Stationary Subspace Analysis

Paul von Bünau, Frank C. Meinecke, and Klaus-Robert Müller

Machine Learning Group, CS Dept., TU Berlin, Germany
{buenau,meinecke,krm}@cs.tu-berlin.de

Abstract. Non-stationarities are an ubiquitous phenomenon in time-series data, yet they pose a challenge to standard methodology: classification models and ICA components, for example, cannot be estimated reliably under distribution changes because the classic assumption of a stationary data generating process is violated. Conversely, understanding the nature of observed non-stationary behaviour often lies at the heart of a scientific question. To this end, we propose a novel unsupervised technique: Stationary Subspace Analysis (SSA). SSA decomposes a multivariate time-series into a stationary and a non-stationary subspace. This factorization is a universal tool for furthering the understanding of non-stationary data. Moreover, we can robustify other methods by restricting them to the stationary subspace. We demonstrate the performance of our novel concept in simulations and present a real world application from Brain Computer Interfacing.

Keywords: Non-Stationarities, Source Separation, BSS, Dimensionality Reduction, Covariate Shift, Brain-Computer-Interface, BCI.

1 Introduction

The assumption that the distribution of the observed data is stationary is a cornerstone of mainstream statistical modelling: ordinary regression and classification models, for instance, rely on the fact that we can generalize from a sample to the population. Even if we are primarily interested in making accurate predictions as in Machine Learning, differences in training and test distributions can cause severe drops in performance [9], because the paradigm of minimizing the expected loss approximated by the training sample is no longer consistent. The same holds true for unsupervised approaches such as PCA and ICA.

As many real data sources are inherently non-stationary, researchers have long tried to account for that. In inferential statistics, the celebrated Heckman [4] bias correction model attempts at obtaining unbiased parameter estimates under sample selection bias; cointegration methods [2] aim at discovering stable relationships between non-stationary time-series. In order to improve predictive performance, several approaches have been developed: explicit modeling of the non-stationary process; tracking rsp. adapting to non-stationarity via online learning [7]; constructing features that are invariant under non-stationarities

T. Adali et al. (Eds.): ICA 2009, LNCS 5441, pp. 1–8, 2009.

[1] and correcting for biased empirical risk estimates under covariate shift by reweighting [10].

In this paper, we propose a novel decomposition paradigm that, to the best of our knowledge, has not been explored so far. The premise is similar in spirit to the ICA setting [5], in that we assume that a multivariate time-series was generated as a mixture of sources. However, instead of assuming independent sources, we suppose that some of the sources are stationary while others are not. The proposed Stationary Subspace Analysis (SSA) algorithm then finds a factorization into a stationary and a non-stationary component from observed data. Apart from shedding light on the nature of the non-stationarities, this decomposition can be used to build robust learning systems by confining them to the stationary subspace as we exemplify in an application to Brain-Computer-Interfacing.

Previous work along similar lines has addressed the question of whether two samples come from the same distribution [3], without trying to find subspaces where the distribution stays the same. While ICA finds independent sources, SSA divides the input space into a stationary and a non-stationary component regardless of independence within or between the subspaces.

2 Problem Formalization

We assume that the non-stationary behaviour of the data generating process is confined to a linear subspace of the D-dimensional data space, i.e. there exists a d-dimensional subspace that is undisturbed by the nonstationarity. Formally, we assume that the analyzed system generates d stationary source signals $s^{\mathfrak{s}}(t) = [s_1(t), s_2(t), \ldots, s_d(t)]^\top$ (also referred to as $\mathfrak{s}$-*sources*) and $D-d$ non-stationary source signals $s^{\mathfrak{n}}(t) = [s_{d+1}(t), s_{d+2}(t), \ldots, s_D(t)]^\top$ (also $\mathfrak{n}$-*sources*).

