pmatrix} \Sigma_j & \Sigma_j \beta_j \\ \beta_j^T \Sigma_j & \Omega_j + \beta_j^T \Sigma_j \beta_j \end{pmatrix},$$

α_j is a $(J-1)$-variate vector, and β_j is a $d \times (J-1)$ matrix.

2.3.2. EM algorithm of generalisation-3

E-step. In the E-step, we can compute the responsibilities as

$$\gamma(y_{nj}) = \frac{\pi_j f_j(x_n, w_n)}{\sum_{j=1}^{J} \pi_j f_j(x_n, w_n)}.$$

M-step. In the M-step, we can update ν_j by setting

$$\nu_j = \frac{\sum_{n=1}^{N} \gamma(y_{nj}) a_n}{\sum_{n=1}^{N} \gamma(y_{nj})},$$

where a_n is a concatenated vector of x_n and w_n. Similarly, the updated covariance matrix is

$$\Psi_j = \frac{\sum_{n=1}^{N} \gamma(y_{nj})(a_n - \nu_j)(a_n - \nu_j)^T}{\sum_{n=1}^{N} \gamma(y_{nj})}.$$

2.3.3. Illustrative example for generalisation-3

A simple example of dependent assessment and feature is illustrated in Figure 4. The joint distribution of assessment and feature follows a bivariate Gaussian distribution with positive non-diagonal elements in the covariance matrix. The y-axis in Figure 4 shows the assessment while the x-axis shows

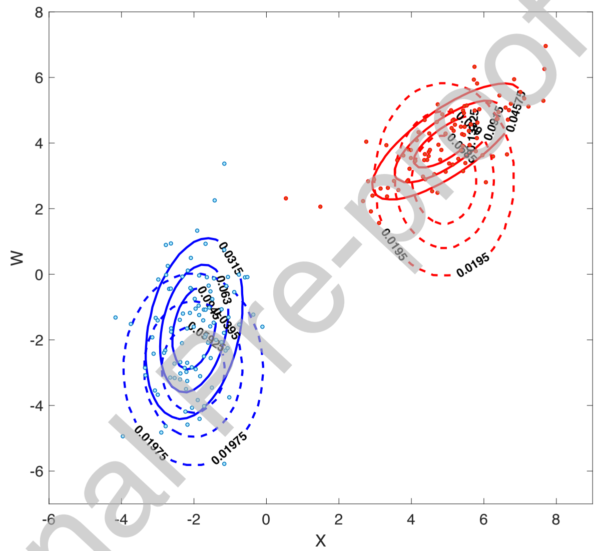


Figure 4: Joint distributions of feature and assessment. Dashed contour plots were estimated by Titterington's original stochastic supervision models. Solid contour plots were estimated by the generalisation-3. Each contour is labelled by its corresponding density.

the feature. The Pearson correlation coefficient between the feature and assessment of the blue class is 0.8378 while that of the red class is 0.2994. It is clear that, compared with Titterington's original model, which assumes the independence between features and assessments, the generalisation-3 fits the joint distribution of the feature and the assessment much better, when they are indeed dependent.

2.4. Generalisation-4: Multi-modal classes

In the original work of Krishnan and Nandy's model [15] and Titterington's model [16] and the three generalisations we have presented, each class is modelled by a Gaussian distribution, implying that there was only a single population for each class, which we call a uni-modal class. In practice, however, the distribution of each class can be much complicated, often having multiple modes, which cannot be described by a standard probabilistic distribution. In this context, we propose our generalisation-4 to model the cases that each class contains multiple subclasses, which makes the class a multi-modal class.

In fact, almost all continuous densities can be approximated with arbitrary accuracy by a mixture of Gaussian distributions [22]. For supervised discriminant analysis, the mixture of Gaussians have been studied well in [24–27]. In the scenario of the stochastic supervision model, which is not deterministically supervised and is itself a mixture of Gaussians, we extend the model to a *mixture of mixtures of Gaussian distributions* [28, 29].

2.4.1. Formulation of generalisation-4

Suppose there are J classes and, for each class j, there are K_j subclasses. The total number of subclasses is $K = \sum_{j=1}^{J} K_j$.

