Prior distribution choices for a mixture of experts *

A. V. Grabovoy[1], **V. V. Strijov**[2]

Abstract: The paper investigates a mixture of expert models. The mixture of experts is a set of experts and the gate function which weighs these experts. Each expert is a linear model. The gate function is a neural network with softmax on the last layer. The paper analyzes different prior distributions for each expert. The authors propose a method that takes into account the relationship between the prior distributions of different experts. The paper uses the EM algorithm for solving the optimisation problem. This paper proposes to use the mixture of experts for the problem of circles parameters estimation. Each expert fits one circle in the image. The experiment uses synthetic and real data to test the proposed method. The real data is a human eye image from the iris detection problem.

Keywords: mixture of Experts; bayesian model selection; prior distribution.

1 Introduction

The paper studies the problem of a mixture of experts constructing. A mixture of experts is multimodel, which are weighed local models that approximate the dataset. The weighting coefficients depend on datum from the dataset.

Examples of multimodel are bagging, gradient boosting [1] and random forest [2]. There are approaches to multimodeling [3] suggests that the contribution of each expert to the answer depends on the object from the dataset. A mixture of experts uses a gate function that weights the prediction of each expert.

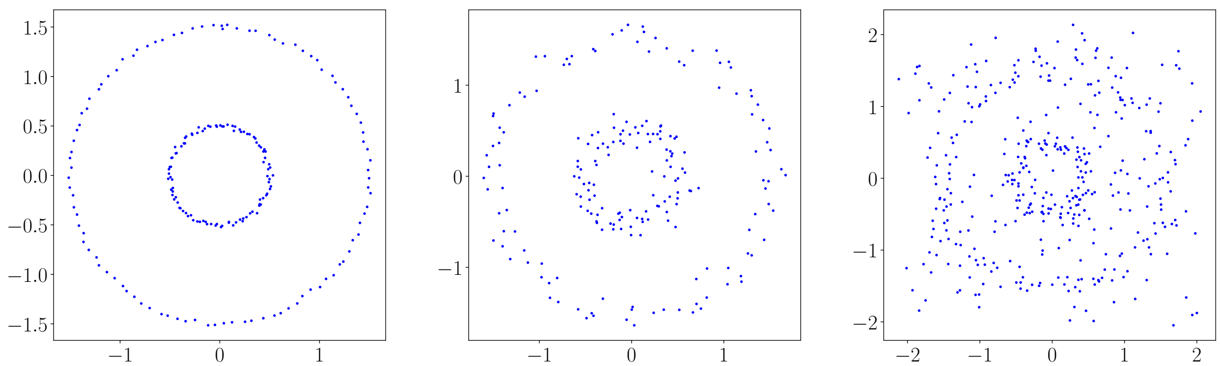
The main problem of multimodal is a convergence dependence on the initial point. We are using the probability approach for finding optimal mixture parameters and local expert

*This research was supported by RFBR (project ???) and NTI (project ???).

[1]Moscow Institute of Physics and Technology, grabovoy.av@phystech.edu

[2]Moscow Institute of Physics and Technology, Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, strijov@phystech.edu

parameters. The paper proposes to use different prior distribution for the parameters to improve convergence. The paper introduces the method which is using a dependence between prior distribution to improve multimodel quality.



(a) Without noise

(b) Noise near circle

(c) Uniform noise

Figure 1: An example of circles with different noise levels: (a) circle without noise; (b) noisy radius circle; (c) noisy radius circle, and uniform noisy on image

The paper analysis multimodel quality depending on the prior distribution for the expert's parameters. There solves a problem of finding circles on a binarized image. Examples of images are shown in fig. 1. In this paper, each expert is a linear model. A gate function is a two-layer fully connected neural network.

Related work. Many papers on a mixture of experts are devoted to the choice of a gateway function: softmax, the Dirichlet process [4], a neural network [5] with softmax function on the last layer. Some papers are devoted to the choice of expert type. Papers [6, 7] analyze linear models as experts. Papers [8, 9] analyze SVMs as experts. The paper [3] contains an overview of different methods for choosing a gate function and expert type.

A mixture of experts has many applications. Papers [10, 11, 12] use a mixture of experts in the task of forecasting time series. The paper [13] uses a mixture of experts in the task of recognizing handwritten numbers. Papers [14, 15, 16, 17] are devoted to methods of text and speech recognition by using a mixture of experts. The paper [18] analyzes a mixture of experts in the task of recognizing three-dimensional human movements.

The paper [19] is devoted to a review of the study results on the iris detection in the image. The methods of highlighting the borders of the iris and pupil are described in papers [20, 21].

2 Problem statement of circle parameters estimation

This data are binary image

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$

where 1 is a black pixel, an image, and 0 is a white pixel, the image background. An image example is shown in fig. 1. The image $\mathbf{M}$ is mapped to a set of coordinates $\mathbf{C} = \{x_i, y_i\}_{i=1}^{N}$. The pair of coordinates (x_i, y_i) is a black pixel in $\mathbf{M}$:

$$\mathbf{C} \in \mathbb{R}^{N \times 2},$$

where N is the number of black pixels.

Let (x_0, y_0) be the center of the circle, and r is radius of the circle. The coordinates $(x_i, y_i) \in \mathbf{C}$ is a circle locus of points defined by

$$\big(x_i - x_0\big)^2 + \big(y_i - y_0\big)^2 = r^2.$$

Expand brackets:

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + \big(r^2 - x_0^2 - y_0^2\big) \cdot 1 = x_i^2 + y_i^2. \tag{2.1}$$

Rewrite equation (2.1) to set the linear regressions problem for all points in the dataset:

$$\mathbf{X}\mathbf{w} \approx \mathbf{y}, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = \big[x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2\big]^{\mathsf{T}}. \tag{2.2}$$

The parameters $\mathbf{w} = [w_1, w_2, w_3]^{\mathsf{T}}$ reconstruct the circle parameters x_0, y_0, r:

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

The solution of problem (2.2) reconstructs the circle parameters only if the number of circles in an image is equal to one. The authors propose to use the multimodel for the image, which consists of several circles. The multimodel is an ensemble of the linear models. Each linear model approximates only one circle in the image. In this paper, multimodel is a mixture of experts.

3 Problem statement of building a mixture of experts

Generalize one-circle approximation problem to the case of several circles. Each circle is a local model. The data for this case is

$$\mathbf{X} \in \mathbb{R}^{N \times n}, \quad \mathbf{y} \in \mathbb{R}^{N} \tag{3.1}$$

where N is the sample size and n is the number of features. In this paper, n is equal to 3.