The observed signals $x(t)$ are then modeled as linear superpositions of these sources

$$x(t) = As(t) = \begin{bmatrix} A^{\mathfrak{s}} \; A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix} \tag{1}$$

where A is an invertible matrix.[1] The columns of $A^{\mathfrak{s}}$ span the subspace that the $\mathfrak{s}$-sources live in, we will refer to this space as $\mathfrak{s}$-*subspace*. Similarly, the span of $A^{\mathfrak{n}}$ will be called $\mathfrak{n}$-*subspace*. The goal is now to estimate a linear transformation B from the observed data $x(t)$ that separates the $\mathfrak{s}$-sources from the $\mathfrak{n}$-sources, i.e. factorizes the observed signals according to eq. (1).

Figure 1 shows an example of partly non-stationary two-dimensional time series, i.e. the upper time course is non-stationary and the lower one is stationary. Due to the non-stationarity, the scatter plots of different parts of the data are quite different. Note, that PCA or ICA would not be able to perform the separation task due to the strong correlations between the two signals.

[1] For simplicity we assume A to be a square matrix. Note that we do *not* assume the sources to be statistically independent or even uncorrelated.

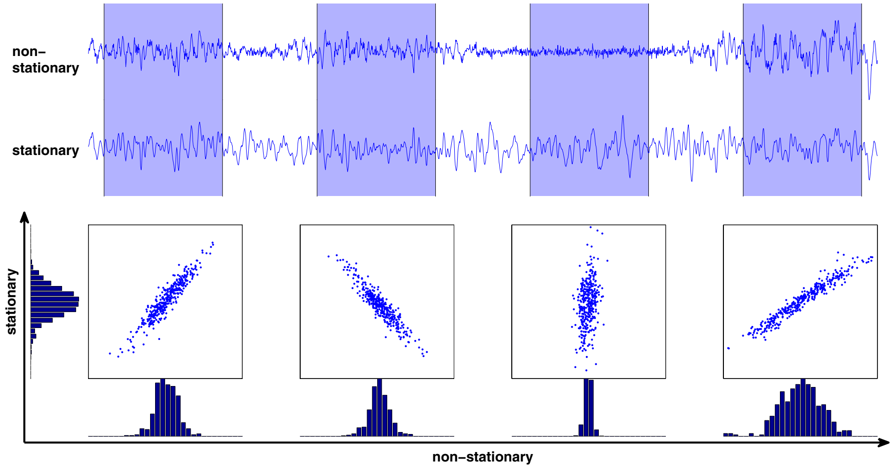

Fig. 1. A two-dimensional time series consisting of a one dimensional stationary and a one-dimensional non-stationary part. The marginal distributions of different sections of the time series are different in the $\mathfrak{n}$-subspace (horizontal histograms), but constant in the $\mathfrak{s}$-subspace (vertical histogram on the left side). Note, that the distributions in these subspaces are in general not independent.

3 Identifiability and Uniqueness of the Solution

Given the mixing model, the first question is whether it is in principle possible to invert it given only the observed (i.e. mixed) signals. In other words: can the mixing matrix A be uniquely identified or are there symmetries that give rise to multiple solutions?[2] Obviously, the basis *within* each of the two subspaces (stationary or not) can only be identified up to arbitrary linear transformations in these subspaces. However, the answer to the question whether or not the two subspaces themselves are uniquely identifiable is less obvious. Let us write the estimated mixing and demixing matrices as

$$\hat{A} = \left[\hat{A}^{\mathfrak{s}}\ \hat{A}^{\mathfrak{n}}\right] \quad \text{and} \quad \hat{B} = \hat{A}^{-1} = \begin{bmatrix} \hat{B}^{\mathfrak{s}} \\ \hat{B}^{\mathfrak{n}} \end{bmatrix}. \tag{2}$$

The matrices $\hat{B}^{\mathfrak{s}} \in \mathbb{R}^{d \times D}$ and $\hat{B}^{\mathfrak{n}} \in \mathbb{R}^{(D-d) \times D}$ denote the projections from the observed data into the estimated $\mathfrak{s}$- and $\mathfrak{n}$-subspaces. If we express the true $\mathfrak{s}$- and $\mathfrak{n}$-subspaces as linear combinations of the respective estimated subspaces

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{s}} M_3 + \hat{A}^{\mathfrak{n}} M_4 \end{aligned} \tag{3}$$

[2] And if so, how do these solutions differ?