We assume for each subclass the features x follow a Gaussian distribution $N(\mu_{jk}, \Sigma_{jk})$, such that each class can be modelled by a mixture of Gaussian distributions $f_j(x)$: $f_j(x_n) = \sum_{k=1}^{K_j} \phi_{jk} N(\mu_{jk}, \Sigma_{jk})$, where $\phi_{jk} = p(t_{njk} = 1|y_{nj} = 1)$ is the mixing weight of subclass k within class j, and $t_{nj} = (t_{nj1}, \ldots, t_{njK_j})$ is a latent vector, such that $t_{njk} \in \{0, 1\}$ indicating the membership of a subclass belonging to a class, and $\sum_{k=1}^{K_j} t_{njk} = 1$.

Given that the real label is also unknown and the instances were generated from J different classes, we have the distribution of features x as a mixture of J different mixtures $f_j(x)$ of Gaussian distributions: $p(x_n) = \sum_{j=1}^{J} \pi_j f_j(x_n)$, where $\pi_j = p(y_{nj} = 1)$ is the mixing weight of class j in the whole dataset, and $y_n = (y_{n1}, \ldots, y_{nJ})$ is a latent variable vector of real class label such that $y_{nj} \in \{0, 1\}$ and $\sum_{j=1}^{J} y_{nj} = 1$.

Moreover, as before, for each class j, the supervisor's assessment w follows a univariate Gaussian distribution $N(\Delta_j, \Omega_j)$.

The notation for the generalisation-4 can be summarised as

- set of features $X = \{x_n\}$, for $n = 1, \ldots, N$;
- set of the supervisor's assessments $W = \{w_n\}$, for $n = 1, \ldots, N$;
- set of the latent class labels $Y = \{y_n\}$, for $n = 1, \ldots, N$;
- set of the latent subclass labels $T = \{t_{njk}\}$, for $n = 1, \ldots, N$, $j = 1, \ldots, J$, $k = 1, \ldots, K_j\}$;

- set of the class mixing weights $\Pi = \{\pi_j\}$, for $j = 1, \ldots, J$;
- set of the subclass mixing weights $\Phi = \{\phi_{jk}\}$, for $j = 1, \ldots, J$, $k = 1, \ldots, K_j$;
- set of the subclass means $U = \{\mu_{jk}\}$, for $j = 1, \ldots, J$, $k = 1, \ldots, K_j$;
- set of the subclass covariances $\Sigma = \{\Sigma_{jk}\}$, for $j = 1, \ldots, J$, $k = 1, \ldots, K_j$;
- set of the assessment means $\Delta = \{\Delta_j\}$, for $j = 1, \ldots, J$; and
- set of the assessment covariances $\Omega = \{\Omega_j\}$, for $j = 1, \ldots, J$.

We also define $\mathcal{X} = \{X, W\}$, $\mathcal{T} = \{Y, T\}$, and $\theta = \{\Pi, \Phi, U, \Sigma, \Delta, \Omega\}$. The complete-data likelihood becomes

$$p(\mathcal{X}, \mathcal{T}|\theta) = \prod_{n=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K_j}[\pi_j \phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk}) N(w_n|\Delta_j, \Omega_j)]^{y_{nj}t_{njk}},$$

and the marginal likelihood of the features becomes

$$p(\mathcal{X}) = \prod_{n=1}^{N}\sum_{j=1}^{J}\left\{\pi_j N(w_n|\Delta_j, \Omega_j)\sum_{k=1}^{K_j}\phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk})\right\}.$$

2.4.2. EM algorithm of generalisation-4

The EM algorithm to fit the model can be derived as follows.

E-step. In the E-step we can update distribution of latent variables by setting $q^{new}(\mathcal{T}) = p(\mathcal{T}|\mathcal{X}, \theta^{old})$. We can update the class responsibilities by setting $\gamma(y_{nj}) = p(y_{nj} = 1|\mathcal{X}, \theta^{old})$, and the subclass responsibilities by setting $r(t_{njk}) = p(t_{njk} = 1|\mathcal{X}, \theta^{old})$, which lead to

$$\gamma(y_{nj}) = \frac{\sum_{k=1}^{K_j} \pi_j \phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk}) N(w_n|\Delta_j, \Omega_j)}{\sum_{j=1}^{J} \sum_{k=1}^{K_j} \pi_j \phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk}) N(w_n|\Delta_j, \Omega_j)}$$

and

$$r(t_{njk}) = \frac{\pi_j \phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk}) N(w_n|\Delta_j, \Omega_j)}{\sum_{j=1}^{J} \sum_{k=1}^{K_j} \pi_j \phi_{jk} N(x_n|\mu_{jk}, \Sigma_{jk}) N(w_n|\Delta_j, \Omega_j)} .$$

