A Machine Learning Approach to Cepheid Variable Star Classification Using Data Alignment and Maximum Likelihood

Ricardo Vilalta, Kinjal Dhar Gupta

Department of Computer Science, University of Houston
501 Philip G. Hoffman Hall, Houston TX, 77204-3010, USA
Ph. 713-743-3614 Fax. 832-842-7532

Lucas Macri

Department of Physics and Astronomy, Texas A&M University
4242 TAMU, College Station, TX 77843-4242501, USA

Abstract

Our study centers on the classification of two subtypes of Cepheid variable stars. Such classification is relatively easy to obtain for nearby galaxies, but as we incorporate new galaxies, the cost of labeling stars calls for some form of model adaptation. Adapting a predictive model to differentiate Cepheids across galaxies is difficult because of the sample bias problem in star distribution (due to the limitation of telescopes in observing faint stars as we try to reach distant galaxies). In addition, estimating the luminosity of a star as we reach distant galaxies carries some inevitable shift in the data distribution. We propose an approach to predict the class of Cepheid stars on a target domain, by first building a model on an "anchor" source domain. Our methodology then shifts the target data until it is well aligned with the source data by maximizing two different likelihood functions. Experimental results with two galaxy datasets (Large Magellanic Cloud as the source domain, and M33 as the target domain), show the efficacy of the proposed method.

Keywords: Machine Learning, Classification, Dataset Shift, Covariate Shift, Cepheid Stars.

1. Introduction

A common assumption pervading most traditional work in classification assigns the same joint probability distribution to training and testing sets. While such assumption has found a plethora of successful real-world applications, recent work in machine learning has revealed an equally rich source of applications when dispensing with such assumption. An interesting example lies in light curve classification from star samples obtained from different galaxies. A classification task set to differentiate different types of stars observed in nearby galaxies, will experience a sample bias as our study encompasses galaxies lying farther away. This is due to our limitation to observe faint stars; as the distance to a galaxy increases, our star sample will inevitable concentrate on higher-luminosity stars. This is an instance of the dataset shift problem [1, 2, 3], and precludes the direct utilization of one single model across galaxies; it calls for a form of model adaptation to compensate for the change in the data distribution.

In this study we show one approach to deal with the dataset shift problem in the context of star classification. The task is to classify Cepheid variable stars into two subtypes according to their pulsation mode: fundamental or first-overtone. From an astronomical view, a robust discrimination between fundamental and first-overtone Cepheids will enable a more precise and accurate determination of distances to nearby galaxies, which in turn impacts the uncertainty in the Hubble constant (current expansion rate of the Universe). A classification model with fairly high accuracy is in fact attainable for nearby galaxies, but the cost of labeling variable stars, together with the large number of galaxy datasets, quickly turn the goal of acquiring predictive models for many galaxies into a daunting task. Additionally, switching to a new galaxy dataset involves a shift in the data distribution, as the proportion of higher-luminosity Cepheids inevitably rises.

Our study suggests a methodology to attack the dataset shift problem in Cepheid star classification by re-using a predictive model previously obtained on a source domain. The model can be invoked again after transforming the testing or target data to account for 1) a shift in apparent mean magnitude, and 2) a sample bias over luminous stars. The alignment is effected using maximum likelihood. Our experiments show results when using the Large Magellanic Cloud as the source domain, and M33 as the target domain; following our proposed methodology, accuracy values are similar to those obtained if class labels were available on the target domain (M33). Our specific contribution lies on a concrete methodology within machine learning and data mining specialized to star classification, where predictive models are re-used across galaxy datasets.

This paper is organized as follows. Section 2 provides basic concepts in classification and a brief introduction to Cepheid stars. Section 3 explains our proposed methodology. Section 4 describes related work. Section 5 shows our empirical results. Lastly, section 6 gives a summary and conclusions.

2. Preliminary Concepts

2.1. Cepheid Variables

Classical Cepheid variables (also commonly referred to as Population I Cepheids) are stars both massive 4-11 M_{Sun} (solar masses) and luminous 1×10^3 - 5×10^4 L_{Sun} (solar luminosities), that undergo regular pulsations with periods ranging from 2 to 100 days. The pulsations are driven by cyclical changes in the opacity of hydrogen and helium in the outer atmosphere of these stars [4]. A tight correlation exists between period (P) and luminosity (L) for Cepheid variables:

$$\log_{10} L = a + b * \log_{10} P \tag{1}$$

where a and b are fitting parameters. Figure 1 shows an example of such relation for a sample of Cepheids populating the Large Magellanic Cloud galaxy. The flux of each star is represented by its apparent magnitude m, defined as $m = -2.5 \times \log_{10} \frac{L}{d^2}$, where d is the distance from Earth to the star measured in parsecs. Hence, smaller numbers correspond to brighter magnitudes (higher fluxes).

Our confidence on the modern version of the "Cepheid P-L relation" is supported by the small intrinsic data dispersion (only 4% of the mean luminosity at fixed period for near- and mid-infrared wavelengths [5]). Briefly, the physical reason behind the P-L relation can be understood as follows. Stars with different masses have different central temperatures and densities, which greatly influence the rate of energy generation via fusion. Stars with different masses will achieve different equilibrium conditions (surface luminosity, "L" in equation 1, as well as temperature

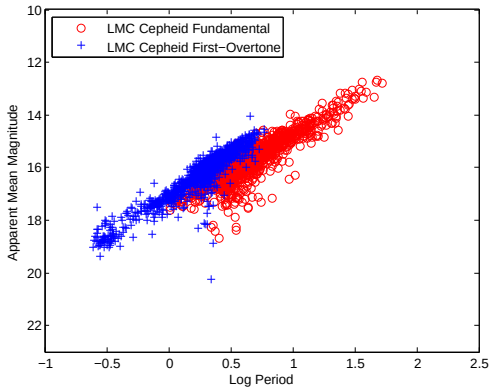

Figure 1: A linear relation exists between the logarithm of the period and the magnitude of a Cepheid star. The graph shows both "fundamental" and "first-overtone" Cepheids from the Large Magellanic Cloud that overlap more than expected due to interstellar dust and poor image resolution.

and radius) at various stages of their evolution. For stars in a narrow range of surface temperature, the internal conditions are propitious for regular, stable pulsations. The period of these pulsations ("P" in equation 1) is inversely related to the density of the star. More massive (i.e. more luminous) stars have lower average densities, leading to longer pulsation periods.