(with $M_1 \in \mathbb{R}^{d \times d}$, $M_2 \in \mathbb{R}^{(D-d) \times d}$, $M_3 \in \mathbb{R}^{d \times (D-d)}$, and $M_4 \in \mathbb{R}^{(D-d) \times (D-d)}$) then the composite transformation (the true mixing followed by the estimated demixing matrix) reads

$$\hat{B}A = \begin{bmatrix} \hat{B}^{\mathfrak{s}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{s}} A^{\mathfrak{n}} \\ \hat{B}^{\mathfrak{n}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} M_1 & M_3 \\ M_2 & M_4 \end{bmatrix} \tag{4}$$

The estimated $\mathfrak{s}$- and $\mathfrak{n}$-sources can now be written in terms of the true sources:

$$\begin{aligned} \hat{s}^{\mathfrak{s}} &= M_1 s^{\mathfrak{s}} + M_3 s^{\mathfrak{n}} \\ \hat{s}^{\mathfrak{n}} &= M_2 s^{\mathfrak{s}} + M_4 s^{\mathfrak{n}} \end{aligned} \tag{5}$$

Given that the estimated transformation separates stationary from non-stationary signals, M_3 must vanish since estimated signals that contain non-stationary source signals will be non-stationary as well. The estimated $\mathfrak{n}$-sources, on the other hand, might contain contributions from the true $\mathfrak{s}$-sources. So the matrices M_1, M_2, M_4 remain unconstrained[3].

From equation (3) and $M_3 = 0$ we see, that the estimated $\mathfrak{n}$-subspace is identical to the true $\mathfrak{n}$-subspace while the estimated $\mathfrak{s}$-subspace is a linear combination of true $\mathfrak{s}$- and $\mathfrak{n}$-subspaces[4]. Since this linear combination is arbitrary, the estimated $\mathfrak{s}$-subspace can always be chosen such that it is orthogonal to the $\mathfrak{n}$-subspace. If we additionally chose orthogonal bases *within* each of these estimated subspaces, we have effectively restricted ourselves to the estimation of an *orthogonal* mixing matrix.

This means that we can restrict our search for the mixing matrix to the space of orthogonal matrices even if the model allows general (non-orthogonal) mixing matrices. As a result, the estimated $\mathfrak{s}$-sources will be linear combinations of the true $\mathfrak{s}$-sources but well separated from the $\mathfrak{n}$-sources.

4 Estimating the Stationary Subspace

We will now formulate an optimization criterion that allows us to estimate a demixing matrix $\hat{B}$ given N sets of data $\mathcal{X}_1, \ldots, \mathcal{X}_N \subset \mathbb{R}^D$. These datasets may for example correspond to epochs of a time series as in Figure 1. More precisely, we want to find an orthogonal transformation $\hat{B}$ such that the first d components of the transformed data $\hat{s}(t) = \hat{B}x(t)$ are as stationary as possible. Since these components are completely determined by the sub-matrix $\hat{B}^{\mathfrak{s}}$, the cost function can only depend on this sub-matrix, even if the optimization takes place in the full space of orthogonal matrices. For the latter calculations it is convenient to express $\hat{B}^{\mathfrak{s}}$ in terms of the complete demixing matrix as $\hat{B}^{\mathfrak{s}} = I^d \hat{B}$, where $I^d \in \mathbb{R}^{d \times D}$ is the identity matrix truncated to the first d rows.

[3] Apart from the mere technical assumption of invertibility of $\hat{A}$.

[4] Note that for the estimated sources this is the other way round: while the estimated $\mathfrak{s}$-sources are mixtures of the true $\mathfrak{s}$-sources only, the estimated $\mathfrak{n}$-sources are mixtures of both the true $\mathfrak{n}$- and $\mathfrak{s}$-sources.