Definition 3.1. *A model* $\mathbf{f}$ *is a local model on dataset* $\boldsymbol{X}$ *if* $\mathbf{f}$ *approximates some no-empty subset* $\boldsymbol{X}' \subset \boldsymbol{X}$.

Definition 3.2. *Call the multimodel* $\hat{\mathbf{f}}$ *a mixture of experts*

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k\big(\mathbf{x}, \mathbf{V}\big) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k\big(\mathbf{x}, \mathbf{V}\big) = 1, \tag{3.2}$$

where $\mathbf{f}_k$ *is a local model,* π_k *is a gate function, vector* $\mathbf{w}_k$ *is some parameters of the local model and* $\mathbf{V}$ *is some parameters of the gate function.*

This paper asserts the local model be linear model. The gate function is the two–layer fully connected neural network

$$\mathbf{f}_k(\mathbf{x}) = \mathbf{w}_k^\mathsf{T}\mathbf{x}, \quad \boldsymbol{\pi}(\mathbf{x}, \mathbf{V}) = \mathrm{softmax}\left(\mathbf{V}_1^\mathsf{T}\boldsymbol{\sigma}\left(\mathbf{V}_2^\mathsf{T}\mathbf{x}\right)\right), \tag{3.3}$$

where $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2\}$ is a set of the gate function parameters.

The paper proposes to use a probabilistic approach to describe a mixture of experts. Let $\mathbf{y}$ be a random variable with density function $p(\mathbf{y}|\mathbf{X})$. Let density $p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{f}})$ approximate truth density $p(\mathbf{y}|\mathbf{X})$:

$$p(\mathbf{y}|\mathbf{X}, \hat{\mathbf{f}}) = \prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k p_k\left(y_i|\mathbf{f}_k(\mathbf{x}_i)\right)\right), \tag{3.4}$$

where $\hat{\mathbf{f}}$ is the mixture of experts and $\mathbf{f}_k, \boldsymbol{\pi}$ are defined by (3.3).

Suppose that $\mathbf{w}_k$ is the random variable with density function $p^k(\mathbf{w}_k)$, and get joint probability distribution of the target and the parameters:

$$p(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}) = \prod_{k=1}^{K} p^k(\mathbf{w}_k)\prod_{i=1}^{N}\left(\sum_{k=1}^{K}\pi_k p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right), \tag{3.5}$$

where $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\}$. The optimal parameters is delivered by the evidence maximisation:

$$\hat{\mathbf{V}} = \arg\max_{\mathbf{V}} p(\mathbf{y}|\mathbf{X}, \mathbf{V}).$$

4 Probabilistic statement of mixture of expert

To build the mixture of experts (3.2, 3.5), set the following probabilistic statement for the dataset (3.1):

1) the likelihood $p_k(y_i|\mathbf{w}_k, \mathbf{x}_i) = \mathcal{N}(y_i|\mathbf{w}_k^\mathsf{T}\mathbf{x}_i, \beta^{-1})$, where parameter β is the noise level,

2) the prior distribution for the parameters $p^k(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k)$, where $\mathbf{w}_k^0$ is a vector of size $n \times 1$ and $\mathbf{A}_k$ is a covariance matrix,

3) the prior regularisation $p(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\Xi}) = \mathcal{N}(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi})$, where $\boldsymbol{\Xi}$ is a covariance matrix and $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

The assumption 2) is the prior distribution for the parameters $\mathbf{w}_k$. It sets some restrictions to the model. For example, if $\mathbf{w}_k^0$ is [0,0,1], then the k-th local model fits with the most probability a circle with parameters $x_0 = 0, y_0 = 0, r = 1$.

The assumption 3) is a prior regularisation. It sets restrictions on the prior distribution parameters. For example, if $\mathrm{diag}(\boldsymbol{\Xi}) = [0.001, 0.001, 1]$, then the centers of different circles, which corresponds to different local models are equal.

Combining assumptions 1), 2), 3) and equation (3.5), obtain the new likelihood:

$$
\begin{aligned}
p\big(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = & \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\big(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\big) \right) \cdot \\
& \cdot \prod_{k=1}^{K} \mathcal{N}\big(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\big) \cdot \prod_{k,k'=1}^{K} \mathcal{N}\big(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\big),
\end{aligned}
\tag{4.1}
$$

where $\mathbf{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_K\}$.

Introduce a binary matrix $\mathbf{Z}$. Its element z_{ik} is equal to 1 if and only if the i-th object corresponds to the k-th local model. Using the binary matrix $\mathbf{Z}$ in (4.1) take logarithm and obtain

$$
\begin{aligned}
\log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) = & \\
= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} & \left[\log \pi_k\big(\mathbf{x}_i, \mathbf{V}\big) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
+ \sum_{k=1}^{K} & \left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} \mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
+ \sum_{k=1}^{K}\sum_{k'=1}^{K} & \left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}} \boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned}
\tag{4.2}
$$

Set the new optimisation problem to optimise the evidence function. The function is obtained by integrating equation (4.2) over parameters $\mathbf{W}, \mathbf{Z}$:

$$
\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \int_{\mathbf{W}, \mathbf{Z}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\big) d\mathbf{W} d\mathbf{Z}. \tag{4.3}
$$

5 EM–algorithm as a solver of optimisation problem

Let $q(\mathbf{W}, \mathbf{Z})$ be some density distribution of parameters $\mathbf{W}, \mathbf{Z}$. Rewrite the evidence:

$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) &= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) d\mathbf{W} d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)}{p(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)} d\mathbf{W} d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) q(\mathbf{W}, \mathbf{Z})}{p(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) q(\mathbf{W}, \mathbf{Z})} d\mathbf{W} d\mathbf{Z} = \\
&= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)}{q(\mathbf{W}, \mathbf{Z})} d\mathbf{W} d\mathbf{Z} + \\
&+ \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \frac{q(\mathbf{W}, \mathbf{Z})}{p(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)} d\mathbf{W} d\mathbf{Z} = \\
&= \mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) + \mathsf{D}_{KL}\left(q(\mathbf{W}, \mathbf{Z}) || p(\mathbf{W}, \mathbf{Z}|\mathbf{y}, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)\right)
\end{aligned}
\tag{5.1}
$$

Using (5.1), we get lower bound of evidence:

$$
\log p(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) \geq \mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta),
$$

where $\mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$ is called lower bound of evidence.