Cepheid variables can be classified according to their pulsation mode(s)[1]. Our study centers on the two most abundant classes, which pulsate in the fundamental and first-overtone modes (figure 1). Their Period-Luminosity relations are shifted in luminosity by 70% and thus should be easily separable. In practice, individual Cepheids within a galaxy will encounter varying amounts of interstellar dust, which attenuates their brightness to the point that the first-overtone and fundamental P-L relations are blurred into a single relation. Additionally, the superposition of multiple stars within a single resolution element of an image could bias the measured flux of a fundamental-mode Cepheid, artificially increasing its brightness to the point where it lies within the first-overtone relation. If one is unable to discriminate between fundamental and first-overtone Cepheids, a hard period cut must be imposed to ensure no overtones are present in the sample (the cut would be made at $P = 8$ days, or 0.9 in logarithm of the period in figure 1). Otherwise, first-overtones will contaminate the sample and bias the distance estimate, because first-overtone pulsators are more luminous than fundamental-mode pulsators at fixed period. Overall this is not a satisfactory solution because such a period cut will eliminate $> 50\%$ of the fundamental mode sample and greatly reduce the precision of the distance determination.

The observational limitations described above will blur the fundamental and first-overtone relations of short-period ($P < 8$ days) Cepheids into a single relation. Since overtone Cepheids are naturally more luminous at fixed period, their inclusion in a sample of fundamental mode Cepheids will bias the slope of the P-L relation towards shallower (i.e., smaller) values. The astronomical community needs an efficient way to classify the pulsation mode of a Cepheid into fundamental and first-overtone to better understand the intrinsic slope of the P-L relation ("b" in equation 1). Specifically, a robust discrimination between fundamental-mode and first-overtone pulsators would enable a comparison of the slope of the period-luminosity relation of each type, which can be compared with the predictions of stellar evolution and pulsation models.

[1]Information about data sources and plotted data can be found in section 5.

These models compute changes in the internal structure and chemical compositions of stars from their initial conditions (in the so-called "main sequence") until the end of the Cepheid phase and beyond. Incorrect assumptions in the initial conditions, or flaws in the modeling, become readily apparent when comparing the predicted and observed properties of the Cepheid period-luminosity relation. For such a study, covering as much of the entire range of Cepheid periods (4 days $\lesssim P \lesssim$ 100 days) as possible is critical to reliably determine the value of the slope. Nearby galaxies (such as the Large Magellanic Cloud, Messier 33, Messier 81 and Messier 101) are good laboratories for such a study because their Cepheids are bright enough to be studied in detail, even at the shortest periods (lowest luminosities).

2.2. Basic Concepts In Classification and Dataset Shift

We now introduce basic notation in classification. We assume a classifier receives as input a set of training examples $T_{\rm tr} = \{(\mathbf{x}^k, w^k)\}_{k=1}^{p}$, where $\mathbf{x}$ is a vector in the input space $\mathcal{X}$, and w is a value in the (discrete) output space $\mathcal{W}$. We assume the training or source sample $T_{\rm tr}$ consists of independently and identically distributed (i.i.d.) examples obtained according to a fixed but unknown joint probability distribution, $P_{\rm tr}(\mathbf{x}, w)$, in the input-output space $\mathcal{X} \times \mathcal{W}$. The outcome of the classifier is a hypothesis or function $f(\mathbf{x}|\theta)$ (parameterized by θ) mapping the input space to the output space, $f : \mathcal{X} \rightarrow \mathcal{W}$. In our study, $T_{\rm tr}$ is a dataset corresponding to a source galaxy (e.g., Large Magellanic Cloud), where vector $\mathbf{x}$ contains light-curve properties; we focus exclusively on period (x_1) and apparent mean magnitude (x_2), such that $\mathbf{x} = (x_1, x_2)$. Variable w can take on any of the two classes we wish to predict (fundamental or first-overtone).

We will also assume a testing or target sample with no class labels, corresponding to a different galaxy $T_{\rm te} = \{\mathbf{x}^k\}_{k=1}^{q}$, that consists of i.i.d. examples obtained from the marginal distribution $P_{\rm te}(\mathbf{x})$ according to a different joint distribution, $P_{\rm te}(\mathbf{x}, w)$, over $\mathcal{X} \times \mathcal{W}$. In the simplest scenario, we assume that the only difference between the training and test distributions is in the marginal distributions $P_{\rm tr}(\mathbf{x}) \neq P_{\rm te}(\mathbf{x})$, while the class posteriors (i.e., conditional distributions) remain identical $P_{\rm tr}(w|\mathbf{x}) = P_{\rm te}(w|\mathbf{x})$. This is known as the *covariate shift problem*. The output will still be a function $f : \mathcal{X} \rightarrow \mathcal{W}$, but with the goal of minimizing generalization error over the "test" distribution (as estimated from $T_{\rm te}$).

3. Addressing The Data Alignment and Sample Bias Problems

Although accurate classification of Cepheids into fundamental and first-overtone can be attained for nearby galaxies (e.g., Large Magellanic Cloud), the high cost of manually labeling variable stars suggests a mode of operation where a predictive model obtained on a data set from a source galaxy $T_{\rm tr}$, is later used on a test set from a target galaxy $T_{\rm te}$. Such scenario is not immediately attainable because of two main problems: 1) a shift in the data distribution along apparent magnitude m as we reach galaxies lying farther away, and 2) a significant degree of sample bias where $P_{\rm tr}(\mathbf{x}) \neq P_{\rm te}(\mathbf{x})$, that increases the probability density over higher-luminosity stars. An example of the effect of these two problems is shown in figure 2 (left), where the distribution of Cepheids in the Large Magellanic Cloud (top sample), deviates significantly from that of M33 (bottom sample). Both the offset along apparent magnitude and the significant degree of sample bias are mostly due to the fact that M33 is $\sim 16\times$ farther than the LMC. At these greater distances, the shorter-period (i.e, less luminous) Cepheids fall below the detection threshold and longer-period (i.e., more luminous) stars are preferentially detected. A less likely, but still possible scenario could occur when deep observations of a faraway galaxy are compared to shallow

