



City Research Online

City St George's, University of London

Citation: Lu, X., Qiao, Y., Zhu, R., Wang, G., Ma, Z. & Xue, J-H. (2021). Generalisations of stochastic supervision models. *Pattern Recognition*, 109, 107575. doi: 10.1016/j.patcog.2020.107575

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/24968/>

Link to published version: <https://doi.org/10.1016/j.patcog.2020.107575>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

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 two 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 and the CIFAR-10 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

22 classification scheme called stochastic supervision, which, in contrast to all
23 the cases aforementioned, provides no deterministic labels $\{y_n\}_{n=1}^N$ but only
24 probabilistic assessments $\{z_n\}_{n=1}^N$ for inference of y . In other words, only
25 some side information about the output is provided.

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

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

46 In [15], the class-dependent joint data-generating distribution $f_j(x, z)$ was

47 further factorised as $f_j(x, z) = f_j(x)q_j(z)$, by assuming that the features
 48 x and the assessment z are independent of each other in each class. By
 49 supposing the features x are continuous random variables in the range of
 50 $(-\infty, \infty)$, it was assumed that $x|y = 1 \sim N(\mu_1, \Sigma)$ and $x|y = 2 \sim N(\mu_2, \Sigma)$,
 51 two class-dependent d -variate Gaussian distributions. We denote the pdfs of
 52 $x|y = 1$ and $x|y = 2$ as $f_1(x)$ and $f_2(x)$, respectively. In the meantime, as
 53 the probabilistic assessment z is a continuous random variable in the range
 54 of $[0, 1]$, it was assumed that $z|y = 1 \sim \text{Beta}(a, b)$ and $z|y = 2 \sim \text{Beta}(b, a)$,
 55 two Beta distributions symmetric between the two classes. We denote the
 56 pdfs of $z|y = 1$ and $z|y = 2$ as $q_1(z)$ and $q_2(z)$, respectively. That is to say,
 57 the model in [15] assumes that the data-generating process in class j follows
 58 a Gaussian distribution $f_j(x)$ for features x and a Beta distribution $q_j(z)$ for
 59 assessment z . Although the assessment z is given for each training instance
 60 x , the real label (denoted by y) is unknown, which leads the likelihood of
 61 the training instance, or say the joint distribution of x and z , as $p(x, z) =$
 62 $\pi_1 f_1(x, z) + \pi_2 f_2(x, z)$. Hence this is a latent variable (finite mixture) problem,
 63 and the model was fitted by an EM algorithm in [15].

64 However, there are two technical issues with Krishnan and Nandy’s stochas-
 65 tic supervision model. Firstly, it cannot accept any assessment that $z > 1$
 66 or $z < 0$, while in some real problems the assessment can be a random vari-
 67 able in the range of $(-\infty, \infty)$. Secondly, the EM algorithm for this model is
 68 complicated, because there is no exact solution in the M-step for the estima-
 69 tion of certain parameters due to the adoption of the Beta distributions for
 70 assessment z .

71 In order to overcome the two issues above, Titterington [16] introduced

72 a new supervisor’s assessment $w = \log \frac{z}{1-z}$ to replace the original z . This
73 transformation is called additive logistic transformation [20], which extends
74 the range of the assessment from $[0, 1]$ to the real line and thus the assess-
75 ment can be modelled by Gaussian distributions. In Titterington’s model,
76 supervisor assessments $q_1(w)$ and $q_2(w)$ are assumed to follow two univariate
77 Gaussian distributions $N(-\Delta, \Omega)$ and $N(\Delta, \Omega)$, respectively, where $\Delta > 0$
78 and $\Omega > 0$. In this model, the constraints of equal variances and symme-
79 try in the assessment distributions between the two classes are preserved.
80 Then Titterington [16] provided an EM algorithm to estimate parameters
81 $\{\pi_1, \mu_1, \mu_2, \Sigma, \Omega, \Delta\}$.

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

- 88 1. *Asymmetric assessments.* In both Krishnan and Nandy’s and Titter-
89 ington’s models, the two class-dependent distributions of assessments
90 $q_j(z)$ (or $q_j(w)$) were symmetric and with equal variances. Our first
91 generalisation aims to relax this restriction on the parameter setting of
92 supervisor’s assessments.
- 93 2. *Multiple classes.* The past models were for two-class discrimination.
94 Our second generalisation is designed for classification of multiple classes.
- 95 3. *Feature-dependent assessments.* In Krishhan and Nandy’s [15] and Tit-
96 ington’s [16] work, the assessment and the features were modelled

97 independent of each other. Our third generalisation aims to model their
98 dependence.

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

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

108 2. Generalised models and their EM algorithms

109 2.1. Generalisation-1: asymmetric stochastic supervision

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

117 2.1.1. Formulation of generalisation-1

118 Our notation is established as follows. The observable dataset is denoted
119 by $\mathcal{X} = \{X, W\}$, the latent variable set by $\mathcal{Y} = \{Y\}$, and the parameter set

120 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
 121 $Y = \{y_n\}$, for $n = 1, \dots, N$, are N instances, assessments and real labels
 122 of the instances, respectively. For each instance, $y_n = (y_{n1}, y_{n2})$ is a latent
 123 variable vector (representing its real label) such that for class j we have
 124 $y_{nj} \in \{0, 1\}$ and for two classes together we have $\sum_{j=1}^2 y_{nj} = 1$. That is, y_n
 125 is a latent indicator vector with only one element being true.