Use the expectation-maximization [22, 23] algorithm to find the solution to optimisation problem (4.3). The EM-algorithm instead of $\log p(\mathbf{y}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)$ optimizes the lower bound $\mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$.

E-step. E-step solves the optimisation problem:

$$
\mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) \to \max_{q(\mathbf{W},\mathbf{Z})},
$$

where $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$ are fixed.

Let the joint distribution $q(\mathbf{Z}, \mathbf{W})$ satisfies the assumption of independence $q(\mathbf{Z}, \mathbf{W}) = q(\mathbf{Z})q(\mathbf{W})$ [23]. The symbol $\propto$ means that both sides are equal to up to an additive constant. First, find the distribution $q(\mathbf{Z})$:

$$
\begin{aligned}
\log q(\mathbf{Z}) = \mathsf{E}_{q/\mathbf{Z}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) \propto \\
\propto \sum_{i+1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2}\left(y_i^2 - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i\right) + \frac{1}{2} \log \frac{\beta}{2\pi}\right] \\
p(z_{ik} = 1) = \frac{\exp\left(\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K} \exp\left(\log \pi_{k'}(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_{k'} \mathbf{w}_{k'}^{\mathsf{T}} \mathbf{x}_i - \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_{k'}\right)\right)}.
\end{aligned}
\tag{5.2}
$$

Using (5.2), we get that distribution $q(z_{ik})$ is the Bernoulli distribution with probability z_{ik} from equation (5.2). Second, find the distribution $q(\mathbf{W})$:

$$\begin{aligned}
\log q(\mathbf{W}) &= \mathsf{E}_{q/\mathbf{W}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) \propto \\
&\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E}z_{ik} \left[\log \pi_k(\mathbf{x}_{i,\mathbf{V}}) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] \\
&\propto \sum_{k=1}^{K}\left[\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E}z_{ik}\right) - \frac{1}{2}\mathbf{w}_k^{\mathsf{T}}\left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)\mathbf{w}_k\right].
\end{aligned} \tag{5.3}$$

Using (5.3), we get that distribution $q(\mathbf{w}_k)$ is the normal distribution with mean $\mathbf{m}_k$ and covariance matrix $\mathbf{B}_k$:

$$\mathbf{m}_k = \mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathsf{E}z_{ik}\right), \qquad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1}.$$

M-step. M-step solves the optimisation problem:

$$\mathcal{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) \to \max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta},$$

where $q(\mathbf{W}, \mathbf{Z})$ are fixed density function. The distribution $q(\mathbf{Z}, \mathbf{W})$ is fixed, while the lower bound $\mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$ is maximized with respect to the parameters $\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta$:

$$\begin{aligned}
\mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) &= \mathsf{E}_q \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta) = \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\mathsf{E}z_{ik}\left[\log\pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2}\mathsf{E}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\
&+ \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\boldsymbol{\Xi}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].
\end{aligned} \tag{5.4}$$

First, to find the optimal parameters $\mathbf{V}$, use the gradient optimisation algorithm. It convergences to some local maximum. Second, using (5.4), we get optimal value for $\mathbf{A}_k$

$$\begin{aligned}
\frac{\partial\mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial\mathbf{A}_k^{-1}} &= \frac{1}{2}\mathbf{A}_k - \frac{1}{2}\mathsf{E}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}} = 0, \\
\mathbf{A}_k &= \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}} - \mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}} - \mathsf{E}\mathbf{w}_k\mathbf{w}_k^{0\mathsf{T}} + \mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}}.
\end{aligned}$$

Similarly, we get optimal value for β and for $\mathbf{w}_k^0$

$$\frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \beta} = \sum_{k=1}^{K}\sum_{i=1}^{N}\left(\frac{1}{\beta}\mathsf{E}z_{ik} - \frac{1}{2}\mathsf{E}z_{ik}\left[y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\right) = 0,$$

$$\frac{1}{\beta} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik}.$$

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{w}_k^0} &= \mathbf{A}_k^{-1}\left(\mathsf{E}\mathbf{w}_k - \mathbf{w}_k^0\right) + \mathbf{\Xi}\sum_{k'=1}^{K}\left[\mathbf{w}_{k'}^0 - \mathbf{w}_k^0\right] = 0,\\ \mathbf{w}_k^0 &= \left[\mathbf{A}_k^{-1} + (K-1)\,\mathbf{\Xi}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k + \mathbf{\Xi}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right).\end{aligned} \tag{5.5}$$

Formulas (5.2–5.5) define an iterative procedure, which convergence to some local maximum of the optimisation problem (4.3).

6 Computational experiment

The computational experiment analyzes quality of various multimodels for circle approximation. The experiment analyzes next multimodels: multimodel $\hat{\mathbf{f}}_1$ without prior distribution for the parameters, multimodel $\hat{\mathbf{f}}_2$ with prior distribution (6.2) for the parameters and multimodel $\hat{\mathbf{f}}_3$ with prior regularisation. The approximation quality of model $\hat{\mathbf{f}}_i$ is

$$\mathcal{S}_{\hat{\mathbf{f}}_i} = \sum_{k=1}^{K}\left(x_0^k - x_{\text{pr}}^k\right)^2 + \left(y_0^k - y_{\text{pr}}^k\right)^2 + \left(r^k - r_{\text{pr}}^k\right)^2, \tag{6.1}$$

where x_0^k, y_0^k, r^k are the true center and radius for k-th circle, $x_{\text{pr}}^k, y_{\text{pr}}^k, r_{\text{pr}}^k$ are the predicted center and radius for k-th circle.

To compare models with various prior distribution, use the log-likelihood without any priors (3.4). The prior distribution for the parameters in the experiment is

$$p^1(\mathbf{w}_1) \sim \mathcal{N}(\mathbf{w}_1^0, \mathbf{I}), \quad p^2(\mathbf{w}_2) \sim \mathcal{N}(\mathbf{w}_2^0, \mathbf{I}), \tag{6.2}$$

where $\mathbf{w}_1^0 = [0, 0, 0.1]$, $\mathbf{w}_2^0 = [0, 0, 2]$.

Synthetic data with various types of noise in the image. The computational experiment compares quality of mixture of experts $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ on a synthetic dataset.

The synthetic data were generated as two concentric circles with different noise levels. The Synthetic 1 is the image without any noises, the Synthetic 2 is the image with a noise radius and the Synthetic 3 is an image with uniform noise.

The fig. 2 shows results of multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$. All multimodels run 50 iterations of the EM-algorithm. Multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ approximate circles better than multimodel $\hat{\mathbf{f}}_1$. Tab. 1 shows the approximation quality (6.1) for all multimodels.