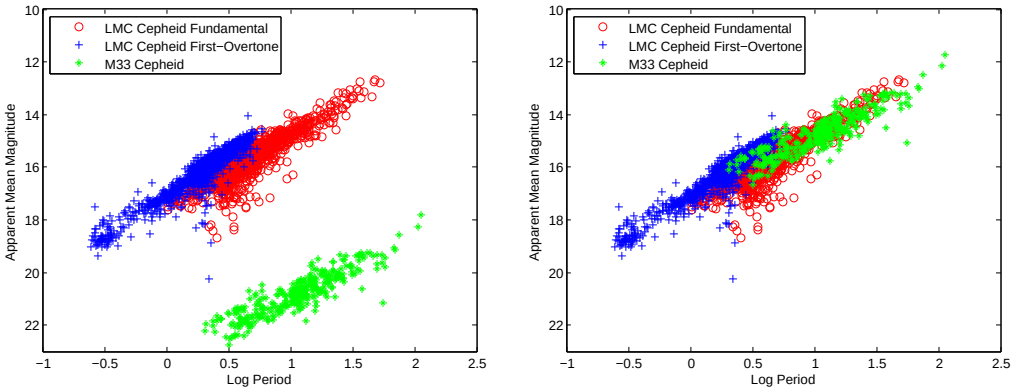

Figure 2: Left. The distribution of Cepheids along the Large Magellanic Cloud LMC (top sample), deviates significantly from M33 (bottom sample). Right. M33 is aligned with LMC by shifting along mean magnitude.

observations of a nearby galaxy. In this case, a sample bias would still take place, but it would occur in the Cepheids of the closer galaxy.

Our general solution is to shift the data on T_{te} to the point where the model obtained over T_{tr} is directly applicable. We have named this the *data-alignment problem* in star classification. This differs from previous methodologies where the goal is to generate a model afresh by re-weighting examples on the training set [1]; instead we leave the model intact, and transform the testing set until the model becomes applicable. An example of a successful alignment is shown in figure 2 (right), where shifting the M33 sample along m by the right amount, enables us to do classification with the same model obtained on T_{tr}.

3.1. Problem Formulation

We formalize our problem as follows. We assume a training set T_{tr} and a predictive model $f(\mathbf{x}|\theta)$ obtained by training on dataset T_{tr}. We assume also an unlabeled testing set $T_{\text{te}} = \{\mathbf{x}\}$, where each feature vector $\mathbf{x} = (x_1, x_2)$, is made of two features[2] corresponding to period (x_1) and apparent mean magnitude (x_2). We assume feature x_2 has suffered a shift (because the galaxy corresponding to T_{te} lies farther away than the galaxy corresponding to T_{tr}). We wish to generate a new dataset $T'_{\text{te}} = \{x'\}$, where $\mathbf{x}' = (x_1, x_2 + \delta)$, to achieve the goal of aligning T_{te} with T_{tr}. Our real focus is on the value of δ; we aim at an optimal value that best achieves such alignment. Note that even if the alignment is done appropriately, the joint distributions between training and testing could differ, because $P_{\text{tr}}(\mathbf{x}) \neq P_{\text{te}}(\mathbf{x})$ due to a form of sample bias. In our case, the sample bias favors highly luminous Cepheids.

3.2. Step 1: Maximum Likelihood Based On Shift

Our methodology to attack the problem above consists of first shifting the data T_{te} using maximum likelihood to generate a first estimate of δ at low computational cost. A second step will

[2]Besides period and magnitude, many other features have been proposed in the literature to characterize light curves [6, 7].

provide a more refined estimate by considering the expected proportion of class labels in T_{te} as predicted by the model $f(\mathbf{x}|\theta)$ (section 3.3), albeit at a higher computational cost.

Density Estimation. We start by modeling the training set T_{tr} as a mixture of two Gaussians, based on the bimodal distribution observed along the two existing classes: fundamental and first-overtone (figure 1), and assume two features (x_1, x_2), corresponding to period and apparent mean magnitude respectively[3]:

$$P_{\text{tr}}(x_1, x_2) = \sum_{i=1}^{2} \phi_i \frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-\rho_i^2}} e^{\frac{-v_i}{2(1-\rho_i^2)}} \tag{2}$$

where $\{\phi_i\}$ are the mixing components ($0 \leq \phi_i \leq 1$, $\phi_1 + \phi_2 = 1$), μ_{ij} and σ_{ij} are the mean and standard deviation of the ith component along the jth feature, ρ_i is the correlation between x_1 and x_2 in the ith component, and

$$v_i = \frac{(x_1 - \mu_{i1})^2}{\sigma_{i1}^2} - \frac{2\rho_i(x_1 - \mu_{i1})(x_2 - \mu_{i2})}{\sigma_{i1}\sigma_{i2}} + \frac{(x_2 - \mu_{i2})^2}{\sigma_{i2}^2} \tag{3}$$

We estimate parameters $\{\mu_i\}$ and $\{\phi_i\}$ and $\{\sigma_{ij}\}$ directly from our sample T_{tr}, since we know all class labels (i.e., we know which points belong to each component or Gaussian distribution). This enables us to have a complete characterization of the training set distribution $P_{\text{tr}}(\mathbf{x})$.

Alignment. Our next step is to define a new testing set $T'_{\text{te}} = \{x'\}$, where $\mathbf{x}' = (x_1, x_2 + \delta)$, since we know a shift has occurred along apparent mean magnitude. In this case we redefine v_i as z_i:

$$z_i = \frac{(x_1 - \mu_{i1})^2}{\sigma_{i1}^2} - \frac{2\rho_i(x_1 - \mu_{i1})(x_2 + \delta - \mu_{i2})}{\sigma_{i1}\sigma_{i2}} + \frac{(x_2 + \delta - \mu_{12})^2}{\sigma_{i2}^2} \tag{4}$$

Our approach is to find the δ that maximizes the log likelihood of T'_{te} with respect to distribution $P_{\text{tr}}(\mathbf{x})$:

$$\mathcal{L}(\delta|T'_{\text{te}}) = \log \Pi_{k=1}^{q} P_{\text{tr}}(\mathbf{x}^k) = \sum_{k=1}^{q} \log P_{\text{tr}}(x_1^k, x_2^k) \tag{5}$$

To proceed, we express our joint distribution alternatively as follows:

$$P_{\text{tr}}(x_1^k, x_2^k) = \sum_{i=1}^{2} \phi_i \alpha_i e^{-\beta_i z_i^k}, \quad \text{where} \quad \alpha_i = \frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-\rho_i^2}} \text{ and } \beta_i = \frac{1}{2(1-\rho_i^2)} \tag{6}$$

After some algebraic manipulation, it can be shown that:

$$e^{-\beta_i z_i^k} = e^{c_{i1}^k + \delta c_{i2}^k + \delta^2 c_{i3}^k} \tag{7}$$

where

$$c_{i1}^k = -\beta_i \left(\frac{(x_1^k - \mu_{i1})^2}{\sigma_{i1}^2} - \frac{2\rho_i(x_1^k - \mu_{i1})(x_2^k - \mu_{i2})}{\sigma_{i1}\sigma_{i2}} + \frac{(x_2^k - \mu_{i2})^2}{\sigma_{i2}^2} \right)$$

