Direct Importance Estimation with a Mixture of Probabilistic Principal Component Analyzers

Makoto YAMADA††††, Masashi SUGIYAMA††††, Members, Gordon WICHERN††††, and Jaak SIMM†, Nonmembers

SUMMARY Estimating the ratio of two probability density functions (a.k.a. the importance) has recently gathered a great deal of attention since importance estimators can be used for solving various machine learning and data mining problems. In this paper, we propose a new importance estimation method using a mixture of probabilistic principal component analyzers. The proposed method is more flexible than existing approaches, and is expected to work well when the target importance function is correlated and rank-deficient. Through experiments, we illustrate the validity of the proposed approach.

1. Introduction

Recently, the ratio of two probability density functions (a.k.a. the importance function) has been shown to be useful for solving various data processing tasks [1], including non-stationarity adaptation [2], [3], transfer learning [4], multi-task learning [5], outlier detection [6], change detection [7], feature selection [8], feature extraction [9], independent component analysis [10], causal inference [11], conditional density estimation [12], and probabilistic classification [13]. For this reason, importance estimation has attracted considerable attention these days.

A naive approach to approximating the importance function is to separately estimate the densities corresponding to the numerator and denominator of the ratio, and then take the ratio of the estimated densities. However, density estimation itself is a difficult problem and taking the ratio of estimated densities can magnify the estimation error. To overcome this problem, a direct importance estimator called the Kullback-Leibler importance estimation procedure (KLIEP) was proposed [14]. A typical implementation of KLIEP employs a linear combination of spherical Gaussian kernels for modeling the importance function, and the (common) Gaussian width is chosen by cross-validation.

Although KLIEP was shown to work well in experiments, using ellipsoidal Gaussian kernels would be more suitable when the true importance function has a correlated profile. Following this idea, an extension of KLIEP using a Gaussian mixture model called Gaussian-mixture KLIEP (GM-KLIEP) was proposed [15]. In GM-KLIEP, the Gaussian mixture model is efficiently trained using an expectation-maximization algorithm, in which the covariance matrices of Gaussian kernels are adaptively adjusted. Although GM-KLIEP was shown to be promising in experiments, its training algorithm involves estimation of inverse covariance matrices, which makes GM-KLIEP numerically unstable when data samples are (locally) rank-deficient.

A standard approach to coping with rank deficiency would be to use dimensionality reduction methods such as principal component analysis (PCA). Following this idea, we propose to combine GM-KLIEP with PCA, resulting in a new direct importance estimation procedure using a mixture of probabilistic principal component analyzers (PPCAs) [16]. We illustrate the usefulness of the proposed approach called PPCA-mixture KLIEP (PM-KLIEP) through experiments.

2. Background

In this section, we formulate the problem of importance estimation, and briefly review the KLIEP method.

2.1 Problem Formulation

Let \( D \in \mathbb{R}^d \) be the data domain and suppose we are given i.i.d. samples \( \{x_i^{de}\}_{i=1}^{n_a} \) from a “denominator” data distribution with density \( p_{de}(x) \) and i.i.d. samples \( \{x_j^{nu}\}_{j=1}^{n_u} \) from a “numerator” data distribution with density \( p_{nu}(x) \). We assume that \( p_{de}(x) > 0 \) for all \( x \in D \). The goal of this paper is to develop a method of estimating the importance function \( w(x) \) from \( \{x_i^{de}\}_{i=1}^{n_a} \) and \( \{x_j^{nu}\}_{j=1}^{n_u} \):

\[
w(x) = \frac{p_{nu}(x)}{p_{de}(x)}.
\]

Our key restriction is that we avoid estimating densities \( p_{de}(x) \) and \( p_{nu}(x) \) when estimating \( w(x) \).

2.2 Kullback-Leibler Importance Estimation Procedure (KLIEP)

KLIEP allows one to directly estimate \( w(x) \) without going
through density estimation [14]. In KLIEP, the following linear importance model is used:

$$
\hat{w}(x) = \sum_{i=1}^{b} \alpha_i \varphi_i(x),
$$

(1)

where \(\{\alpha_i\}_{i=1}^{b}\) are parameters, \(b\) is the number of parameters, and \(\varphi_i(x)\) is a basis function. In the original KLIEP paper [14], the Gaussian kernel was used as the basis function:

$$
\varphi_i(x) = \exp \left( -\frac{||x - c_i||^2}{2\tau^2} \right),
$$

where \(\tau > 0\) is the Gaussian width and \(c_i\) is the Gaussian center randomly chosen from \([x_{ij}]_{j=1}^{m}\) of (approximately) the same size. Then obtain an importance estimator \(\hat{w}_k(x)\) from \([x_{ij}]_{j=1}^{m}\) and approximate the score \(KL'\) using \(X_{\alpha}^{nu}\) as

$$
\hat{KL}_k = \frac{1}{|X_{\alpha}^{nu}|} \sum_{x \in X_{\alpha}^{nu}} \ln \hat{w}_k(x).
$$

This procedure is repeated for \(k = 1, \ldots, K\) and the average of \(\hat{KL}_k\) over all \(k\) is used as an estimate of \(KL'\):

$$
\hat{KL}' = \frac{1}{K} \sum_{k=1}^{K} \hat{KL}_k.
$$

For model selection, \(\hat{KL}'\) is computed for all model candidates (the Gaussian width \(\tau\) in the current setting) and the candidate that maximizes \(KL'\) is chosen.