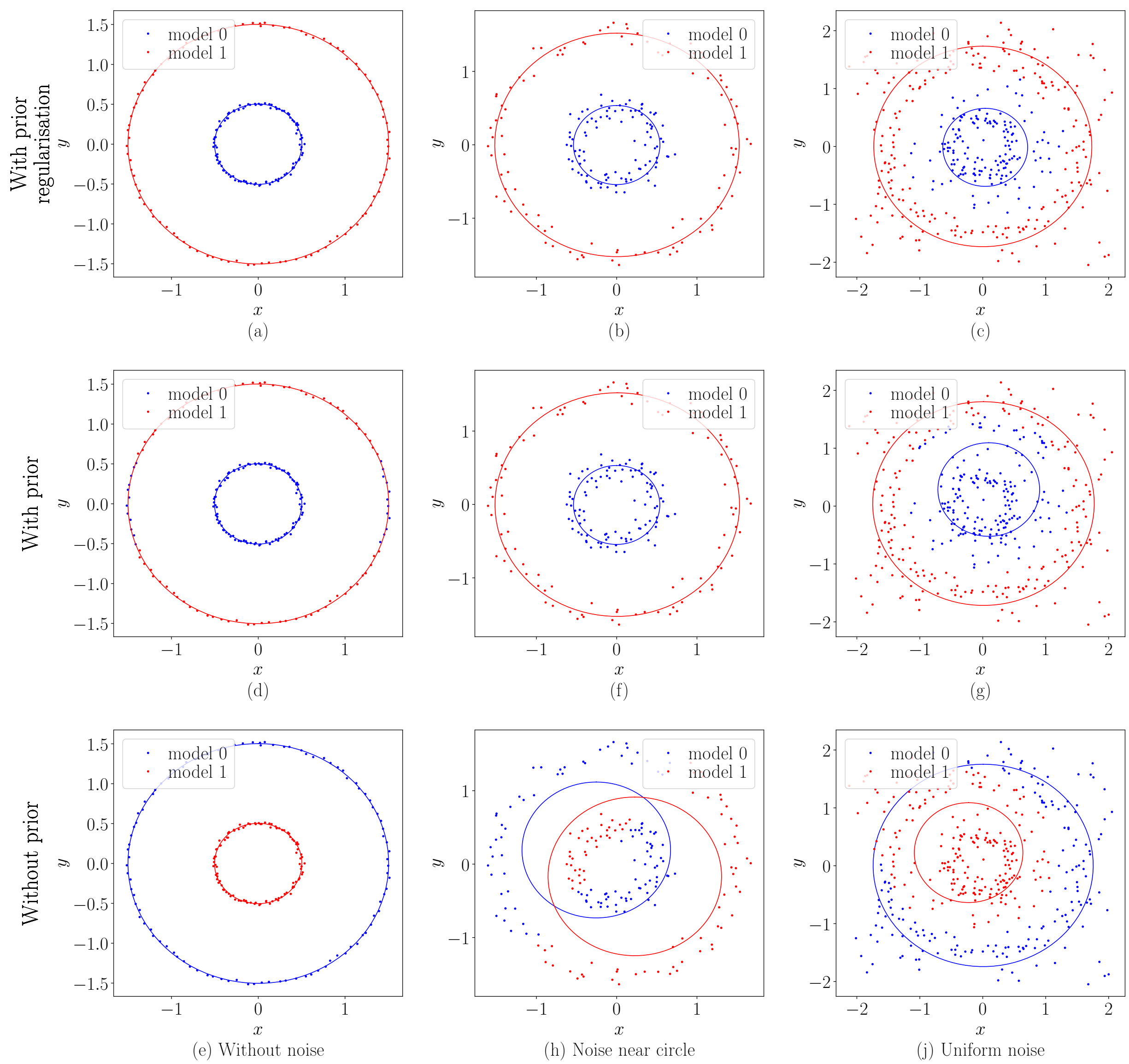


Figure 2: The multimodel dependes on different prior distribution and dependes on different noise levels: (a)–(c) multimodel with prior regularization; (d)–(g) multimodel with simple prior; (e)–(j) multimodel without any priors.

Table 1: The approximation quality (6.1) for all multimodels

Dataset	$\mathcal{S}_{\hat{\mathbf{f}}_1}$	$\mathcal{S}_{\hat{\mathbf{f}}_2}$	$\mathcal{S}_{\hat{\mathbf{f}}_3}$
Synthetic 1	10^{-5}	10^{-5}	10^{-5}
Synthetic 2	0.6	10^{-3}	10^{-3}
Synthetic 3	0.6	10^{-3}	10^{-3}