[3]Our limited feature space obviates a vectorial representation for the joint density (mixture of two Gaussians).

$$c_{i2}^k = -\beta_i \left(\frac{2(x_2^k - \mu_{i2})}{\sigma_{i2}^2} - \frac{2\rho_i(x_1^k - \mu_{i1})}{\sigma_{i1}\sigma_{i2}} \right) \quad \text{and} \quad c_{i3}^k = \frac{-\beta_i}{\sigma_{i2}^2}$$

To find the δ that maximizes $\mathcal{L}(\delta|T'_{\text{te}})$, we use a derivative approach (see appendix for details):

$$\frac{d}{d\delta}\left(\sum_{k=1}^{q} \log P_{\text{tr}}(x_1^k, x_2^k) \right) = \sum_{k=1}^{q} \left(\frac{\gamma_{11}^k e^{\theta_1^k(\delta)} + \gamma_{21}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}} + \delta \frac{\gamma_{12}^k e^{\theta_1^k(\delta)} + \gamma_{22}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}} \right) = 0 \tag{8}$$

where

$$\theta_i^k(\delta) = \delta c_{i2}^k + \delta^2 c_{i3}^k, \;\; \gamma_{i1} = \phi_i \alpha_i c_{i2}^k e^{c_{i1}^k}, \;\; \gamma_{i2}^k = 2\phi_i \alpha_i c_{i3}^k e^{c_{i1}^k}, \;\; \text{and} \;\; \Gamma_i^k = \phi_i \alpha_i e^{c_{i1}^k}$$

The algebraic manipulations above convert an optimization problem into a simple equation-solving problem. Specifically, solving for δ (equation 8) enables us to know the exact shift necessary to generate the new testing set T'_{te}.

3.3. Step 2: Maximum Likelihood Based On The Induced Class Distribution

The previous section explains how to obtain an initial approximation to the alignment problem using maximum likelihood after estimating the probability density of $P_{\text{tr}}(\mathbf{x})$. We now describe a more refined search that takes into account class distributions and the predictions of model $f(\mathbf{x}|\theta)$. The rationale is as follows. While a testing set T_{te} lacks any class label information, it is possible to determine its likelihood according to the class distribution induced by model $f(\mathbf{x}|\theta)$. Our aim is to find the data shift δ that maximizes the likelihood of the induced class distributions. We explain this methodology next.

Class Prior Modeling. Our second search focuses on the class distribution of the predictions made by $f(\mathbf{x}|\theta)$, but not in the original class priors, since we assume an unlabeled testing set T_{te}. We model this distribution by generating multiple bootstrap samples from T_{tr}, and invoking $f(\mathbf{x}|\theta)$ on each sample to predict the class of each observation. The result is an estimation of the distribution on the proportion of examples assigned to each class as dictated by $f(\mathbf{x}|\theta)$. Since we deal with a binary classification problem, we can concentrate on the distribution for one class only (e.g., fundamental), as clearly the two distributions have the same variance. The distribution is characterized by a mean μ and variance σ^2 on the proportion of examples classified as class w. We assume a normal distribution $Q_{\text{tr}}(y) \sim \mathcal{N}(\mu, \sigma^2)$, where y is the proportion of training examples predicted as class w by $f(\mathbf{x}|\theta)$. Figure 3 shows density functions on the proportion of examples assigned to each class using LMC data and a normal distribution.

Alignment. We then proceed to find the shift δ that maximizes the log likelihood of T'_{te} (starting with the value of δ found in the previous section). We use an iterative approach, where the previous value of δ is updated using fixed increments $\delta = \delta_{\text{old}} \pm \tau$. At iteration k, the log likelihood of T'_{te} is computed as follows:

$$\mathcal{L}(\delta|T'_{\text{te}}) = \log Q_{\text{tr}}(y^k) \tag{9}$$

where y^k is the proportion of testing examples predicted as class w by $f(\mathbf{x}|\theta)$ at step k. The need to run model $f(\mathbf{x}|\theta)$ to compute the log likelihood precludes finding an analytical solution for δ. Instead we resort to a greedy-best search technique where we iterate on increments τ until we reach a maximum number of iterations, or the difference in $\mathcal{L}(\delta|T'_{\text{te}})$ between the current and previous steps is ≤ 0. The sign of the increments is decided on the first iteration, after which they always remain either positive or negative.

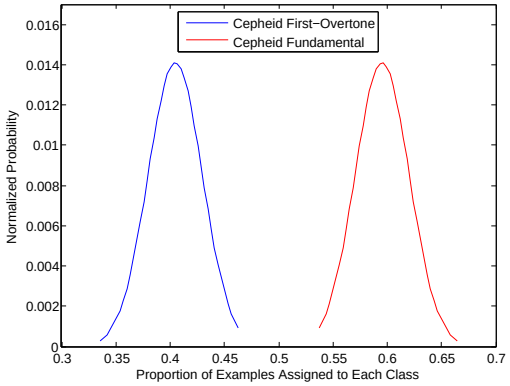

Figure 3: An estimation of the distribution on the proportion of examples assigned to each class by $f(\mathbf{x}|\theta)$ on LMC data. We assume a normal distribution for both classes.

3.4. The Sample Bias Problem

As mentioned before, even if the alignment is done appropriately, the joint distributions between training and testing could differ due to a form of sample bias. As shown in figure 2 (right), M33 appears aligned with LMC, but the concentration of stars in M33 lies at higher periods; this is considered a case of sample bias. Sample bias is one reason for the presence of covariate shift, which is characterized by a difference in the marginal distributions of the training and testing sets (i.e., occurs whenever $P_{\text{tr}}(\mathbf{x}) \neq P_{\text{te}}(\mathbf{x})$). In our case, the reason for such bias is due to our limitation to observe faint stars as we try to reach distant galaxies.

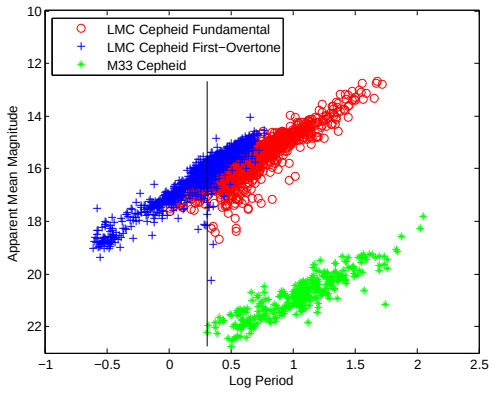

Figure 4: We address the sample bias problem by focusing on those examples on LMC that fall within the same range of values along period for M33 (right side of the vertical line).