M-step. In the M-step, we can update θ by $\theta^{new} = \arg\max_\theta \sum_{\mathcal{T}} q^{new}(\mathcal{T}) \log p(\mathcal{T}, \mathcal{X}|\theta)$. It follows that

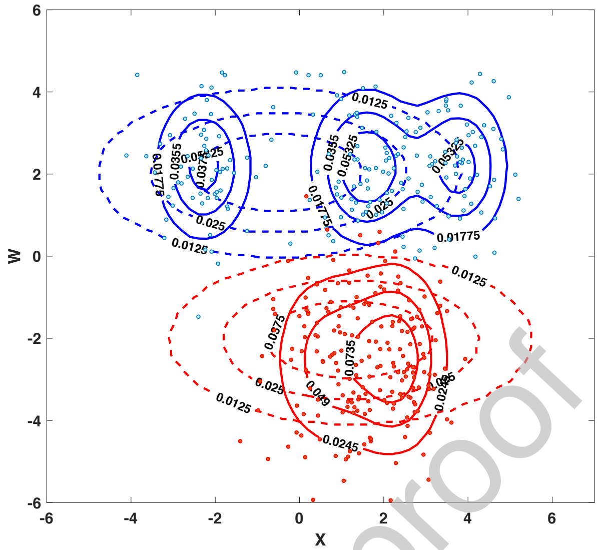
$$\pi_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj})}{N} , \ \phi_{jk}^{new} = \frac{\sum_{n=1}^{N} r(t_{njk})}{\sum_{n=1}^{N} \gamma(y_{nj})} , \ \mu_{jk}^{new} = \frac{\sum_{n=1}^{N} r(t_{njk}) x_n}{\sum_{n=1}^{N} r(t_{njk})} ,$$

$$\Delta_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj}) w_n}{\sum_{n=1}^{N} \gamma(y_{nj})}, \ \Sigma_{jk}^{new} = \frac{\sum_{n=1}^{N} r(t_{njk})(x_n - \mu_{jk})(x_n - \mu_{jk})^T}{\sum_{n=1}^{N} r(t_{njk})} ,$$

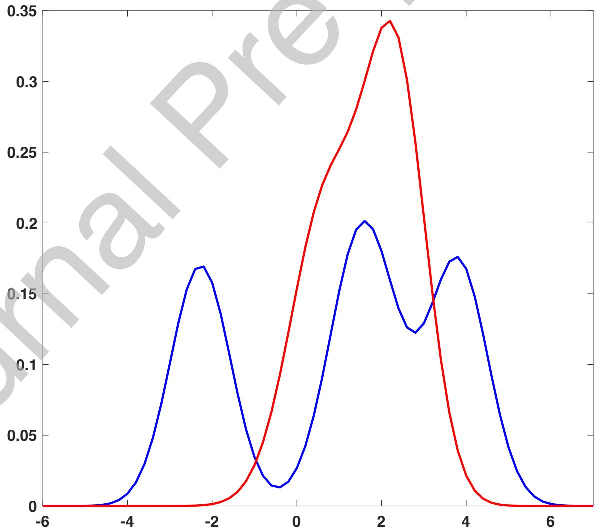
$$\Omega_j^{new} = \frac{\sum_{n=1}^{N} \gamma(y_{nj})(w_n - \Delta_j)(w_n - \Delta_j)^T}{\sum_{n=1}^{N} \gamma(y_{nj})}.$$

2.4.3. Illustrative example for generalisation-4

Figure 5(a) and Figure 5(b) illustrate an example of generalisation-4 for two classes, Class-A with a mixture of two Gaussian subclasses while Class-B with a mixture of three Gaussian subclasses. In this case Class-A and Class-B are difficult to be modelled well by a single Gaussian distribution, if



(a)



(b)