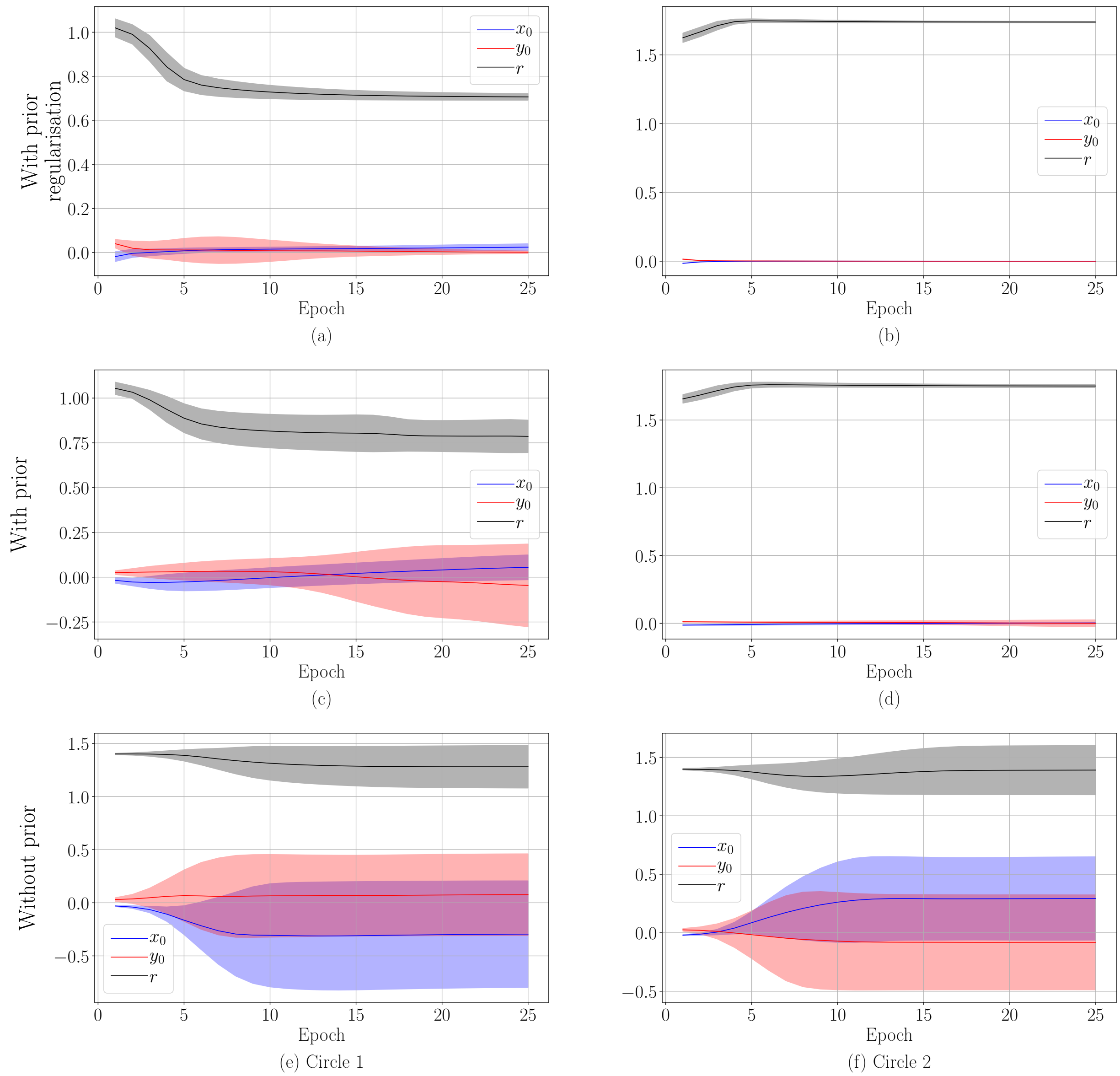


Figure 3: The dependence of center and radius on the iteration number: (a)–(b) multimodel with prior regularization; (e)–(d) multimodel with simple prior; (e)–(f) multimodel without any priors.

Analysis of convergence on synthetic data. The experiment analyzes multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ during the EM-algorithm convergence. Multimodels were analysed on the synthetic dataset Synthetic 3.

The fig. 3 shows the dependence of the radius and center on the EM-algorithm iteration number. The multimodel with prior distribution for the parameters $\hat{\mathbf{f}}_2$ approximates circles better than multimodel $\hat{\mathbf{f}}_1$ without any priors. The multimodel with prior regularisation $\hat{\mathbf{f}}_3$ approximates circles more stable than multimodel $\hat{\mathbf{f}}_2$.

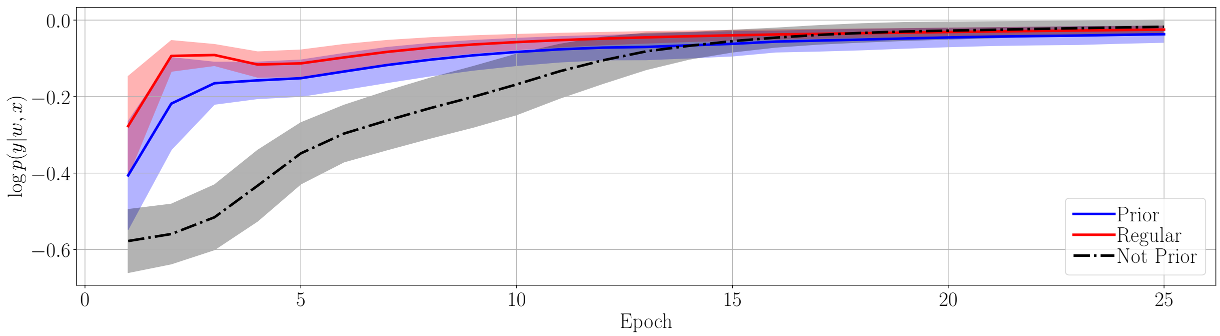


Figure 4: The dependence of log-likelihood (3.4) on the iteration number.

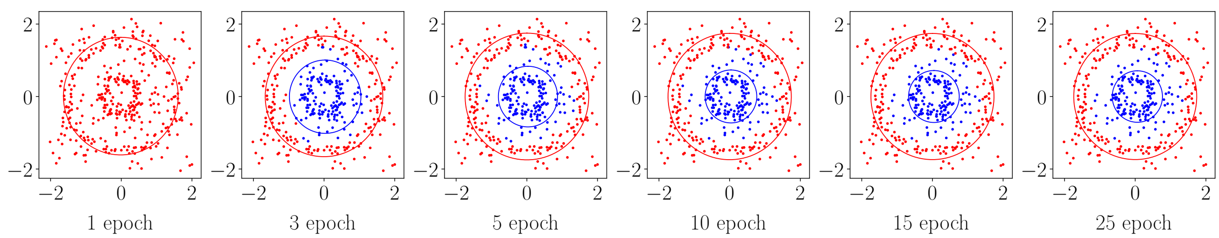


Figure 5: Visualization of convergence for the multimodel with prior regularisation.

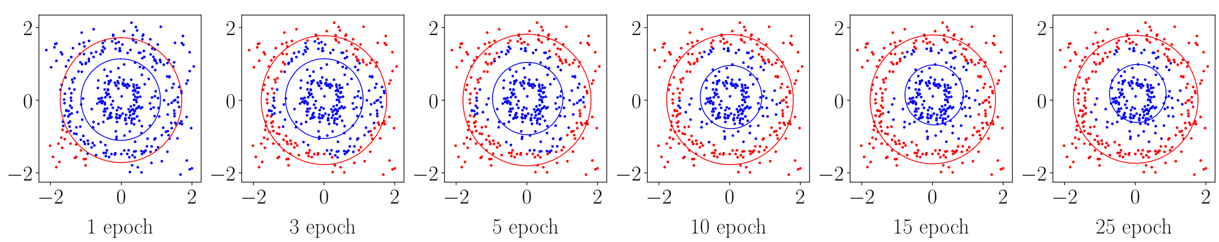


Figure 6: Visualization of convergence for the multimodel with simple prior.