3. KLIEP with a Mixture of Probabilistic Principal Component Analyzers (PPCsAs)

In this section, we propose a new method, PPCA-mixture KLIEP (PM-KLIEP).

Instead of the linear model (1), we use a mixture of PPCAs as the importance model:

$$
\hat{w}(x) = \sum_{i=1}^{b} \pi_i p_i(x),
$$

where \(b\) is the number of mixture components, \(\{\pi_i\}_{i=1}^{b}\) are mixing coefficients, and \(\{p_i(x)\}_{i=1}^{b}\) are PPCAs:

$$
p_i(x) = (2\pi\sigma^2)^{-\frac{d}{2}} \det(C_i)^{\frac{1}{2}}
\times \exp \left( -\frac{1}{2} (x - \mu_i)^\top C_i^{-1} (x - \mu_i) \right).
$$

’det’ denotes the determinant, \(C_i = \sigma_i^2 I_d + W_i W_i^\top\), \(I_d\) is the \(d\)-dimensional identity matrix, \(^\top\) denotes the transpose, \(W_i \in \mathbb{R}^{d \times m}, \mu_i \in \mathbb{R}^{d}, \sigma_i > 0\) are parameters, \(d\) is the dimensionality of the input space, and \(m \leq d\) is the dimensionality of the latent space. Then the PM-KLIEP optimization problem becomes

$$
\max_{\{\pi, \omega, \mu, \sigma\}_{i=1}^{b}} \left[ \sum_{i=1}^{b} \ln \left( \sum_{i=1}^{b} \pi_i p_i(x_{ij}^{nu}) \right) \right]
$$

s.t. \(1\)

\(\frac{1}{n_{de}} \sum_{i=1}^{n_{de}} \sum_{i=1}^{b} \alpha_i \varphi_i(x_{ij}^{de}) = 1\) and \(\alpha_1, \ldots, \alpha_b \geq 0\).

2.3 Model Selection by Likelihood Cross-Validation

In the KLIEP algorithm, the choice of the Gaussian width \(\tau\) is critical. Since KLIEP is based on the maximization of the score \(KL'\), it would be natural to determine \(\tau\) so that \(KL'\) is maximized.

The expectation over \(p_m(x)\) involved in \(KL'\) can be approximated by likelihood cross-validation (LCV) as follows [14]: First divide the “numerator” samples \(X_{nu}^{nu} = [x_{ij}]_{j=1}^{m}\) into \(K\) disjoint subsets \(X_{\alpha}^{nu}\) of (approximately) the same size. Then obtain an importance estimator \(\hat{w}_k(x)\) from \(X_{\alpha}^{nu}\) and approximate the score KL using \(X_{\alpha}^{nu}\) as

$$
\hat{KL}_k = \frac{1}{|X_{\alpha}^{nu}|} \sum_{x \in X_{\alpha}^{nu}} \ln \hat{w}_k(x).
$$

This procedure is repeated for \(k = 1, \ldots, K\) and the average of \(\hat{KL}_k\) over all \(k\) is used as an estimate of \(KL'\):

$$
\hat{KL}' = \frac{1}{K} \sum_{k=1}^{K} \hat{KL}_k.
$$

For model selection, \(\hat{KL}'\) is computed for all model candidates (the Gaussian width \(\tau\) in the current setting) and the candidate that maximizes \(KL'\) is chosen.
Initialization step: Initialize the parameter values \( \{\pi_i, W_i, \mu_i, \sigma_i\}^b_{i=1} \).

E-step: Compute the ‘responsibility’ values \( \gamma_{ij}^b_{j=1} \) using the current parameters \( \{\pi_i, W_i, \mu_i, \sigma_i\}^b_{i=1} \):

\[
\gamma_{ij} = \frac{\pi_j p_j(x_i^m)}{\sum_{k=1}^b \pi_k p_k(x_i^m)}.
\]