To address the sample bias problem, we narrow the training set T_{tr} (Large Magellanic Cloud) to those examples where the period P is within the range of values found in T_{te}, (M33). Figure 4

shows the range of values under consideration; the idea is to focus the analysis to regions in the feature space with highest density based on T_{te}. The strategy followed here looks at the range of values along period P covered by the test set T_{te}. Specifically, we model the source domain as a mixture of Gaussians only along the range of values embedded by the target domain. Steps 1 and 2 as specified in sections 3.2 and 3.3 are then applied directly, without any further modification. The advantage gained with this process is to narrow our attention to local regions over the input-output space where we wish to specialize our predictive model.

4. Related Work

The covariate shift problem has been studied extensively in the past years. A common solution is to generate a new model afresh with the source training set $T_{\text{tr}} \sim P_{\text{tr}}(\mathbf{x})$, through a re-weighting scheme that gives more importance to regions of high probability density relative to $P_{\text{te}}(\mathbf{x})$ as estimated from the testing set T_{te} [2, 8, 3, 9]. Different from previous work, our problem does not initially fall into the covariate shift paradigm due to the shift in apparent mean magnitude between T_{tr} and T_{te}; this causes $P_{\text{tr}}(w|\mathbf{x}) \neq P_{\text{te}}(w|\mathbf{x})$. We attack this problem using a data-alignment solution based on maximum likelihood, after which we are able to exploit the original model $f(\mathbf{x}|\theta)$ obtained from T_{tr}, since the goal of the alignment step is to make $P_{\text{tr}}(w|\mathbf{x}) = P_{\text{te}}(w|\mathbf{x})$.

The re-weighting scheme mentioned above follows two main assumptions: 1) we can generate accurate estimations of $P_{\text{tr}}(\mathbf{x})$ and $P_{\text{te}}(\mathbf{x})$ [10, 11, 12], and 2) the support of P_{te} is contained in the support of P_{tr} (i.e., the testing set is encompassed or covered by the training set). Regarding the first assumption, nonparametric approaches have been proposed too, where the weights $\frac{P_{\text{te}}(\mathbf{x})}{P_{\text{tr}}(\mathbf{x})}$ can be obtained indirectly, without explicit estimation of the training and testing set densities [13]. The second assumption places a severe restriction on the applicability of the weighting scheme; when it does not hold, different techniques have been proposed to deal with concepts that change over time, most notably under the umbrella of concept drift [14], or concept evolution [15]. Our work avoids the difficulty linked to a change in the class posterior probabilities through the alignment step, by reinstating the original model built on T_{tr}.

5. Experiments

Our analysis uses data obtained by two surveys: (1) the third phase of the Optical Gravitational Lensing Experiment (OGLE-III)[4] and (2) the M33 Synoptic Stellar Survey[5]. OGLE [16] is a collaboration of mostly Polish astronomers originally designed to search for dark matter using gravitational microlensing. The project has been observing the Magellanic Clouds and the Galactic Bulge using a dedicated 1.3-m telescope at Las Campanas Observatory in Chile. As a side product of the search for microlensing events, OGLE has discovered tens of thousands of variable stars, including Cepheids. We analyze the third catalog of OGLE Large Magellanic Cloud Cepheids [17], on both the visible (V) and infrared (I) bands; the training set for the visible band contains 3,059 points (i.e., light curves), while the training set for the infrared band contains 3,056 points. Data is of excellent quality, with statistical uncertainties of 0.001% and 0.00035% on each band respectively. The approximate detection thresholds of these observations are $V = 21.5$ and $I = 21$ mag, well below the faintest apparent magnitude of any Cepheid

[4]Data repository available at http://ogledb.astrouw.edu.pl/ ogle/CVS/

[5]Data repository available at http://faculty.physics.tamu.edu/lmacri/M33SSS/

Table 1: Classification Performance under Visible Band. Numbers in parentheses represent standard deviations. Last two columns correspond to proposed methodology.

Learning Algorithm	Accuracy on LMC with class labels	Accuracy on M33 (target domain)			
		with class labels	with no class labels using data alignment		
			Standard	Step 1	Step 2
Neural Networks	94.77 (1.18)	96.18 (0.16)	90.09 (1.33)	96.78 (0.90)	95.72 (0.98)
SVM Polynomial 1	94.48 (1.19)	92.55 (0.26)	95.30 (0.28)	95.76 (0.88)	97.41 (1.10)
SVM Polynomial 2	94.59 (1.23)	95.28 (0.19)	93.69 (0.61)	95.98 (0.80)	95.39 (0.88)
SVM Polynomial 3	94.62 (1.21)	95.08 (0.20)	93.38 (0.66)	95.98 (0.80)	97.09 (0.88)
Decision Trees	94.32 (1.27)	97.61 (0.13)	94.49 (0.52)	97.27 (0.75)	97.09 (0.80)
Random Forest	93.80 (1.26)	97.78 (0.13)	94.12 (0.41)	96.13 (0.78)	96.40 (0.88)

in the Large Magellanic Cloud. The M33 Synoptic Stellar Survey [18] is a follow-up study of Cepheids and other variables initially discovered by the DIRECT project [19, 20]. These testing sets contain 323 points in the visible band, and 322 points in the infrared band (statistical uncertainties of 0.0125% and 0.0056% respectively). The approximate detection thresholds of these observations are $V = 23$ and $I = 21.5$ mag, which are adequate to detect all fundamental mode Cepheids with $P > 3$ days and first-overtone Cepheids with $P > 2$ days.

We refer to data from the Large Magellanic Cloud galaxy as the source domain, and to data from M33 galaxy as the target domain. As a pre-processing step, parameters for the mixture of Gaussians on the source domain are estimated directly through the data. Since class labels are available, it is straightforward to estimate means, covariance matrices and mixing proportions. When narrowing the training set T_{tr} (source domain) to those examples where the period is within the range of values found in the target domain (section 3.4), the size of the set decreases to $2,313$ points. Implementation of the learning algorithms can be found in WEKA [21] using default parameters. Specifically, neural networks are invoked with one hidden layer, two internal nodes, a learning rate of 0.3, momentum of 0.2, and 500 epochs. Support Vector Machines are invoked with polynomial kernels of degrees 1, 2, and 3 respectively. Decision trees are invoked with a confidence factor of 0.25. Random Forests combine 10 decision trees on each run. Regarding the optimization of equation 8, we solve for δ by invoking the "solve function" in the "symbolic math toolbox" in Matlab.