Figure 5: (a) Joint distributions of feature and assessment for two classes with subclasses: Class-A with two subclasses (red); Class-B with three subclasses (blue). Dashed contour plots were estimated by Titterington's original stochastic supervision models. Solid contour plots were estimated by the generalisation-4. Each contour is labelled by its corresponding density. (b) Distributions of feature for two classes with subclasses: Class-A with two subclasses (red); Class-B with three subclasses (blue).

the original Titterington's model is adopted. Our generalisation-4, however, can handle such a complicated dataset, as shown in Figure 5(a). Moreover, comparing Figure 5(a) and Figure 5(b), we can also observe that the data became more separable when the assessment information is added to the model: in Figure 5(b) there is a large overlap between the two classes when only the feature is used while in Figure 5(a) the two groups of points became separable when the feature and assessment are jointly modelled.

3. Real-data experiments

In stochastic supervision, as no deterministic labels were available to training, we cannot compare its classification performance to supervised learning methods such as linear discriminant analysis and support vector machines; on the other hand, it would also be unfairly to favour stochastic supervision if we evaluate it with unsupervised clustering methods such as k-means, given the latter does not even provide any assessment information. Hence we only compare our generalisations with other stochastic supervisors like Titterington's model, the comparison with which has been demonstrated in the previous sections with simulated data, and in the following experiments with real-world data.

In our experiments, the generalisation-1 and the generalisation-2 are not evaluated in the real-data experiments because their asymmetric and multi-class settings are also covered by the generalisation-3 and the generalisation-4.

3.1. Real-world datasets

We use three famous real-world datasets in our experiments: the MNIST dataset [30] is used to evaluate the effectiveness of the generalisation-3, the CIFAR-10 dataset [31] is used to evaluate that of the generalisation-4 and the EMNIST dataset [32] is used to evaluate both generalisations.

In MNIST, we aim to classify handwritten digits 3 and 5, which are hard to distinguish. The assessment and features show strong linear relationship in these two classes, as shown in Table 1. In CIFAR-10, we divide the whole dataset into two large classes: the animal class (which includes bird, cat, deer, dog, frog and horse) and the transportation class (which includes airplane, automobile, ship and truck). This setting is reasonable for the generalisation-4, because the two large classes contain several subclasses. In EMNIST, we aim to classify three large classes: the digits class, the capital letters class and the lower cases class. These three classes have 47 subclasses, including 10 digits subclasses, 26 capital letters subclasses and 11 lowercases subclasses. The linear relationship between the assessment and features are shown in Table 1. Thus the EMNIST data is a mixture of feature-dependent assessments and multi-modal classes and is suitable to test both generalisations 3 and 4.

Table 1: Adjusted R^2 when regressing the assessment against the features for the MNIST and EMNIST datasets.

Dataset	MNIST		EMNIST		
	Digit 5	Digit 3	Capital Letters	Digits	Lowercases
Adjusted R^2	0.9801	0.9585	0.5585	0.6021	0.6050

3.2. Experiment settings

3.2.1. Assessments generation

Considering that stochastic supervision has assessments only and thus is not a supervised learning model, during the model training we need to ignore the labelling information and before the training we need to 'generate' the supervisor's assessments.

For the MNIST data, to generate such assessments we use logistic regression to generate the probabilities that an instance belongs to two classes as appropriate assessments. Note that the dependency between features and assessments in the generalisation-3 is satisfied when such an approach is adopted to generate assessments, because the posterior probabilities generated are dependent on the features. For the EMNIST data with more than two classes, we use Naive Bayes to generate the posterior probabilities as assessments.

Based on the assessments only, a simple intuitive approach to inferring y is to directly compare different elements of assessments. For example, for a two-class problem, let $y = 1$ if $w > 0$ and $y = 0$ otherwise; and for a J-class problem, set $y = \arg\max_{j\in\{1,\dots,J\}} z_j$ (or $y = \arg\max_{j\in\{1,\dots,J-1\}} w_j$ if at least one $w_j > 0$, and $y = J$ otherwise).

3.2.2. Parameters initialisation

Note that in the following initialisation settings, the samples that belong to class j are determined by assessments rather than true labels, because we cannot use true-label information for stochastic supervision methods.