M-step: Re-estimate the parameters \( \{\pi_i, W_i, \mu_i, \sigma_i\}^b_{i=1} \) using the current responsibility values \( \gamma_{ij}^b_{j=1} \):

\[
\pi_j = \frac{n_{de} \sum_{j=1}^{n_{mu}} \gamma_{ij}}{n_{mu} \sum_{j=1}^{b} \gamma_{ij} p_i(x_i^m)},
\]

\[
W_j = \left( \sum_{j=1}^{n_{mu}} \gamma_{ij} (x_i^m - \mu_j) (\sum_{j=1}^{n_{mu}} \gamma_{ij} C_{ij}) \right)^{-1},
\]

\[
\mu_j = \frac{\sum_{j=1}^{n_{mu}} \gamma_{ij} (x_i^m - W_j z_{ij})}{\sum_{j=1}^{n_{mu}} \gamma_{ij}},
\]

\[
\sigma_j^2 = \frac{1}{d \sum_{j=1}^{n_{mu}} \gamma_{ij}} \left( \sum_{j=1}^{n_{mu}} \gamma_{ij} (x_i^m - \mu_j)^2 - 2 \gamma_{ij} z_{ij}^\top W_j (x_i^m - \mu_j) + \gamma_{ij} \text{tr}(C_{ij} W_j^\top W_l) \right),
\]

where \( z_{ij} \) and \( C_{ij} \) are the expectations of the latent variable and its covariance [16]:

\[
z_{ij} = M_j^{-1} W_j (x_i^m - \mu_j),
\]

\[
C_{ij} = \sigma_j^2 M_j^{-1} + z_{ij} z_{ij}^\top,
\]

with \( M_j = \sigma_j^2 I + W_j^\top W_l \).

Evaluation step: Repeat the E- and M-steps until the log-likelihood converges:

\[
\sum_{j=1}^{n_{mu}} \ln \left( \sum_{j=1}^{b} \pi_j p_j(x_i^m) \right).
\]

When the dimensionality of the latent space, \( m \), is equal to the entire dimensionality \( d \), the proposed PM-KLIEP is reduced to the Gaussian-mixture KLIEP (GM-KLIEP) [15]. Thus PM-KLIEP can be regarded as an extension of GM-KLIEP to (locally) rank-deficient data.

4. Experiments

In this section, we compare the performance of PM-KLIEP with the original KLIEP and GM-KLIEP.

4.1 Illustrative Example

Let us first consider a locally rank-deficient two-dimensional importance estimation problem. The “denominator” and “numerator” densities are defined as

\[
p_{de}(x) = \frac{1}{2} N \left( x; \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & \epsilon \end{bmatrix} \right) - \frac{1}{2} N \left( x; \begin{bmatrix} 0 & \epsilon & 0 & 2 \end{bmatrix} \right),
\]

where \( N(\cdot; \mu, \Sigma) \) denotes the Gaussian density with mean \( \mu \) and covariance matrix \( \Sigma \), and \( \epsilon = 10^{-15} \). We set \( n_{de} = 100 \) and \( n_{mu} = 1000 \). In KLIEP, we set \( b = 100 \) and use the Gaussian kernel as the basis function; the kernel width is chosen based on 5-fold LCV. In GM-KLIEP and PM-KLIEP, we use the k-means clustering algorithm for parameter initialization [17], and we choose the number of mixture components based on 5-fold LCV. In PM-KLIEP, the dimension of the latent space is also chosen based on 5-fold LCV.

Figure 1 depicts the true importance function, along with the importance functions estimated by KLIEP, GM-KLIEP, and PM-KLIEP, respectively. As can be seen, PM-KLIEP can approximate the true importance function more favorably than KLIEP and GM-KLIEP for this rank-deficient example.

4.2 Application to Inlier-Based Outlier Detection

Next, we compare the performance of the proposed PM-KLIEP method with KLIEP and GM-KLIEP in inlier-based outlier detection [6].

The problem of inlier-based outlier detection is to identify outliers in an “evaluation” dataset based on a “model” dataset that only contains inliers. If the density ratio of two datasets is considered, the importance values for inliers are close to one, while those for outliers tend to significantly deviate from one. Thus, the values of the importance may be used as an outlier score.

The IDA benchmark datasets [19] are used for performance evaluation; we exclude the “splice” dataset since it is discrete. The original datasets are binary classification and each set consists of positive/negative and training/test samples. We regard the positive samples as inliers and the
negative samples as outliers. We form the model dataset by including all the positive training samples, and the evaluation dataset by including all the positive test samples and the first 5% of the negative test samples in the database. We assign the evaluation dataset to the denominator of the ratio and the model dataset to the numerator of the ratio. Thus, a sample with the importance value significantly less than one is likely to be an outlier.

In the evaluation of outlier detection performance, it is important to take into account both the detection rate (the amount of true outliers that an outlier detection algorithm can find) and the detection accuracy (the amount of true inliers that an outlier detection algorithm misjudges as outliers). Since there is a trade-off between the detection rate and the detection accuracy, we adopt the area under the ROC curve (AUC) as our error metric.

The results are summarized in Table 1, showing that PM-KLIEP clearly outperforms KLIEP and compares favorably with GM-KLIEP with a small margin.

### 5. Conclusions

In this paper, we proposed a new importance estimation method using a mixture of probabilistic principal component analyzers (PPCA). Optimization of the proposed algorithm, PM-KLIEP, can be efficiently carried out by the expectation-maximization algorithm. The usefulness of the proposed approach was illustrated through experiments.

We thank SCAT, AOARD, the JST PRESTO program, and Monbukagakusho Scholarship.

### References