Our first set of experiments compute predictive accuracy when the model obtained on LMC is applied to the transformed target domain (when a shift δ is applied to apparent magnitude). Accuracy is obtained by creating bootstrap samples of the target or test set (M33); each sample is generated by sampling with replacement. We generated 10 samples, and averaged our results after applying model $f(\mathbf{x}|\theta)$ on each transformed sample. Table 1 shows our results when the data refers to the visible band. The first column corresponds to the learning algorithms. The second column corresponds to accuracy on the source domain (as reference only). The third column shows accuracy on the target domain when class labels are available. The next columns show accuracy on M33 when class labels are unavailable. The fourth column (labeled "Standard") shows results using the values of δ listed in Table 2 of [18]. We take into consideration that the

Table 2: Classification Performance under Infrared Band. Numbers in parentheses represent standard deviations. Last two columns correspond to proposed methodology.

Learning Algorithm	Accuracy on LMC with class labels	Accuracy on M33 (target domain)			
		with class labels	with no class labels using data alignment		
			Standard	Step 1	Step 2
Neural Networks	96.89 (0.16)	97.05 (0.14)	94.25 (2.45)	96.99 (0.81)	94.32 (1.68)
SVM Polynomial 1	97.00 (0.17)	93.18 (0.26)	96.80 (0.15)	95.50 (1.11)	96.32 (1.11)
SVM Polynomial 2	96.98 (0.17)	93.18 (0.26)	94.97 (0.58)	96.83 (0.79)	95.53 (1.11)
SVM Polynomial 3	97.02 (0.17)	93.18 (0.26)	95.15 (0.82)	96.71 (0.78)	95.03 (0.97)
Decision Trees	96.35 (0.17)	96.27 (0.17)	95.77 (0.49)	97.62 (1.08)	96.15 (1.28)
Random Forest	96.47 (0.17)	95.78 (0.16)	95.62 (0.71)	96.86 (0.84)	97.26 (0.56)

apparent magnitudes of the OGLE LMC Cepheids are not corrected for the average amount of interstellar dust towards this galaxy, while [18] applies a correction for dust[6]. The fifth column (labeled "Step 1") shows results when δ is determined using maximum likelihood based on density estimation (Section 3.2), while the sixth column (labeled "Step 2") shows results when we exploit the induced class distribution (Section 3.3, $\tau = 0.01$).

Compared to previous work [18], we observe a significant increase in accuracy when δ is computed using either step 1 or 2; this is true for all learning algorithms under consideration, which validates our proposed methodology. Furthermore, results for both step 1 and 2 show accuracies similar to those obtained on M33 when class labels are present, evidencing how an appropriate shift in magnitude renders the model trained on the source domain valid under the target domain. The difference between steps 1 and 2 is mixed; in two cases step 2 outperforms step 1, while in two cases the opposite is true, and in two other cases the difference is not significant ($p = 0.05$ level using a t-student distribution). Further investigation, however, shows some clear benefits with step 2: the range of values along δ for which a maximum in the likelihood function is obtained is consistently narrower for step 2, which helps us pinpoint more precisely our final estimation for δ ($\delta_{\text{visible}-\text{final}} = -6.24$).

Table 2 displays results similar to the ones describe above, but using the infrared band. Compared to previous work, we observe a general improvement in accuracy when δ is computed using maximum likelihood, but the improvement is less clear. The difference between steps 1 and 2 is also similar. Our final estimation for δ ($\delta_{\text{infrared}-\text{final}} = -6.08$) is different from the one obtained using the visible band, because the effects of reddening due to interstellar dust are significantly reduced at infrared waves [22]. Overall, empirical results support our proposed methodology. In addition, we provide an alternative means to compute the correction shift δ under a new approach based on maximum likelihood and sample bias adjustment. Our values are lower than those derived by [18], but statistically consistent, given the scatter in their Period-Luminosity relations (0.4 mag in V and 0.27 mag in I).

[6]For reference, the adjustment values derived in [18] are $\delta = 6.45 \pm 0.02$ for the visible band, and 6.31 ± 0.02 for the infrared band.

6. Summary and Conclusions

Our study shows an approach to deal with the automated prediction of Cepheid stars obtained from a target galaxy, after a predictive model has already been built on a source galaxy. Our methodology is based on manipulating the target data by shifting along magnitude to correct for uncertainty in luminosity and sample bias. Our first form of alignment uses maximum likelihood based on density estimation on the source domain. The second form of alignment refines the previous one by exploiting the class distribution over the source domain as predicted by the learning model. Our experiments show accuracy values similar to those obtained if class labels were available on the target galaxy.

Our approach is not limited to the star classification problem. The alignment problem can be found on other domains where there is a shift in data distribution due to a systematic change during data collection, caused mainly by a change in the physical conditions surrounding an experiment. In particle physics, for example, a model built to identify a particle may by re-used on a new sample obtained through a more sophisticated accelerator, by simply shifting the data along the set of parameters where reaching out to higher energies brings with it a displacement in the data distribution. Our study suggests, however, that the success of dataset shift strategies depends on the global properties of the model built over the source distribution. Specifically, if both marginal distributions differ, $P_{\text{tr}}(\mathbf{x}) \neq P_{\text{te}}(\mathbf{x})$, but the class posteriors remain identical, $P_{\text{tr}}(w|\mathbf{x}) = P_{\text{te}}(w|\mathbf{x})$, then in general the bias produced by training on the source domain is counterbalanced by a distribution exhibiting global patterns (i.e., exhibiting no local irregularities). In our case, this corresponds to the P-L linear relationship (section 2.1), which is invariant across galaxies. When that is the case, model $f(\mathbf{x}|\theta)$ can be re-used on different regions of the input space.

As future work, we plan to extend our ideas to the context of transfer learning [23]. Dataset shift resembles transfer learning, but while the former deals with changes on the same task, the latter comprises a broader setting where target tasks can differ substantially from the source task [24]. We plan to explore when a model can be re-used, based on the presence of invariant global data properties across tasks.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable suggestions. Lucas Macri acknowledges support for this project by the following sources: NASA, through Hubble Fellowship grant HST-HF-01153 from the Space Telescope Science Institute; the National Science Foundation, through a Goldberg Fellowship from the National Optical Astronomy Observatory and through grant number 1211603; and by Texas A&M University, through a faculty start-up fund.