We will consider a set of d estimated sources as stationary, if the joint distribution of these sources stays the same over all sets of samples. Therefore, our objective function is based on minimizing the pairwise distance between the distributions of the projected data which we measure using the Kullback-Leibler divergence. For technical reasons, we only consider differences in the first two moments, i.e. two distributions are assumed to be the same if they have the same mean and covariance matrix. Consequently, our method is ignorant of any non-stationary process that does not change the first two moments. Even though the rationale for this restriction is purely technical, its practical consequences should be limited.[5] In order to compute the KL divergence, we need to estimate both densities. Since we have restricted ourselves to distributions whose sufficient statistics consist of the mean and covariance matrix, a natural choice is to use a Gaussian approximation according to the maximum entropy principle [6].

Let $(\hat{\mu}_i, \hat{\Sigma}_i)$ be the estimate of mean and covariance for dataset $\mathcal{X}_i$. Then we find the demixing matrix as the solution of the optimization problem

$$\hat{B} = \operatorname*{argmin}_{BB^\top=I} \sum_{i<j} \mathrm{KL}\left[\mathcal{N}(I^dB\hat{\mu}_i, I^dB\hat{\Sigma}_i(I^dB)^\top) \;||\; \mathcal{N}(I^dB\hat{\mu}_j, I^dB\hat{\Sigma}_j(I^dB)^\top)\right].$$

To stay on the manifold of orthogonal matrices (or, more formally speaking: the special orthogonal group SO(N)), we will employ a multiplicative update scheme: starting with $B = I$, we multiply B in each iteration by an orthogonal matrix, $B^{\mathrm{new}} \leftarrow RB$. At each step, the estimated projection to the $\mathfrak{s}$-subspace is then given by I^dRB and we can write the projected mean and covariance matrix of data set $\mathcal{X}_i$ as

$$\hat{\mu}_i^{\mathfrak{s}} = I^dRB\hat{\mu}_i \quad \text{and} \quad \hat{\Sigma}_i^{\mathfrak{s}} = I^dRB\hat{\Sigma}_i(I^dRB)^\top,$$

and the loss function as

$$L_B(R) = \sum_{i<j} \log \frac{\det(\hat{\Sigma}_j^{\mathfrak{s}})}{\det(\hat{\Sigma}_i^{\mathfrak{s}})} + \mathrm{tr}\left((\Sigma_j^{\mathfrak{s}})^{-1}\Sigma_i^{\mathfrak{s}}\right) + (\mu_j^{\mathfrak{s}} - \mu_i^{\mathfrak{s}})^\top(\Sigma_j^{\mathfrak{s}})^{-1}(\mu_j^{\mathfrak{s}} - \mu_i^{\mathfrak{s}}).$$

The rotation R can be parametrized as the matrix exponential of an antisymmetric matrix, $R = e^M$ with $M = -M^\top$, where each element M_{ij} can be interpreted as generalized rotation angle (i.e. rotating axis i towards axis j). Using this, we can express the gradient of the function $L_B(R) = L_B(e^M)$ w.r.t. M in terms of the corresponding gradient w.r.t. R (see e.g. [8]) as

$$\partial L_B/\partial M|_{M=0} = (\partial L_B/\partial R)R^\top - R(\partial L_B/\partial R)^\top. \tag{6}$$

The gradient of the loss function with respect to R is

$$\begin{aligned}\frac{\partial L_B}{\partial R} = I^{d\top}I^dR\sum_{i<j} B\Big(&\hat{\Sigma}_iQ\hat{\Sigma}_j^{-1} + \hat{\Sigma}_j^{-1}Q\hat{\Sigma}_i + \hat{\Sigma}_jQ\hat{\Sigma}_i^{-1} + \hat{\Sigma}_i^{-1}Q\hat{\Sigma}_j \\ &+ \left(\hat{\Sigma}_j^{-1} + \hat{\Sigma}_i^{-1}\right)QD_\mu + D_\mu Q\left(\hat{\Sigma}_j^{-1} + \hat{\Sigma}_i^{-1}\right)\Big)B^\top\end{aligned} \tag{7}$$

[5] Real world nonstationary processes will hardly go unnoticed by the first two moments. Moreover, estimates of higher-order moments are less stable themselves.