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

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

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

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

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

139 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

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

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

141 *M-step.* In the M-step, we take partial differential of $l(\theta) = Q(\theta, \theta^{old})$ with
 142 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
 143 obtain updated parameters θ^{new} . It follows that

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

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

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

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

149 Also, the updated covariance matrix of the features is

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

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

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

152 Since the two mixing weights have to satisfy $\pi_0 + \pi_1 = 1$, we can set
 153 $\partial l(\theta)/\partial \pi_j + \lambda = 0$, where λ is a Lagrange multiplier. It then follows that
 154 $\pi_1^{new} = \frac{1}{N} \sum_{n=1}^N \gamma(y_{n1})$, $\pi_2^{new} = 1 - \pi_1^{new}$, indicating that each of the updated
 155 mixing weights is an average of the responsibilities.

156 *2.1.3. Illustrative example for generalisation-1*

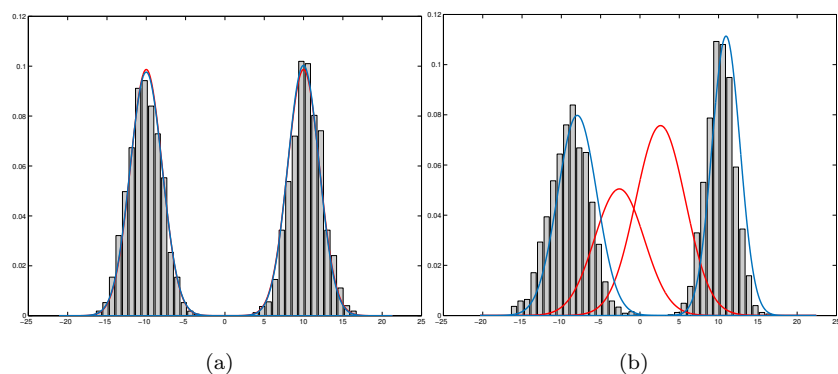


Figure 1: (a) Supervisor assessments with *equal* variances and *symmetrical* means between the two classes. Red curve: assessments density estimated by Titterington’s model. Blue curve: assessments density estimated by the generalisation-1. (b) Supervisor assessments with *unequal* variances and *asymmetrical*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

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

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

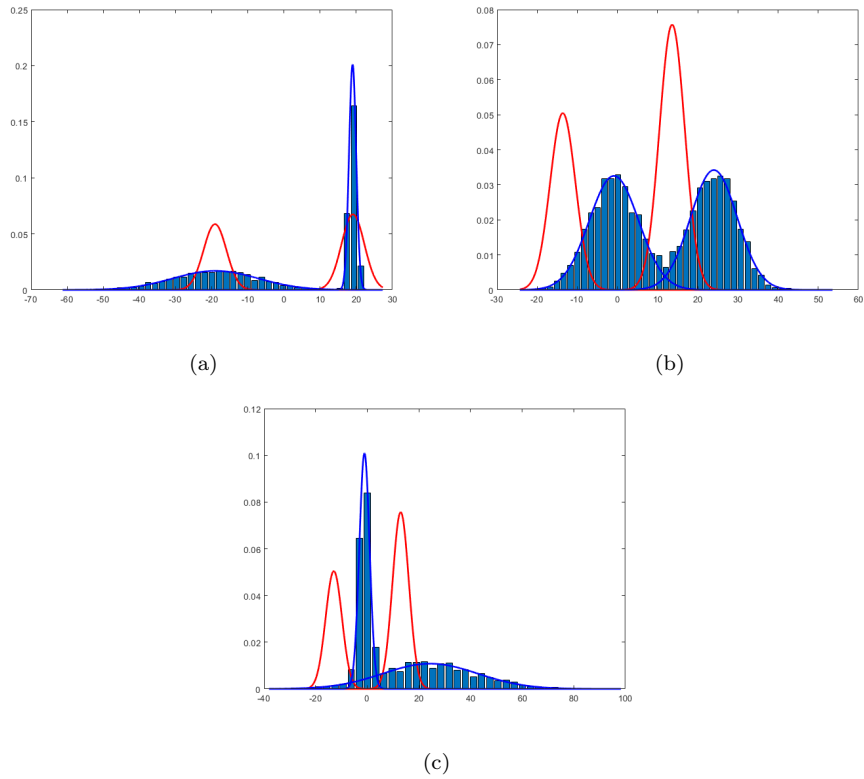


Figure 2: Three extreme cases of supervisor assessments. (a) Supervisor assessments with large *unequal* variances and *symmetrical* means between the two classes. Red curve: assessments density estimated by Titterington's model. Blue curve: assessments density estimated by the generalisation-1. (b)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

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

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

179 2.2. Generalisation-2: multi-class stochastic supervision

180 Original stochastic supervision models were only for two-class discrim-
181 ination. In practice multi-class classification problems are also prevailing.
182 Hence here we extend Titterington’s model to multi-class cases, as suggested
183 by Titterington [16].