References

[1] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset Shift in Machine Learning, MIT Press, 2009.

[2] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, Journal of Statistical Planning and Inference 90 (2000) 227–244.

[3] M. Sugiyama, N. Rubens, K. R. Muller, A conditional expectation approach to model selection and active learning under covariate shift, in: J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009, pp. 107–130.

[4] J. P. Cox, Theory of Stellar Pulsation, Princeton University Press, 1980.

[5] S. E. Persson, B. F. Madore, W. KrzemiÅski, W. L. Freedman, M. Roth, D. C. Murphy, New cepheid period-luminosity relations for the large magellanic cloud: 92 near-infrared light curves, The Astronomical Journal 128 (5) (2004) 2239–2264.
[6] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, M. Rischard, H. Shimodaira, On machine-learned classification of variable stars with sparse and noisy time-series data, The Astrophysical Journal 733:10.
[7] J. Debosscher, L. M. Sarro, C. Aerts, J. Cuypers, B. Vandenbussche, R. Garrido, E. Solano, Automated supervised classification of variable stars, Astronomy & Astrophysics 475 (2007) 1159–1183.
[8] T. Kanamori, H. Shimodaira, Geometry of covariate shift with applications to active learning, in: J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009, pp. 87–105.
[9] S. Bickel, M. Bruckner, T. Scheffer, Discriminative learning under covariate shift, Journal of Machine Learning Research 10 (2009) 2137–2155.
[10] M. Sugiyama, K. R. Muller, Model selection under covariate shift, in: Proceedings of the International Conference on Artificial Neural Networks, 2005, pp. 235–240.
[11] Y. Lin, Y. Lee, G. Wahba, Support vector machines for classification in nonstandard situations, Machine Learning 46 (2002) 191–202.
[12] B. Zadrozny, Learning and evaluating classifiers under sample selection bias, in: Proceedings of the International Conference on Machine Learning, 2004, pp. 114–122.
[13] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Scholkopf, Covariate shift by kernel mean matching, in: J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009, pp. 131–160.
[14] S. H. Bach, M. A. Maloof, A bayesian approach to concept drift, in: Proceedings of the International Conference on Neural Information Processing Systems, 2010, pp. 127–135.
[15] M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, B. Thuraisingham, Addressing concept-evolution in concept-drifting data streams, in: Proceedings of the IEEE International Conference on Data Mining, 2010, pp. 929–934.
[16] A. Udalski, The optical gravitational lensing experiment (ogle): Bohdan's and our great adventure, in: K. Z. Stanek (Ed.), The Variable Universe: A Celebration of Bohdan Paczynski, Vol. 403 of Astronomical Society of the Pacific Conference Series, 2009, p. 110.
[17] I. Soszynski, R. Poleski, A. Udalski, M. K. Szymanski, M. Kubiak, G. Pietrzynski, L. Wyrzykowski, O. Szewczyk, K. Ulaczyk, The optical gravitational lensing experiment. the ogle-iii catalog of variable stars. i. classical cepheids in the large magellanic cloud, Acta Astronomica 58 (2008) 163–185.
[18] A. Pellerin, L. M. Macri, The m33 synoptic stellar survey. i. cepheid variables, Astrophysical Journal Supplement Series 193 (2011) 26.
[19] K. Z. Stanek, J. Kaluzny, M. Krockenberger, D. D. Sasselov, J. L. Tonry, M. Mateo, Direct distances to nearby galaxies using detached eclipsing binaries and cepheids. ii. variables in the field m31a, Astronomical Journal 115 (1998) 1894–1915.
[20] L. M. Macri, K. Z. Stanek, D. D. Sasselov, M. Krockenberger, J. Kaluzny, Direct distances to nearby galaxies using detached eclipsing binaries and cepheids. vi. variables in the central part of m33, Astronomical Journal 121 (2001) 870–890.
[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: An update, SIGKDD Explorations 11 (1).
[22] B. Madore, W. L. Freedman, The cepheid distance scale, Publications of the Asronomical Society of the Pacific 103 (1991) 933–957.
[23] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.
[24] A. Storkey, When training and test sets are different, in: J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence (Eds.), Dataset Shift in Machine Learning, MIT Press, 2009, pp. 3–28.

Appendix

We provide a detailed derivation of the maximum likelihood approach presented in Section 3.2. The joint distribution of the data is expressed as follows:

$$P_{\text{tr}}(x_1^k, x_2^k) = \sum_{i=1}^{2} \phi_i \alpha_i e^{-\beta_i z_i^k}, \quad \text{where} \quad \alpha_i = \frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-\rho_i^2}} \text{ and } \beta_i = \frac{1}{2(1-\rho_i^2)}$$

where

$$e^{-\beta_i z_i^k} = e^{c_{i1}^k + \delta c_{i2}^k + \delta^2 c_{i3}^k}$$

$$c_{i1}^k = -\beta_i \left(\frac{(x_1^k - \mu_{i1})^2}{\sigma_{i1}^2} - \frac{2\rho_i (x_1^k - \mu_{i1})(x_2^k - \mu_{i2})}{\sigma_{i1}\sigma_{i2}} + \frac{(x_2^k - \mu_{i2})^2}{\sigma_{i2}^2} \right)$$

$$c_{i2}^k = -\beta_i \left(\frac{2(x_2^k - \mu_{i2})}{\sigma_{i2}^2} - \frac{2\rho_i (x_1^k - \mu_{i1})}{\sigma_{i1}\sigma_{i2}} \right) \quad \text{and} \quad c_{i3}^k = \frac{-\beta_i}{\sigma_{i2}^2}$$

The derivative of the log likelihood function can be written as this:

$$\frac{d}{d\delta}\left(\sum_{k=1}^{q} \log P_{\text{tr}}(x_1^k, x_2^k) \right) = \frac{d}{d\delta}\left(\sum_{k=1}^{q} \log \sum_{i=1}^{2} \phi_i \alpha_i e^{-\beta_i z_i^k} \right) = \frac{d}{d\delta}\left(\sum_{k=1}^{q} \log(\phi_1 \alpha_1 e^{-\beta_1 z_1^k} + \phi_2 \alpha_2 e^{-\beta_2 z_2^k}) \right)$$

$$= \sum_{k=1}^{q} \left(\frac{d}{d\delta}\left(\log(\phi_1 \alpha_1 e^{-\beta_1 z_1^k} + \phi_2 \alpha_2 e^{-\beta_2 z_2^k}) \right) \right) = \sum_{k=1}^{q} \left(\frac{1}{\phi_1 \alpha_1 e^{-\beta_1 z_1^k} + \phi_2 \alpha_2 e^{-\beta_2 z_2^k}} \left(\frac{d}{d\delta}(\phi_1 \alpha_1 e^{-\beta_1 z_1^k} + \phi_2 \alpha_2 e^{-\beta_2 z_2^k}) \right) \right)$$