In Titterington's model, the EM algorithm needs initial values of parameters π_j, μ_j, Σ, Δ and Ω. Here we use the sample estimates to initialise these parameters: π_j is the fraction of the estimated number of samples in class j over the total number of samples N, μ_j is the sample mean of the samples, Δ is the sample mean of the assessments of class 1 and $-\Delta$ for class 2, and Σ and Ω are the pooled covariance matrices of the features and the assessments over all J classes, respectively.

In the generalisation-3, α_j and β_j are obtained from the linear regression of the samples in the jth class against their associated w. The EM algorithm of this model needs initial values of π_j, μ_j, Σ_j and Ω_j. We use the same initialisation settings of π_j and μ_j as those for Titterington's model. Similarly, Σ_j and Ω_j are initialised as the sample covariances of the features and the assessments of class j, respectively.

In the generalisation-4, for CIFAR-10 there are 6 subclasses for animal and 4 for transportation and for EMNIST there are 10 subclasses for digits, 26 for capital letters and 11 for lowercases. The EM algorithm of this model needs initial values of the following parameters: π_j, ϕ_{jk} μ_{jk}, Σ_{jk}, Δ_j and Ω_j. The initialisation of π_j and Ω_j is the same as that for the generalisation-3; Δ_j is initialised as the sample mean of the assessments of samples in class j.

To initialise the subclass mean μ_{jk}, covariance matrix Σ_{jk} and mixing weight ϕ_{jk}, we apply k-means to class j: μ_{jk} and Σ_{jk} are set to the subclass means and covariance matrices estimated by k-means on class j, respectively, and ϕ_{jk} is set to the fraction of the number of samples in subclass k of class j over the total number of samples in class j.

3.2.3. Validation settings

In the MNIST dataset, we perform 20 training/test splits; for each split, 70% samples are randomly selected from each class to form the training set and the rest are for the test set. We record the classification accuracies on the test sets for all splits.

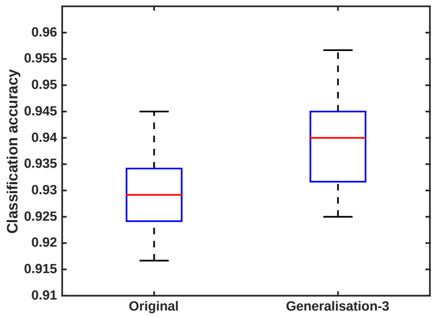
In the CIFAR-10 dataset, we use the training/test split provided by Krizhevsky and Hinton [31], where the training set contains 50000 images with 30000 for the animal class and 20000 for the transportation class and the test set contains 10000 images with 6000 for the animal class and 4000 for the transportation class. For each experiment, we use all the training samples to train the model and randomly select 1000 images from the rest to test. We repeat the procedure 20 times and record the 20 classification accuracies on the test sets. All images are transformed to greyscale in the experiments.

In the EMNIST dataset, the number of training samples is large and using all the samples is time consuming. For illustrative purposes, we randomly sample 1200 images for each subclass, which makes the whole training set contain 1200×47 images. For each experiment, we use all training samples to train the model and randomly select 1000 images from the rest to test. We repeat the procedure 20 times and record the 20 classification accuracies on the test sets. The pixel values of the margin part of images in EMNIST are zeros, which leads to singular covariance matrices. Thus we add small white noises to these images to make the covariance matrices invertible. Since Titterington's model is used for binary classification and we have three classes here, the one-versus-all strategy [33] is applied here for Titterington's model.

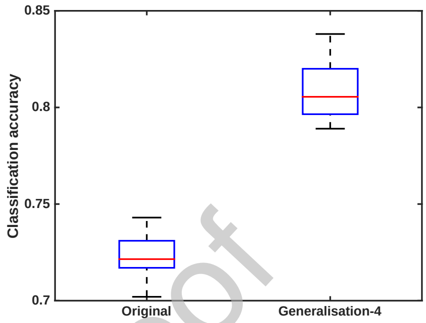
3.3. Results

Classification accuracies on the 20 test sets of MNIST, CIFAR-10 and EMNIST are boxplotted in Figure 6(a), Figure 6(b) and Figure 6(c), respectively. It is clear that the generalisation-3 and the generalisation-4 have higher boxes than Titterington's model in Figure 6(a) and Figure 6(b). This indicates the effectiveness of our generalisations when the data satisfy the associated conditions: in our experiments, the MNIST dataset satisfies the feature-assessment dependency condition in the generalisation-3 and the CIFAR-10 dataset satisfies the multi-modality condition in the generalisation-4.