184 2.2.1. Formulation of generalisation-2

185 Suppose there are J classes. As with [16], the supervisor’s assessment of
186 an instance x is now a J -variate vector of ‘probabilities’, $z = (z_1, \dots, z_J)$,
187 and we can define a new assessment vector $w_j = \log \frac{z_j}{z_J}$ for $j = 1, \dots, J - 1$,
188 which extends the supervisor’s assessments from $(0, 1)$ to $(-\infty, \infty)$. Then we

189 can assume that, for each class j , the assessments $w = (w_1, \dots, w_{J-1})$ follow
 190 $(J - 1)$ -variate Gaussian distributions: $q_j(w) \sim N(\Delta_j, \Omega_j)$, where $q_j(w)$ is
 191 the pdf of $w|y = j$.

192 Then, given the real label $y_n = (y_{n1}, \dots, y_{nJ})$ is unknown, the joint dis-
 193 tribution of the observed features x_n and assessment w_n of the n th instance
 194 becomes $p(x_n, w_n) = \sum_{j=1}^J \pi_j f_j(x_n, w_n)$, where $f_j(x_n, w_n) = f_j(x_n)q_j(w_n)$
 195 and $\pi_j = p(y_{nj} = 1)$ is the mixing weight of class j .

196 Before going further, we recall some notation to be used for the generalisation-
 197 2:

- 198 • set of the latent labels $Y = \{y_n\}$, for $n = 1, \dots, N$, where y_n is a
 199 J -variate latent vector of real labels, and we have $y_{nj} \in \{0, 1\}$ and
 200 $\sum_{j=1}^J y_{nj} = 1$;
- 201 • set of the class mixing weights $\Pi = \{\pi_j\}$, for $j = 1, \dots, J$, where π_j is
 202 a scalar;
- 203 • set of the class means $U = \{\mu_j\}$, for $j = 1, \dots, J$, where μ_j is a d -variate
 204 vector;
- 205 • set of the class covariances $\Sigma = \{\Sigma_j\}$, for $j = 1, \dots, J$, where Σ_j is a
 206 $d \times d$ matrix;
- 207 • set of the assessment means $\Delta = \{\Delta_j\}$, for $j = 1, \dots, J$, where Δ_j is a
 208 $(J - 1)$ -variate vector; and
- 209 • set of the assessment covariances $\Omega = \{\Omega_j\}$, for $j = 1, \dots, J$, where Ω_j
 210 is a $(J - 1) \times (J - 1)$ matrix.

211 In this notation, the parameter set for the generalisation-2 is $\theta = \{\Pi, U, \Sigma, \Delta, \Omega\}$;
 212 the complete-data likelihood of observed data \mathcal{X} and latent data \mathcal{Y} is $p(\mathcal{Y}, \mathcal{X}|\theta) =$
 213 $\prod_{n=1}^N \sum_{j=1}^J y_{nj} [\pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)]$, and the marginal likelihood of
 214 observed data \mathcal{X} is $p(\mathcal{X}|\theta) = \prod_{n=1}^N \sum_{j=1}^J \pi_j N(x_n|\mu_j, \Sigma_j) N(w_n|\Delta_j, \Omega_j)$.

215 *2.2.2. EM algorithm of generalisation-2*

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

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

218 we have the class responsibilities as

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

219

220 *M-step.* In the M-step, we update θ by $\theta^{new} = \arg \max_{\theta} \sum_{\mathcal{Y}} q^{new}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}|\theta)$.

221 Since the mixing weights π_j satisfy the sum-to-one constraint, as in section 2.1

222 we introduce a Lagrange multiplier λ and set $\partial l(\theta)/\partial \pi_j + \lambda(\sum_{j=1}^J \pi_j - 1) = 0$,

223 which results in the updated mixing weights as $\pi_j^{new} = \frac{1}{N} \sum_{n=1}^N \gamma(y_{nj})$, which is

224 again an average of the responsibilities over all the data points. Similarly to

225 the M-step in section 2.1, we can obtain the updated means and covariance

226 matrices as

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

227

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

228

229 2.2.3. Illustrative example for generalisation-2

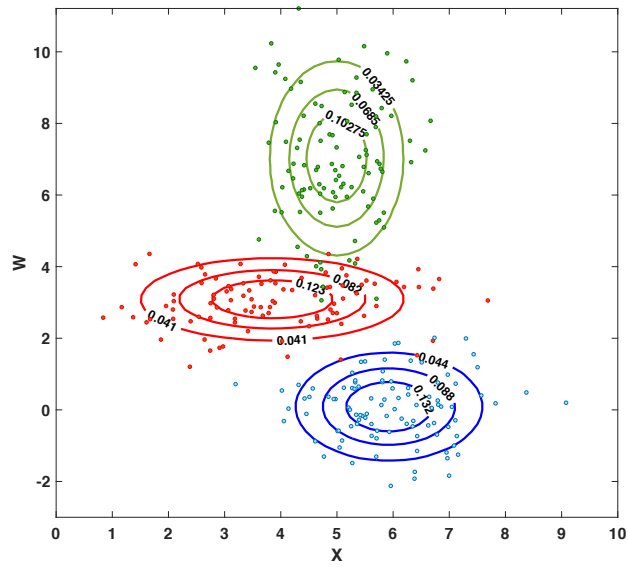
230 In Figure 3(a), we depict a simple example of three classes with a one-
 231 dimensional feature x (in the horizontal axis) and one dimension of the as-
 232 sessment w (in the vertical axis). The joint distribution of the feature and
 233 the assessment is thus a three-component mixture of Gaussian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