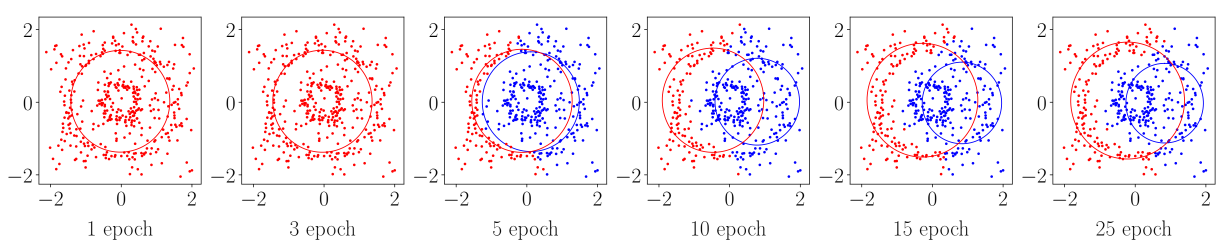
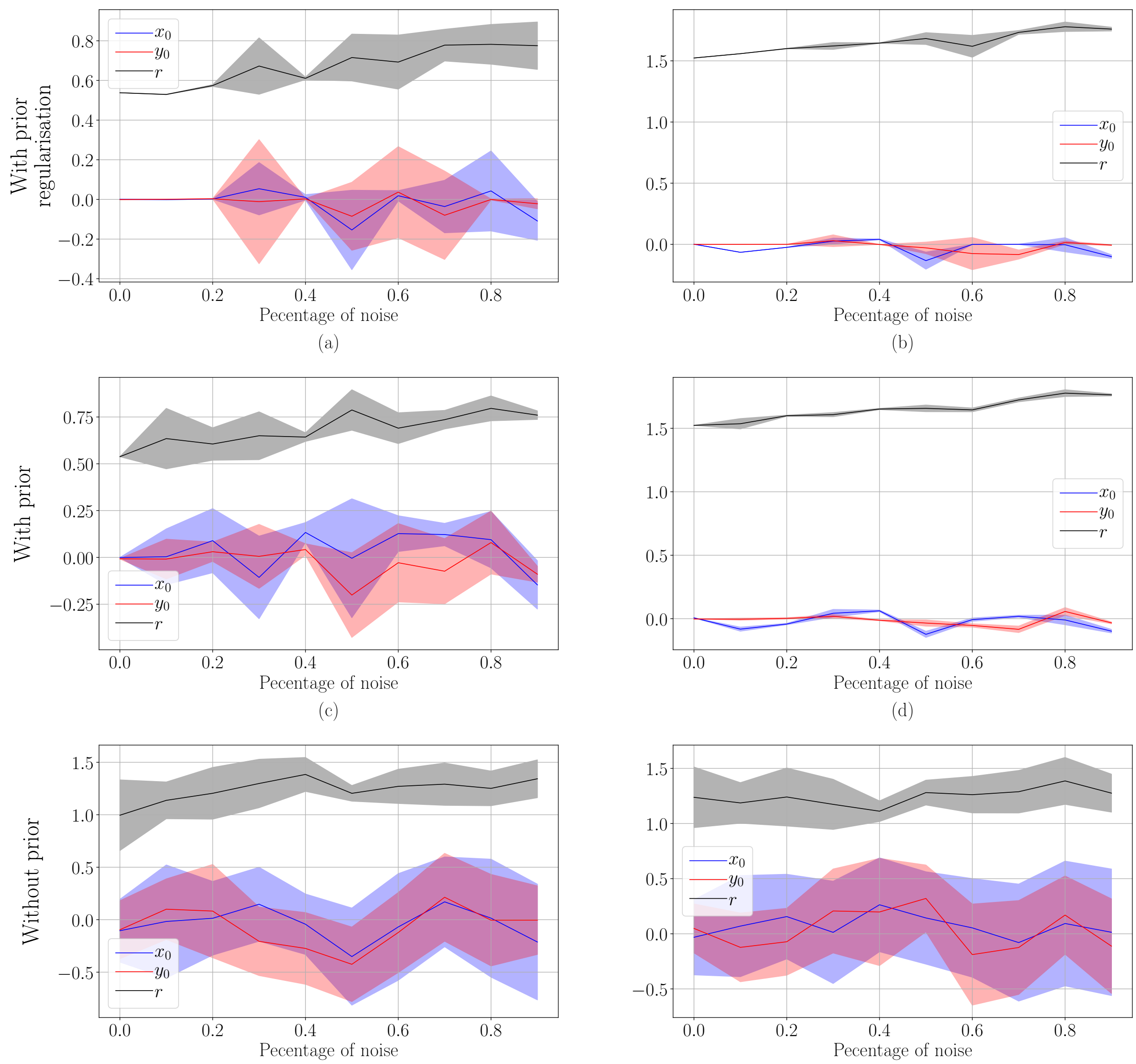


Figure 7: Visualization of convergence for the multimodel without any priors.

The fig. 4 shows the dependence of log-likelihood (3.4) on the EM-algorithm iteration number. The log-likelihood of multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ is growing faster than multimodel $\hat{\mathbf{f}}_1$. After the 20-th iteration all three multimodels have the same log-likelihood.

The fig. 5-7 show learning process for multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$. The fig. 7 shows multimodel $\hat{\mathbf{f}}_1$, which does not approximate circles correctly. The fig. 5-6 show multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$, which approximate circles correctly.

The experiment shows that multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ with prior distribution for the parameters approximate circles better than multimodel $\hat{\mathbf{f}}_1$ without any priors.



(a) (b)

(c) (d)

(e) Circle 1 (f) Circle 2

Figure 8: The dependence of center and radius on the noise level: (a)–(b) multimodel with prior regularization; (c)–(d) multimodel with simple prior; (e)–(f) multimodel without any priors.