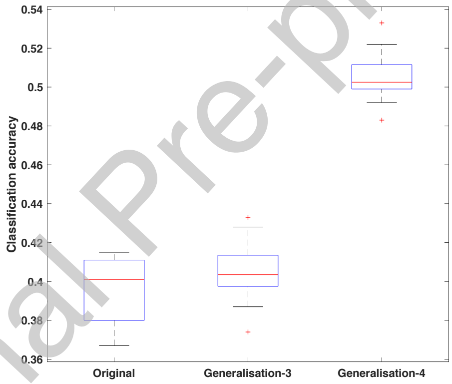
For the EMNIST data, the generalisation-3 and generalisation-4 produce higher boxes than Titterington's model and the generalisation-4 has the best classification performance. This also shows the effectiveness of our models. Note that here the generalisation-4 has much better classification performance than the generalisation-3. One possible reason is that the multi-modal classes have more effect on the final results than the feature-dependent assessment, since the subclasses in each large class are clearly defined while the linear relationship between the assessment and features is not strong, as shown in Table 1. We also note that there is a large space for improvement in classification accuracy of EMNIST. By developing a new method that can deal with feature-dependent assessments and multi-modal classes together, we may further improve the classification performance on complex data such as EMNIST. We list this as our future work in the conclusions section.



(a) MNIST



(b) CIFAR-10



(c) EMNIST

Figure 6: (a) Classification accuracies of Titterington's model and the generalisation-3 on 20 test sets of MNIST. (b) Classification accuracies of Titterington's model and the generalisation-4 on 20 test sets of CIFAR-10. (c) Classification accuracies of Titterington's model, generalisation-3 and generalisation-4 on 20 test sets of EMNIST.

4. Conclusions

In this paper, we extended stochastic supervision models in four aspects, generalising them to asymmetric assessments, multiple classes, feature-dependent assessments and multi-modal classes, respectively, to enhance their applicability. The experiments on both simulated data and real-world data demonstrate the effectiveness of our generalisations. In the future, to enhance further our models' flexibility and generality, we shall explore non-linear modelling for the relationship between assessments and features, as well as more sophisticated techniques for multi-modality modelling. Moreover, instead of using a fixed threshold of w to infer y, we propose to learn this threshold from data. Since we use the transformation $w_i = \log z_i/z_J$ to transform a softmax vector to a $(J-1)$ dimensional normal distributed random variable, learning the threshold of w is equivalent to giving different weights to different classes. By utilising the learned threshold, our model can adapt to more real-world scenarios where different classes have different importance. In addition, we propose to develop new algorithms that can provide superior classification performances under more complex situations, e.g. with both feature-dependent assessment and multi-modal classes.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grant 61628301. We thank the reviewers for their constructive comments to improve our manuscript.

References

[1] G. J. McLachlan, Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis, Journal of the American Statistical Association 70 (350) (1975) 365–369.

[2] T. J. O'neill, Normal discrimination with unclassified observations, Journal of the American Statistical Association 73 (364) (1978) 821–826.

[3] F. Schwenker, E. Trentin, Pattern classification and clustering: A review of partially supervised learning approaches, Pattern Recognition Letters 37 (2014) 4–14.

[4] F. Schwenker, E. Trentin, Partially supervised learning for pattern recognition, Pattern Recognition Letters 37 (2014) 1–3.

[5] D. Ahfock, G. J. McLachlan, On missing label patterns in semi-supervised learning, arXiv preprint arXiv:1904.02883.

[6] X. Zhu, A. B. Goldberg, Introduction to Semi-Supervised Learning, Morgan and Claypool Publishers, 2009.

[7] O. Chapelle, B. Schlkopf, A. Zien, Semi-Supervised Learning, The MIT Press, 2010.

[8] C. Chittineni, Learning with imperfectly labeled patterns, Pattern Recognition 12 (5) (1980) 281–291.