238 2.3. Generalisation-3: feature-dependent stochastic supervision

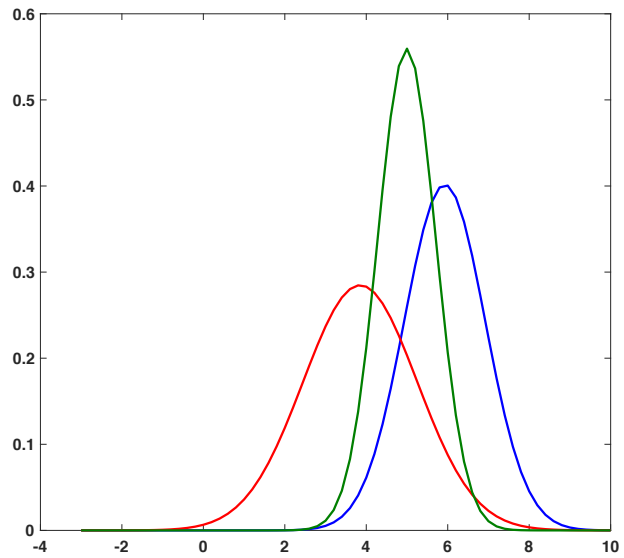
239 Titterington [16] suggested to generalise the stochastic supervision model
 240 to the scenarios that the supervisor's assessment w is dependent on the fea-
 241 tures x . In the generalisation-3, we assume that there is a linear relationship
 242 between the assessment and the features. To check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

246 2.3.1. Formulation of generalisation-3

247 The formulation of this generalisation is quite similar to that of the origi-
 248 nal 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.

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

251 As suggested in [16], a simple way to model $q_j(w_n|x_n)$ is to use the Gaus-
 252 sian distribution $N(\alpha_j + \beta_j^T x_n, \Omega_j)$, and in this case the joint distribution
 253 $f_j(x_n, w_n)$ is simply another Gaussian distribution $N(\nu_j, \Psi_j)$, where

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

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

255 2.3.2. EM algorithm of generalisation-3

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

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

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

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

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

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

260 2.3.3. Illustrative example for generalisation-3

261 A simple example of dependent assessment and feature is illustrated in
 262 Figure 4. The joint distribution of assessment and feature follows a bivariate
 263 Gaussian distribution with positive non-diagonal elements in the covariance
 264 matrix. The y-axis in Figure 4 shows the assessment while the x-axis shows
 265 the feature. The Pearson correlation coefficient between the feature and

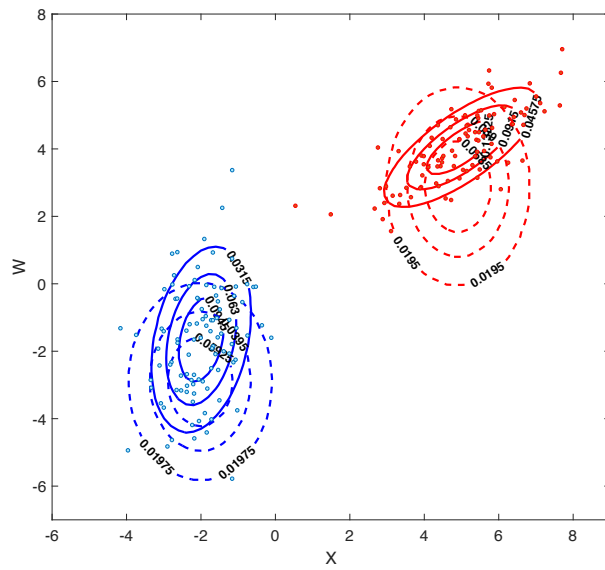


Figure 4: Joint distributions of feature and assessment. Dashed contour plots were estimated by Titterington's original stochastic supervision models. Solid contour plots were estimated by the generalisation-3. Each contour is labelled by its corresponding density.

266 assessment of the blue class is 0.8378 while that of the red class is 0.2994.
267 It is clear that, compared with Titterington’s original model, which assumes
268 the independence between features and assessments, the generalisation-3 fits
269 the joint distribution of the feature and the assessment much better, when
270 they are indeed dependent.

271 2.4. Generalisation-4: Multi-modal classes

272 In the original work of Krishnan and Nandy’s model [15] and Tittering-
273 ton’s model [16] and the three generalisations we have presented, each class
274 is modelled by a Gaussian distribution, implying that there was only a sin-
275 gle population for each class, which we call a uni-modal class. In practice,
276 however, the distribution of each class can be much complicated, often hav-
277 ing multiple modes, which cannot be described by a standard probabilistic
278 distribution. In this context, we propose our generalisation-4 to model the
279 cases that each class contains multiple subclasses, which makes the class a
280 multi-modal class.

281 In fact, almost all continuous densities can be approximated with arbi-
282 trary accuracy by a mixture of Gaussian distributions [22]. For supervised
283 discriminant analysis, the mixture of Gaussians have been studied well in [24–
284 27]. In the scenario of the stochastic supervision model, which is not deter-
285 ministically supervised and is itself a mixture of Gaussians, we extend the
286 model to a *mixture of mixtures of Gaussian distributions* [28, 29].