Multimodels analysis depending on the noise level. The experiment analyzes dependence of multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ on the noise level. Multimodels were analysed on the

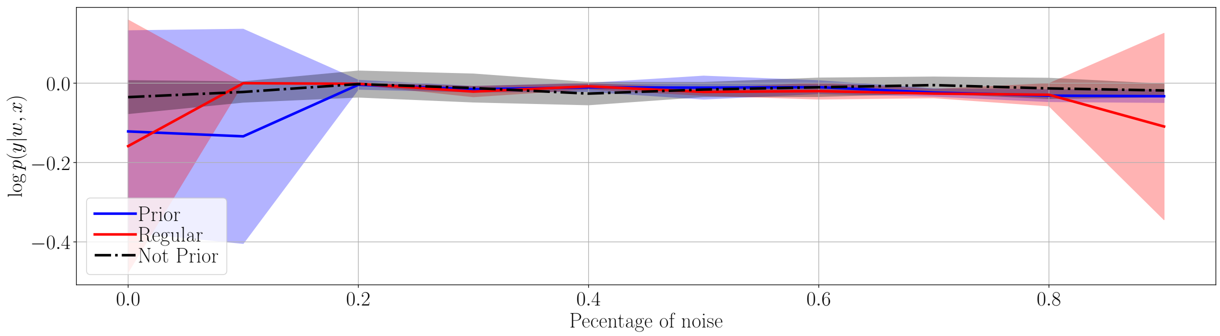


Figure 9: The dependence of log-likelihood (3.4) on the noise level.

synthetic dataset Synthetic 1, with adding various noise level. The lowest noise level is equal to 0, when no noise. The highest noise level is equal to 1, when number of noise datum is equal to number of point in circles.

The fig. 8 shows the dependence of center and radius on the noise level. It shows that circle radius increases with increasing noise level. The multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ approximate circles center correctly, but the multimodel $\hat{\mathbf{f}}_3$ is more stable.

The fig. 9 shows the dependence of log-likelihood (3.4) on the noise level. It shows that log-likelihood (3.4) is the same for all multimodels, but the fig. 8 shows that the approximation quality (6.1) depends on multimodels.

This part of the experiment shows that multimodel $\hat{\mathbf{f}}_3$ with prior regularisation is the most stable model.

Real data. This part of the experiment analyzes multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ on the real data.

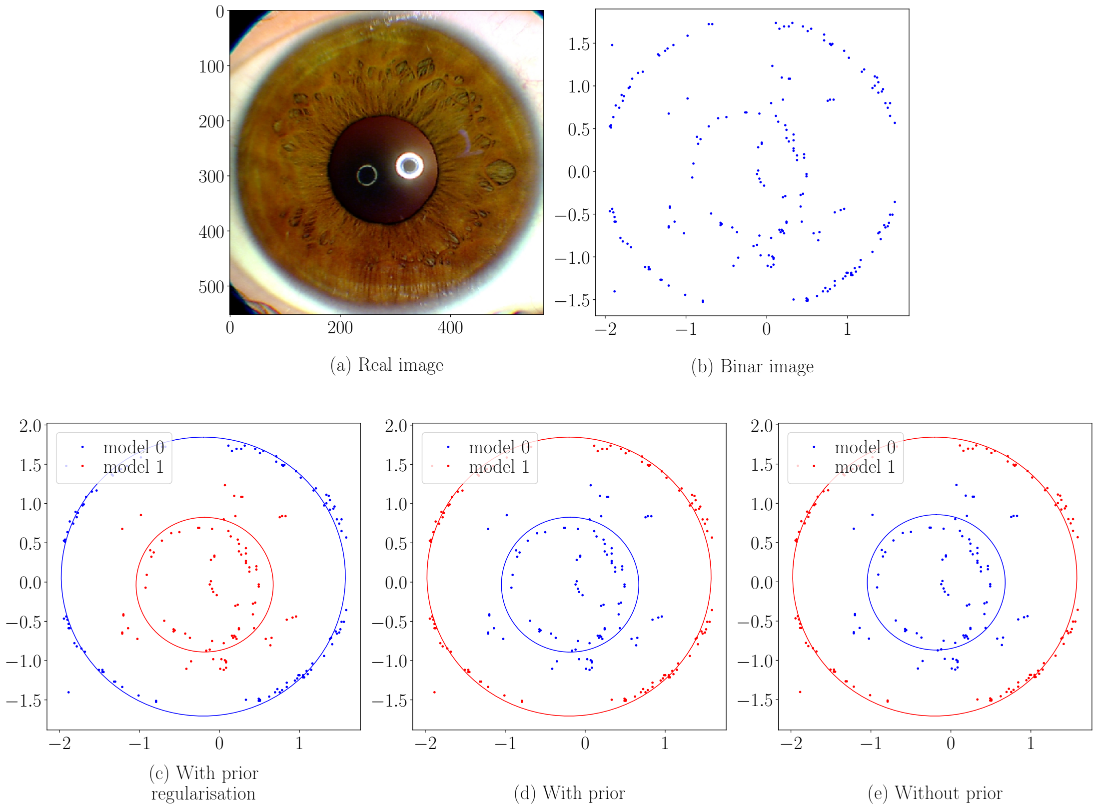
The fig. 10 shows the result of different multimodels. The multimodel $\hat{\mathbf{f}}_1$ approximates incorrectly one of the circles. The multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ approximate both circles correctly.

The fig. 11-13 show learning process for multimodels $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$. The fig. 11 shows multimodel $\hat{\mathbf{f}}_1$. The fig. 12 shows multimodel $\hat{\mathbf{f}}_2$. The fig. **??** shows multimodel $\hat{\mathbf{f}}_3$.

This part of experiment shows that multimodels $\hat{\mathbf{f}}_2, \hat{\mathbf{f}}_3$ approximate circles better than multimodel $\hat{\mathbf{f}}_1$ even for the real images.

7 Conclusion

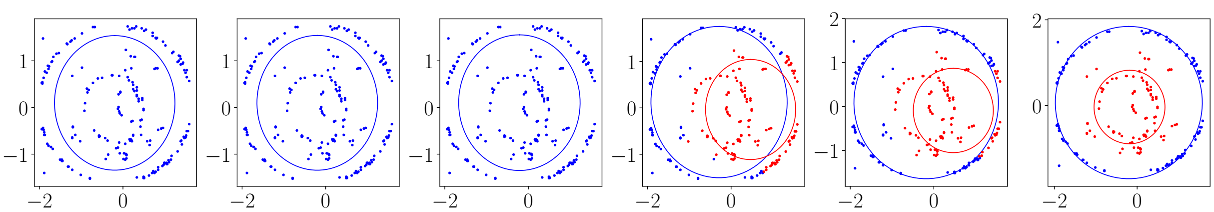
The paper compares multimodels with different prior distributions for the model parameters. Concentric circles with different noises are data in the computational experiment. The linear models were used to approximate the circles in the image. The gate function is a two-layer fully connected neural network. The experiment compares the model with the prior distribution and without it. The multimodel with prior distribution is more accurate compared to the model without prior distribution. Another experiment compares different types of regularization. The experiment showed that multimodel with regularization is



(a) Real image (b) Binar image

(c) With prior regularisation (d) With prior (e) Without prior

Figure 10: Different multimodels on the real image: (a) source image; (b) binarize image; (c) multimodel without any priors; (d) multimodel with simple prior; (e) multimodel with prior regularisation.