with $Q = B^\top R^\top I^{d\top} I^d R B$ and $D_\mu = (\hat{\mu}_i - \hat{\mu}_j)(\hat{\mu}_i - \hat{\mu}_j)^\top$. The factor of $I^{d\top} I^d$ from the left ensures that the lower $D - d$ rows of this matrix gradient vanish. This means, that in the gradient w.r.t. M the lower right block has to be zero. Furthermore, since every summand in the sum is symmetric, the skew-symmetry of eq. (6) makes the upper left block in $\partial L_B / \partial M$ vanish as well. Thus the gradient has the shape

$$\frac{\partial L_B}{\partial M}|_{M=0} = \begin{bmatrix} 0 & Z \\ -Z^\top & 0 \end{bmatrix}$$

where the non-zero part $Z \in \mathbb{R}^{d \times (D-d)}$ corresponds to the rotations between coordinates of the $\mathfrak{s}$- and $\mathfrak{n}$-space. Note that the derivative w.r.t. the rotations within the two spaces must vanish, because they do not change the solution. Thus we can reduce the number of variables to $d(D-d)$. The optimization is then carried out using a standard conjugate gradient procedure in angle space with multiplicative updates.

5 Simulations

We investigate the performance of SSA under different scenarios based on simulated data. In order to make the analysis more concise, we first consider the hypothetical case where the true mean and covariance matrix of the data generating process are known and then examine the impact of the estimation error in a second set of experiments. The experimental setup is as follows: we randomly

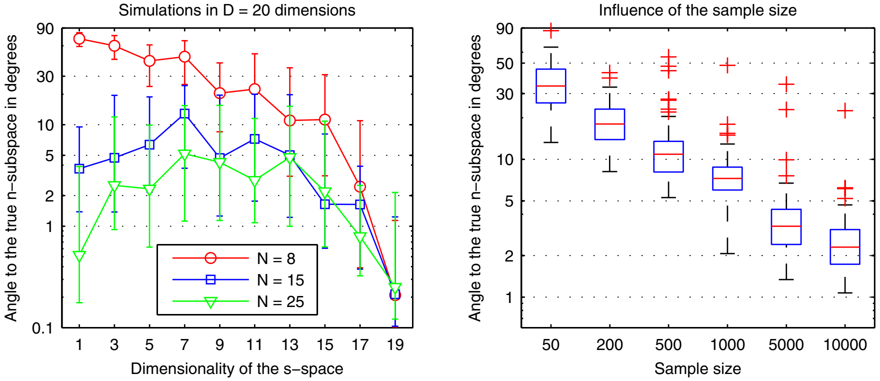

Fig. 2. The left plot shows the results of the simulations for a 20 dimensional input space. The performance of the method is measured in terms of the median angle to the true $\mathfrak{n}$-subspace (vertical axis) for several dimensionalities (horizontal axis) and number of datasets $N = 8, 15, 25$ (red, blue and green curve) over 100 random trials. The error bars stretch from the 25% to the 75% quantile. The right plot shows the performance under varying numbers of samples in each dataset for fixed $d = 5$ and $N = 10$ in $D = 10$ input dimensions.

sample N covariance matrices and means such that the $\mathfrak{s}$-subspace is spanned by the first d coordinates. Moreover, we randomly sample a well-conditioned mixing matrix A that is then applied to each mean and covariance matrix. Note that ICA cannot in principle separate the $\mathfrak{s}$- from the $\mathfrak{n}$-space because we have allowed for arbitrary correlations between them. For each trial, in order to avoid local minima, we restart the optimization procedure five times and then select the solution with the lowest objective function value. The accuracy is measured as angle between the estimated $\mathfrak{n}$-subspace and the ground truth.

From the results shown in the left plot of Figure 2, we can see that the likelihood that the true $\mathfrak{s}$-sources can be found grows with the number of available datasets and scales with the degrees of freedom $d(D-d)$. In order to analyze the influence of the number of samples within each dataset $n = |\mathcal{X}_i|, 1 \leq i \leq N$, we fix the other parameters and vary n. The right plot in Figure 2 shows the result: the performance clearly increases with the number of available samples whereas the method remains applicable even in the small sample case.