287 2.4.1. Formulation of generalisation-4

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

290 We assume for each subclass the features x follow a Gaussian distribution
 291 $N(\mu_{jk}, \Sigma_{jk})$, such that each class can be modelled by a mixture of Gaussian
 292 distributions $f_j(x)$: $f_j(x_n) = \sum_{k=1}^{K_j} \phi_{jk} N(\mu_{jk}, \Sigma_{jk})$, where $\phi_{jk} = p(t_{njk} =$
 293 $1 | y_{nj} = 1)$ is the mixing weight of subclass k within class j , and $t_{nj} =$
 294 $(t_{nj1}, \dots, t_{njK_j})$ is a latent vector, such that $t_{njk} \in \{0, 1\}$ indicating the
 295 membership of a subclass belonging to a class, and $\sum_{k=1}^{K_j} t_{njk} = 1$.

296 Given that the real label is also unknown and the instances were generated
 297 from J different classes, we have the distribution of features x as a mixture of
 298 J different mixtures $f_j(x)$ of Gaussian distributions: $p(x_n) = \sum_{j=1}^J \pi_j f_j(x_n)$,
 299 where $\pi_j = p(y_{nj} = 1)$ is the mixing weight of class j in the whole dataset,
 300 and $y_n = (y_{n1}, \dots, y_{nJ})$ is a latent variable vector of real class label such that
 301 $y_{nj} \in \{0, 1\}$ and $\sum_{j=1}^J y_{nj} = 1$.

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

304 The notation for the generalisation-4 can be summarised as

- 305 • set of features $X = \{x_n\}$, for $n = 1, \dots, N$;
- 306 • set of the supervisor's assessments $W = \{w_n\}$, for $n = 1, \dots, N$;
- 307 • set of the latent class labels $Y = \{y_n\}$, for $n = 1, \dots, N$;
- 308 • set of the latent subclass labels $T = \{t_{njk}\}$, for $n = 1, \dots, N$, $j =$
 309 $1, \dots, J$, $k = 1, \dots, K_j$;
- 310 • set of the class mixing weights $\Pi = \{\pi_j\}$, for $j = 1, \dots, J$;
- 311 • set of the subclass mixing weights $\Phi = \{\phi_{jk}\}$, for $j = 1, \dots, J$, $k =$
 312 $1, \dots, K_j$;

- 313 • set of the subclass means $U = \{\mu_{jk}\}$, for $j = 1, \dots, J$, $k = 1, \dots, K_j$;
- 314 • set of the subclass covariances $\Sigma = \{\Sigma_{jk}\}$, for $j = 1, \dots, J$, $k =$
- 315 $1, \dots, K_j$;
- 316 • set of the assessment means $\Delta = \{\Delta_j\}$, for $j = 1, \dots, J$; and
- 317 • set of the assessment covariances $\Omega = \{\Omega_j\}$, for $j = 1, \dots, J$.

318 We also define $\mathcal{X} = \{X, W\}$, $\mathcal{T} = \{Y, T\}$, and $\theta = \{\Pi, \Phi, U, \Sigma, \Delta, \Omega\}$.

319 The complete-data likelihood becomes

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

320 and the marginal likelihood of the features becomes

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

321

322 2.4.2. EM algorithm of generalisation-4

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

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

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

328 and

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

329

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

331 It follows that

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

332

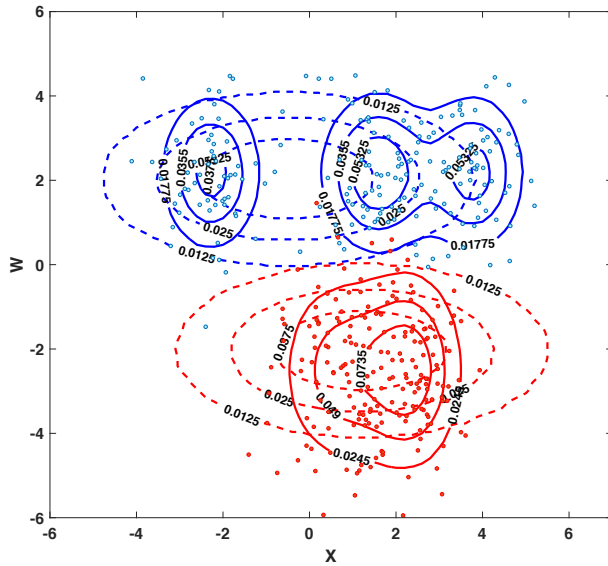
$$\Delta_j^{new} = \frac{\sum_{n=1}^N \gamma(y_{nj})w_n}{\sum_{n=1}^N \gamma(y_{nj})}, \Sigma_{jk}^{new} = \frac{\sum_{n=1}^N r(t_{nj})(x_n - \mu_{jk})(x_n - \mu_{jk})^T}{\sum_{n=1}^N r(t_{nj})},$$