[9] T. Krishnan, Efficiency of learning with imperfect supervision, Pattern Recognition 21 (2) (1988) 183–188.

[10] U. Katre, T. Krishnan, Pattern recognition with an imperfect supervisor, Pattern recognition 22 (4) (1989) 423–431.

[11] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE transactions on neural networks and learning systems 25 (5) (2014) 845–869.

[12] C. Bouveyron, S. Girard, Robust supervised classification with mixture models: Learning from data with uncertain labels, Pattern Recognition 42 (11) (2009) 2649–2658.

[13] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications, Pattern Recognition 77 (2018) 329–353.

[14] J. Aitchison, C. B. Begg, Statistical diagnosis when basic cases are not classified with certainty, Biometrika 63 (1) (1976) 1–12.

[15] T. Krishnan, S. C. Nandy, Discriminant analysis with a stochastic supervisor, Pattern Recognition 20 (4) (1987) 379–384.

[16] D. M. Titterington, An alternative stochastic supervisor in discriminant analysis, Pattern Recognition 22 (1) (1989) 91–95.

[17] T. Krishnan, S. C. Nandy, Efficiency of discriminant analysis when initial samples are classified stochastically, Pattern Recognition 23 (5) (1990) 529–537.

[18] T. Krishnan, S. C. Nandy, Efficiency of logistic-normal stochastic supervision, Pattern Recognition 23 (11) (1990) 1275–1279.

[19] D. M. Titterington, Some recent research in the analysis of mixture distributions, Statistics 21 (4) (1990) 619–641.

[20] J. Aitchison, The Statistical Analysis of Compositional Data, Chapman & Hall, 1986.

[21] G. McLachlan, T. Krishnan, The EM algorithm and Extensions, John Wiley & Sons, 2007.

[22] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006.

[23] H. Theil, Economic forecasts and policy, North-Holland Pub. Co., 1961.

[24] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 155–176.

[25] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American Statistical Association 97 (458) (2002) 611–631.

[26] G. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, 2004.

[27] G. J. McLachlan, S. X. Lee, S. I. Rathnayake, Finite mixture models, Annual Review of Statistics and Its Application 6 (2019) 355–378.

[28] R. P. Browne, P. D. McNicholas, M. D. Sparling, Model-based learning using a mixture of mixtures of gaussian and uniform distributions, IEEE

Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2011) 814–817.

[29] C. Viroli, G. J. McLachlan, Deep gaussian mixture models, Statistics and Computing 29 (1) (2019) 43–51.

[30] Y. LeCun, C. Cortes, MNIST handwritten digit database. URL http://yann.lecun.com/exdb/mnist/

[31] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Tech. rep., Citeseer (2009).

[32] G. Cohen, S. Afshar, J. Tapson, A. Van Schaik, EMNIST: Extending MNIST to handwritten letters, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017, pp. 2921–2926.

[33] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, Vol. 112, Springer, 2013.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author biography:

- Xiao-Ou Lu received the M.Sc degree in computational statistics and machine learning from University College London, London, U.K., in 2014. He is currently pursuing the Ph.D. degree at the Department of Statistical Science, University College London. His research interests include deep generative model, variational inference and domain adaptation.
- Yangqi Qiao received the M.Sc degree in statistics from University College London in 2016. He is currently a corporate account manager in Agricultural Bank of China, Zhejiang.
- Rui Zhu received the Ph.D. degree in statistics from University College London in 2017. She is a lecturer in the Faculty of Actuarial Science and Insurance, City, University of London. Her research interests include spectral data analysis, hyperspectral image analysis, subspace-based classification methods and image quality assessment.
- Guijin Wang received the B.S. degree and Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, China in 1998 and 2003, respectively. In 2003, he joined the Information Technologies Laboratories, Sony Corporation, Japan as a researcher. Since 2006, he has been with the Department of Electronic Engineering, Tsinghua University, China as an associate professor.
- Zhanyu Ma received the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013 he was a post-doctoral research fellow with the School of Electrical Engineering, KTH Royal Institute of Technology. He has been an associate professor with the Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He has also been an adjunct associate professor with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics. He is a senior member of IEEE.

- Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998, and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a senior lecturer in the Department of Statistical Science, University College London. His research interests include statistical machine learning, high-dimensional data analysis, pattern recognition and image analysis.