1 epoch 3 epoch 5 epoch 10 epoch 15 epoch 25 epoch

Figure 11: Visualization of convergence for the multimodel without any priors.

more stable. The experiment shows that all models in this article are sensitive to outburst. To solve this problem, it is proposed to use a local model that approximates noise.

The future work, plannes to improve the multimodel by adding a prior distribution for the gate function parameters. It plannes to add a local model that approximates noise in the data. It assumes that the probability of noise is low.

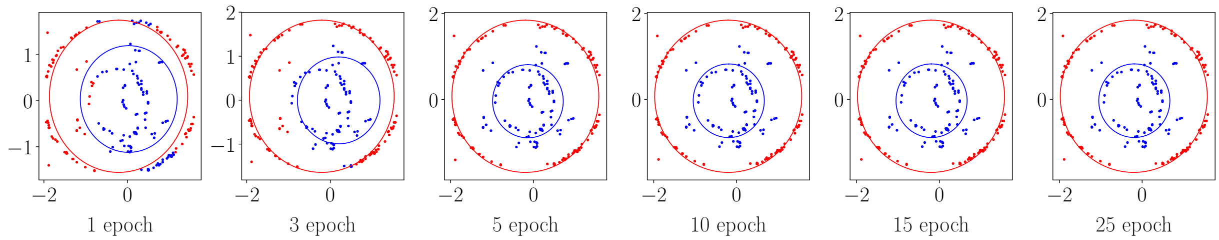


Figure 12: Visualization of convergence for the multimodel with simple prior.

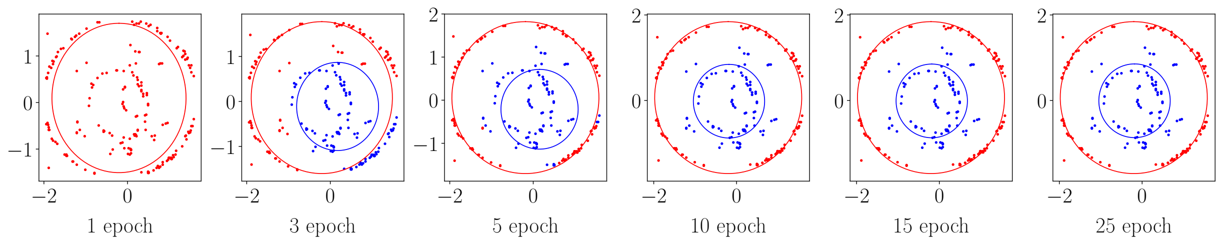


Figure 13: Visualization of convergence for the multimodel with prior regularisation.

References

[1] *Tianqi C., Carlos G.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[2] *Xi C., Hemant I.* Random Forests for Genomic Data Analysis // Genomics. 2012. Issues. 99. No. 6. P. 323–329.

[3] *Esen Y. S., Wilson J., Gader P. D.* Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. Issues. 23. No. 8. P. 1177–1193.

[4] *Rasmussen C. E., Ghahramani Z.* Infinite Mixtures of Gaussian Process Experts // Advances in Neural Information Processing Systems 14. 2002. P. 881–888.

[5] *Shazeer N., Mirhoseini A., Maziarz K.* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // International Conference on Learning Representations. 2017.

[6] *Jordan M. I.* Hierarchical mixtures of experts and the EM algorithm // Neural Comput. 1994. Vol. 6, No. 2. P. 181–214.

[7] *Jordan M. I., Jacobs R. A.* Hierarchies of adaptive experts // In Advances in Neural Information Processing Systems. 1991. P. 985–992.

[8] *Lima C., Coelho A., Zuben F. J.* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci. 2007. Vol. 177. No. 10. P. 2049–2074.

[9] *Cao L.* Support vector machines experts for time series forecasting // Neurocomputing. 2003. Vol. 51. P. 321–339.

[10] *Yumlu M. S., Gurgen F. S., Okay N.* Financial time series prediction using mixture of experts // In Proc. 18th Int. Symp. Comput. Inf. Sci. 2003. P. 553–560.

[11] *Cheung Y. M., Leung W. M., Xu L.* Application of mixture of experts model to financial time series forecasting // On Proc. Int. Conf. Neural Netw. Signal Process. 1995. P. 1–4.

[12] *Weigend A. S., Shi S.* Predicting daily probability distributions of S&P500 returns // J. Forecast. 2000. Vol. 19. No. 4. P. 375–392.

[13] *Ebrahimpour R., Moradian M. R., Esmkhani A., Jafarlou F. M.* Recognition of Persian handwritten digits using characterization loci and mixture of experts // J. Digital Content Technol. Appl. 2009. Vol. 3. No. 3. P. 42–46.

[14] *Estabrooks A., Japkowicz N.* A mixture-of-experts framework for text classification //In Proc. Workshop Comput. Natural Lang. Learn., Assoc. Comput. Linguist. 2001. P. 1–8.

[15] *Mossavat S., Amft O., Petkov Vries B., Kleijn W.* A Bayesian hierarchical mixture of experts approach to estimate speech quality // In Proc. 2nd Int. Workshop Qual. Multimedia Exper. 2010. P. 200–205.

[16] *Peng F., Jacobs R. A., Tanner M. A.* Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition // J. Amer. Stat. Assoc. 1996. Vol. 91. No. 435. P. 953–960.

[17] *Tuerk A.* The state based mixture of experts HMM with applications to the recognition of spontaneous speech. Ph.D. thesis. Cambridge: Univ. Cambridge, 2001.

[18] *Sminchisescu C., Kanaujia A., Metaxas D.* Discriminative density propagation for visual tracking // IEEE Trans. Pattern Anal. Mach. Intell. 2007. Vol. 29. No. 11. P. 2030–2044.

[19] *Bowyer K., Hollingsworth K., Flynn P.* A Survey of Iris Biometrics Research: 2008–2010.

[20] *Matveev I.* Detection of iris in image by interrelated maxima of brightness gradient projections // Appl.Comput. Math. 2010. Vol. 9. No. 2. P. 252–257.

[21] *Matveev I., Simonenko I.*. Detecting precise iris boundaries by circular shortest path method // Pattern Recognition and Image Analysis. 2014. Vol. 24. P. 304–309.

[22] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological). 1977. Vol. 39. No. 1 P. 1–38.

[23] *Bishop C.* Pattern Recognition and Machine Learning. Berlin: Springer, 2006. P. 758.