333

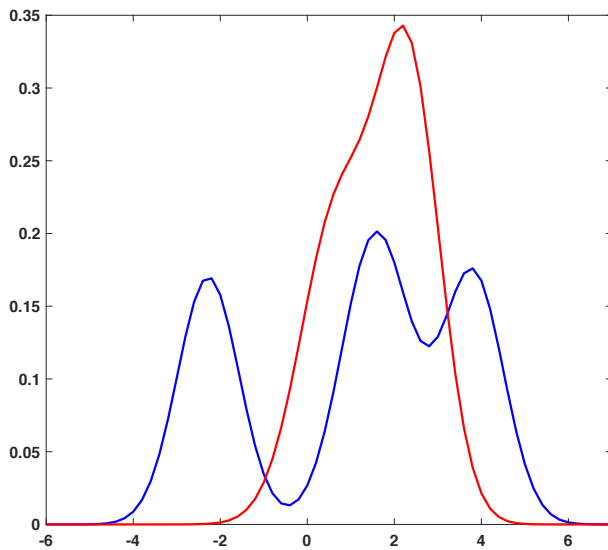
$$\Omega_j^{new} = \frac{\sum_{n=1}^N \gamma(y_{nj})(w_n - \Delta_j)(w_n - \Delta_j)^T}{\sum_{n=1}^N \gamma(y_{nj})}.$$

334 2.4.3. Illustrative example for generalisation-4

335 Figure 5(a) and Figure 5(b) illustrate an example of generalisation-4 for
 336 two classes, Class-A with a mixture of two Gaussian subclasses while Class-
 337 B with a mixture of three Gaussian subclasses. In this case Class-A and
 338 Class-B are difficult to be modelled well by a single Gaussian distribution, if
 339 the original Titterington's model is adopted. Our generalisation-4, however,
 340 can handle such a complicated dataset, as shown in Figure 5(a). Moreover,
 341 comparing Figure 5(a) and Figure 5(b), we can also observe that the data
 342 became more separable when the assessment information is added to the
 343 model: in Figure 5(b) there is a large overlap between the two classes when
 344 only the feature is used while in Figure 5(a) the two groups of points became
 345 separable when the feature and assessment are jointly modelled.



(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 Titterton'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).

346 **3. Real-data experiments**

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

357 In our experiments, the generalisation-1 and the generalisation-2 are not
358 evaluated in the real-data experiments because their asymmetric and multi-
359 class settings are also covered by the generalisation-3 and the generalisation-
360 4.

361 *3.1. Real-world datasets*

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

366 In MNIST, we aim to classify handwritten digits 3 and 5, which are hard
367 to distinguish. The assessment and features show strong linear relationship
368 in these two classes, as shown in Table 1. In CIFAR-10, we divide the whole
369 dataset into two large classes: the animal class (which includes bird, cat, deer,

370 dog, frog and horse) and the transportation class (which includes airplane,
 371 automobile, ship and truck). This setting is reasonable for the generalisation-
 372 4, because the two large classes contain several subclasses. In EMNIST, we
 373 aim to classify three large classes: the digits class, the capital letters class
 374 and the lower cases class. These three classes have 47 subclasses, including 10
 375 digits subclasses, 26 capital letters subclasses and 11 lowercases subclasses.
 376 The linear relationship between the assessment and features are shown in
 377 Table 1. Thus the EMNIST data is a mixture of feature-dependent assess-
 378 ments and multi-modal classes and is suitable to test both generalisations 3
 379 and 4.

380

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

381 3.2. Experiment settings

382 3.2.1. Assessments generation

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

387 For the MNIST data, to generate such assessments we use logistic regres-
 388 sion to generate the probabilities that an instance belongs to two classes as

389 appropriate assessments. Note that the dependency between features and
 390 assessments in the generalisation-3 is satisfied when such an approach is
 391 adopted to generate assessments, because the posterior probabilities gener-
 392 ated are dependent on the features. For the EMNIST data with more than
 393 two classes, we use Naive Bayes to generate the posterior probabilities as
 394 assessments.

395 Based on the assessments only, a simple intuitive approach to inferring y
 396 is to directly compare different elements of assessments. For example, for a
 397 two-class problem, let $y = 1$ if $w > 0$ and $y = 0$ otherwise; and for a J -class
 398 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
 399 one $w_j > 0$, and $y = J$ otherwise).

400 3.2.2. Parameters initialisation

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

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

411 In the generalisation-3, α_j and β_j are obtained from the linear regression
 412 of the samples in the j th class against their associated w . The EM algorithm
 413 of this model needs initial values of π_j , μ_j , Σ_j and Ω_j . We use the same ini-

414 tialisation settings of π_j and μ_j as those for Titterington’s model. Similarly,
415 Σ_j and Ω_j are initialised as the sample covariances of the features and the
416 assessments of class j , respectively.

417 In the generalisation-4, for CIFAR-10 there are 6 subclasses for animal
418 and 4 for transportation and for EMNIST there are 10 subclasses for digits,
419 26 for capital letters and 11 for lowercases. The EM algorithm of this model
420 needs initial values of the following parameters: π_j , ϕ_{jk} , μ_{jk} , Σ_{jk} , Δ_j and Ω_j .
421 The initialisation of π_j and Ω_j is the same as that for the generalisation-3;
422 Δ_j is initialised as the sample mean of the assessments of samples in class j .
423 To initialise the subclass mean μ_{jk} , covariance matrix Σ_{jk} and mixing weight
424 ϕ_{jk} , we apply k -means to class j : μ_{jk} and Σ_{jk} are set to the subclass means
425 and covariance matrices estimated by k -means on class j , respectively, and
426 ϕ_{jk} is set to the fraction of the number of samples in subclass k of class j
427 over the total number of samples in class j .