6 Application to Brain-Computer-Interfacing

We demonstrate that SSA can be used for robust learning on EEG dat (49 channels) from Brain-Computer-Interfacing (BCI) where the task is to distinguish imagined movements of the right/left hand. Imagined movements of the hand are known to produce characteristic spatio-temporal signal patterns on the respective contralateral hemispheres. However, the frequency content of these signals over the motor cortex (μ-rhythm) is in the same range as the occipital α-rhythm. The strength of the α-rhythm is task unrelated and strongly correlated to the tiredness or the exposure to visual stimulation of a subject.

In our experiment, we induce changes in the strength of the α-rhythm by first extracting it from a separate artefact measurement session (using ICA) and

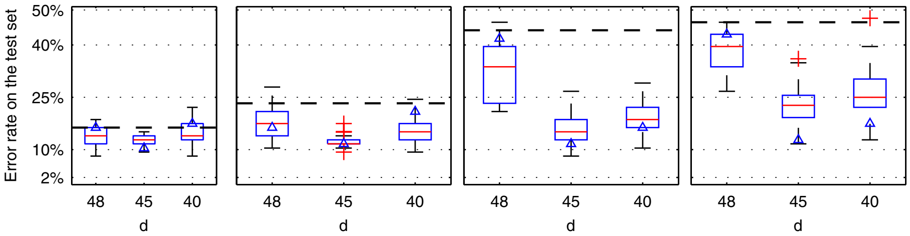

Fig. 3. Results on the BCI data for varying levels of change in the strengths of the α component between test and training data (increasing from left to right). For each scenario, the test error rate of the CSP baseline method is indicated by the dashed black line. The boxplots show the distribution of the test error rates on the $\mathfrak{s}$-subspace for varying dimensionalities $d = 48, 45, 40$ over 100 runs of the optimization. The blue triangle indicates the error rate on the $\mathfrak{s}$-subspace that attained the minimum objective function value over all runs.

then superimpose it on the data in varying strengths in order to induce realistic yet controlled non-stationarities. See [1] for a full exposition of the experimental setup. The data is divided into three parts: a training data set that contains strong α-activity which corresponds to a wakeful resting brain state and a test data set, where we change the strength of the α-components. From the test portion, we set aside the first 30 trials for adaptation. Then we estimate the $\mathfrak{s}$-space over the training and adaptation part and apply the standard CSP algorithm within the $\mathfrak{s}$-space. The performance is measured as the misclassification rate on the test set. The experimental results presented in Figure 3 show that the classification accuracy can be retained even under very strong non-stationarities if the learning algorithm is restricted to a 45 dimensional $\mathfrak{s}$-subspace.

7 Conclusion and Future Work

We have presented the first algorithm for decomposing a multivariate time-series into a stationary and a non-stationary component. A number of interesting questions remain: first of all, how can we choose the dimensionality of the $\mathfrak{s}$-space from data? Secondly, can we extend the algorithm to measure non-stationarities in higher order moments? Finally, we will pursue further applications beyond the neurosciences.

References

1. Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Nikulin, V., Müller, K.-R.: Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 113–120. MIT Press, Cambridge (2008)
2. Engle, R.F., Granger, C.W.J.: Co-integration and error correction: Representation, estimation, and testing. Econometrica 55(2), 251–276 (1987)
3. Friedman, J., Rafsky, L.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. The Annals of Statistics 7(4), 697–717 (1979)
4. Heckman, J.J.: Sample selection bias as a specification error. Econometrica 47(1), 153–162 (1979)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
6. Jaynes, E.T.: Information theory and statistical mechanics. Physical Review 160, 620–630 (1957)
7. Murata, N., Kawanabe, M., Ziehe, A., Müller, K.-R., Amari, S.: On-line learning in changing environments with applications in supervised and unsupervised learning. Neural Networks 15(4-6), 743–760 (2002)
8. Plumbley, M.D.: Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras. Neurocomputing 67, 161–197 (2005)
9. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. (eds.): Dataset Shift in Machine Learning. MIT Press, Cambridge (2008)
10. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference 90(2), 227–244 (2000)