428 3.2.3. Validation settings

429 We divide each dataset into a validation set, a training set and a test set
430 with no overlapping. The validation set is used to train a logistic regression
431 model or a Naive Bayes model, in order to generate assessments for the train-
432 ing set and the test set, which are used to train and evaluate the stochastic
433 supervision models, respectively.

434 In the MNIST dataset, we randomly select 2500 samples from each class
435 to generate the validation set. The training set is generated by randomly
436 selecting 2500 samples from the rest of each class. The rest samples form
437 the test set. For each experiment, we use all the training samples to train
438 the model; 20 tests are performed to evaluate the model, with each test

439 containing 1000 samples randomly selected from the test set; and thus 20
440 classification accuracies are recorded for the tests.

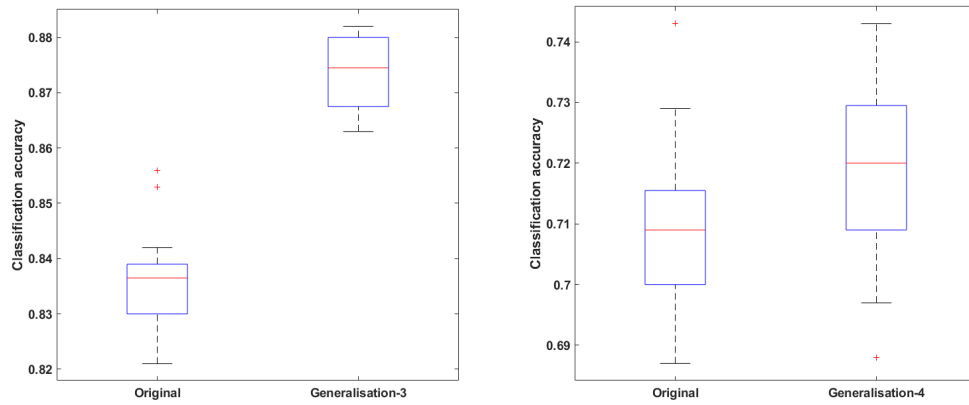
441 In the CIFAR-10 dataset, we use the training-test split provided by
442 Krizhevsky and Hinton [31], where the training set contains 50000 images
443 with 30000 for the animal class and 20000 for the transportation class. In or-
444 der to construct the validation set, we further divide the 50000 images in the
445 training set into two datasets: a validation set of 25000 images and a train-
446 ing set of 25000 images. The test set contains 10000 images with 6000 for
447 the animal class and 4000 for the transportation class. For each experiment,
448 we use all the training images to train the model and randomly select 1000
449 images from the test set to evaluate the model. We repeat the procedure 20
450 times and record 20 classification accuracies. All images are transformed to
451 greyscale in the experiments.

452 In the EMNIST dataset, we divide the 3000 images from each subclass to
453 a training set with 1200 images, a validation set with 1200 images and a test
454 set with 600 images. For each experiment, we use all 1200×47 training images
455 to train the model and randomly select 1000 images from the whole test set
456 with 600×47 images to test. We repeat the procedure 20 times and record
457 20 classification accuracies. The pixel values of the margin part of images in
458 EMNIST are zeros, which lead to singular covariance matrices. Thus we add
459 small white noises to these images to make the covariance matrices invertible.
460 Since Titterington’s model is used for binary classification and we have three
461 classes here, the one-versus-all strategy [33] is applied here for Titterington’s
462 model.

463 3.3. Results

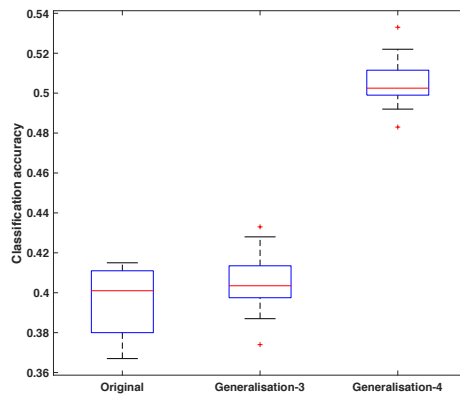
464 Classification accuracies on the 20 test sets of MNIST, CIFAR-10 and EM-
465 NIST are boxplotted in Figure 6(a), Figure 6(b) and Figure 6(c), respectively.
466 It is clear that the generalisation-3 and the generalisation-4 have higher boxes
467 than Titterington’s model in Figure 6(a) and Figure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

472 For the EMNIST data, the generalisation-3 and generalisation-4 produce
473 higher boxes than Titterington’s model and the generalisation-4 has the best
474 classification performance. This also shows the effectiveness of our models.
475 Note that here the generalisation-4 has much better classification perfor-
476 mance than the generalisation-3. One possible reason is that the multi-modal
477 classes have more effect on the final results than the feature-dependent as-
478 sessment, since the subclasses in each large class are clearly defined while
479 the linear relationship between the assessment and features is not strong, as
480 shown in Table 1. We also note that there is a large space for improvement
481 in classification accuracy of EMNIST. By developing a new method that can
482 deal with feature-dependent assessments and multi-modal classes together,
483 we may further improve the classification performance on complex data such
484 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.

485 4. Conclusions

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

503 Acknowledgements

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

507 **References**

- 508 [1] G. J. McLachlan, Iterative reclassification procedure for constructing
509 an asymptotically optimal rule of allocation in discriminant analysis,
510 Journal of the American Statistical Association 70 (350) (1975) 365–
511 369.
- 512 [2] T. J. O’neill, Normal discrimination with unclassified observations, Jour-
513 nal of the American Statistical Association 73 (364) (1978) 821–826.
- 514 [3] F. Schwenker, E. Trentin, Pattern classification and clustering: A review
515 of partially supervised learning approaches, Pattern Recognition Letters
516 37 (2014) 4–14.
- 517 [4] F. Schwenker, E. Trentin, Partially supervised learning for pattern recog-
518 nition, Pattern Recognition Letters 37 (2014) 1–3.
- 519 [5] D. Ahfock, G. J. McLachlan, On missing label patterns in semi-
520 supervised learning, arXiv preprint arXiv:1904.02883.
- 521 [6] X. Zhu, A. B. Goldberg, Introduction to Semi-Supervised Learning,
522 Morgan and Claypool Publishers, 2009.
- 523 [7] O. Chapelle, B. Schlkopf, A. Zien, Semi-Supervised Learning, The MIT
524 Press, 2010.
- 525 [8] C. Chittineni, Learning with imperfectly labeled patterns, Pattern
526 Recognition 12 (5) (1980) 281–291.
- 527 [9] T. Krishnan, Efficiency of learning with imperfect supervision, Pattern
528 Recognition 21 (2) (1988) 183–188.

- 529 [10] U. Katre, T. Krishnan, Pattern recognition with an imperfect supervisor,
530 Pattern recognition 22 (4) (1989) 423–431.
- 531 [11] B. Fréney, M. Verleysen, Classification in the presence of label noise:
532 a survey, IEEE transactions on neural networks and learning systems
533 25 (5) (2014) 845–869.
- 534 [12] C. Bouveyron, S. Girard, Robust supervised classification with mixture
535 models: Learning from data with uncertain labels, Pattern Recognition
536 42 (11) (2009) 2649–2658.
- 537 [13] M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple
538 instance learning: A survey of problem characteristics and applications,
539 Pattern Recognition 77 (2018) 329–353.
- 540 [14] J. Aitchison, C. B. Begg, Statistical diagnosis when basic cases are not
541 classified with certainty, Biometrika 63 (1) (1976) 1–12.
- 542 [15] T. Krishnan, S. C. Nandy, Discriminant analysis with a stochastic su-
543 pervisor, Pattern Recognition 20 (4) (1987) 379–384.
- 544 [16] D. M. Titterington, An alternative stochastic supervisor in discriminant
545 analysis, Pattern Recognition 22 (1) (1989) 91–95.
- 546 [17] T. Krishnan, S. C. Nandy, Efficiency of discriminant analysis when
547 initial samples are classified stochastically, Pattern Recognition 23 (5)
548 (1990) 529–537.
- 549 [18] T. Krishnan, S. C. Nandy, Efficiency of logistic-normal stochastic super-
550 vision, Pattern Recognition 23 (11) (1990) 1275–1279.

- 551 [19] D. M. Titterton, Some recent research in the analysis of mixture
552 distributions, *Statistics* 21 (4) (1990) 619–641.
- 553 [20] J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman
554 & Hall, 1986.
- 555 [21] G. McLachlan, T. Krishnan, *The EM algorithm and Extensions*, John
556 Wiley & Sons, 2007.
- 557 [22] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-
558 Verlag, 2006.
- 559 [23] H. Theil, *Economic forecasts and policy*, North-Holland Pub. Co., 1961.
- 560 [24] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures,
561 *Journal of the Royal Statistical Society. Series B (Methodological)* (1996)
562 155–176.
- 563 [25] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis,
564 and density estimation, *Journal of the American Statistical Association*
565 97 (458) (2002) 611–631.
- 566 [26] G. McLachlan, D. Peel, *Finite Mixture Models*, John Wiley & Sons,
567 2004.
- 568 [27] G. J. McLachlan, S. X. Lee, S. I. Rathnayake, Finite mixture models,
569 *Annual Review of Statistics and Its Application* 6 (2019) 355–378.
- 570 [28] R. P. Browne, P. D. McNicholas, M. D. Sparling, Model-based learning
571 using a mixture of mixtures of gaussian and uniform distributions, *IEEE*

- 572 Transactions on Pattern Analysis and Machine Intelligence 34 (4) (2011)
573 814–817.
- 574 [29] C. Viroli, G. J. McLachlan, Deep gaussian mixture models, *Statistics*
575 *and Computing* 29 (1) (2019) 43–51.
- 576 [30] Y. LeCun, C. Cortes, [MNIST handwritten digit database](http://yann.lecun.com/exdb/mnist/).
577 URL <http://yann.lecun.com/exdb/mnist/>
- 578 [31] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny
579 images, Tech. rep., Citeseer (2009).
- 580 [32] G. Cohen, S. Afshar, J. Tapson, A. Van Schaik, EMNIST: Extending
581 MNIST to handwritten letters, in: 2017 International Joint Conference
582 on Neural Networks (IJCNN), IEEE, 2017, pp. 2921–2926.
- 583 [33] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to sta-
584 tistical learning, Vol. 112, Springer, 2013.