Now,

$$\frac{d}{d\delta}\left(\phi_1 \alpha_1 e^{-\beta_1 z_1^k} + \phi_2 \alpha_2 e^{-\beta_2 z_2^k} \right) = \frac{d}{d\delta}\left(\phi_1 \alpha_1 (e^{c_{11}^k + \delta c_{12}^k + \delta^2 c_{13}^k}) + \phi_2 \alpha_2 (e^{c_{21}^k + \delta c_{22}^k + \delta^2 c_{23}^k}) \right)$$

$$= (c_{12}^k + 2\delta c_{13}^k)\phi_1 \alpha_1 (e^{c_{11}^k + \delta c_{12}^k + \delta^2 c_{13}^k}) + (c_{22}^k + 2\delta c_{23}^k)\phi_2 \alpha_2 (e^{c_{21}^k + \delta c_{22}^k + \delta^2 c_{23}^k})$$

$$= \phi_1 \alpha_1 c_{12}^k e^{c_{11}^k} (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + 2\delta c_{13}^k \phi_1 \alpha_1 e^{c_{11}^k} (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + \phi_2 \alpha_2 c_{22}^k e^{c_{21}^k} (e^{\delta c_{22}^k + \delta^2 c_{23}^k}) + 2\delta c_{23}^k \phi_2 \alpha_2 e^{c_{21}^k} (e^{\delta c_{22}^k + \delta^2 c_{23}^k})$$

$$= \left(\phi_1 \alpha_1 c_{12}^k e^{c_{11}^k} (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + \phi_2 \alpha_2 c_{22}^k e^{c_{21}^k} (e^{\delta c_{22}^k + \delta^2 c_{23}^k}) \right) + \delta \left(2c_{13}^k \phi_1 \alpha_1 e^{c_{11}^k} (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + 2c_{23}^k \phi_2 \alpha_2 e^{c_{21}^k} (e^{\delta c_{22}^k + \delta^2 c_{23}^k}) \right)$$

$$= \left(\gamma_{11}^k (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + \gamma_{21}^k (e^{\delta c_{22}^k + \delta^2 c_{23}^k}) \right) + \delta \left(\gamma_{12}^k (e^{\delta c_{12}^k + \delta^2 c_{13}^k}) + \gamma_{22}^k (e^{\delta c_{22}^k + \delta^2 c_{23}^k}) \right)$$

where

$$\gamma_{11}^k = \phi_1 \alpha_1 c_{12}^k e^{c_{11}^k}, \quad \gamma_{21}^k = \phi_2 \alpha_2 c_{22}^k e^{c_{21}^k} \quad \gamma_{12}^k = 2c_{13}^k \phi_1 \alpha_1 e^{c_{11}^k}, \text{ and } \gamma_{22}^k = 2c_{23}^k \phi_2 \alpha_2 e^{c_{21}^k}$$

Therefore,

$$\frac{d}{d\delta}\left(\sum_{k=1}^{q}\log P_{\text{tr}}(x_1^k, x_2^k)\right) = \sum_{k=1}^{q}\frac{\left(\gamma_{11}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{21}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right) + \delta\left(\gamma_{12}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{22}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right)}{\phi_1\alpha_1 e^{-\beta_1 z_1^k} + \phi_2\alpha_2 e^{-\beta_2 z_2^k}}$$

$$= \sum_{k=1}^{q}\frac{\left(\gamma_{11}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{21}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right) + \delta\left(\gamma_{12}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{22}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right)}{\phi_1\alpha_1(e^{c_{11}^k+\delta c_{12}^k+\delta^2 c_{13}^k}) + \phi_2\alpha_2(e^{c_{21}^k+\delta c_{22}^k+\delta^2 c_{23}^k})}$$

$$= \sum_{k=1}^{q}\frac{\left(\gamma_{11}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{21}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right) + \delta\left(\gamma_{12}^k(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \gamma_{22}^k(e^{\delta c_{22}^k+\delta^2 c_{23}^k})\right)}{\phi_1\alpha_1 e^{c_{11}^k}(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \phi_2\alpha_2 e^{c_{21}^k}(e^{\delta c_{22}^k+\delta^2 c_{23}^k})}$$

$$= \sum_{k=1}^{q}\left(\frac{\gamma_{11}^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \gamma_{21}^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}}{\phi_1\alpha_1 e^{c_{11}^k}(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \phi_2\alpha_2 e^{c_{21}^k}(e^{\delta c_{22}^k+\delta^2 c_{23}^k})} + \delta\frac{\gamma_{12}^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \gamma_{22}^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}}{\phi_1\alpha_1 e^{c_{11}^k}(e^{\delta c_{12}^k+\delta^2 c_{13}^k}) + \phi_2\alpha_2 e^{c_{21}^k}(e^{\delta c_{22}^k+\delta^2 c_{23}^k})}\right)$$

$$= \sum_{k=1}^{q}\left(\frac{\gamma_{11}^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \gamma_{21}^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}}{\Gamma_1^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \Gamma_2^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}} + \delta\frac{\gamma_{12}^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \gamma_{22}^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}}{\Gamma_1^k e^{\delta c_{12}^k+\delta^2 c_{13}^k} + \Gamma_2^k e^{\delta c_{22}^k+\delta^2 c_{23}^k}}\right) \text{ where } \Gamma_1^k = \phi_1\alpha_1 e^{c_{11}^k} \text{ and } \Gamma_2^k = \phi_2\alpha_2 e^{c_{21}^k}$$

$$= \sum_{k=1}^{q}\left(\frac{\gamma_{11}^k e^{\theta_1^k(\delta)} + \gamma_{21}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}} + \delta\frac{\gamma_{12}^k e^{\theta_1^k(\delta)} + \gamma_{22}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}}\right) \text{ where } \theta_1^k(\delta) = \delta c_{12}^k + \delta^2 c_{13}^k \text{ and } \theta_2^k(\delta) = \delta c_{22}^k + \delta^2 c_{23}^k$$

Finally,

$$\frac{d}{d\delta}\left(\sum_{k=1}^{q}\log P_{\text{tr}}(x_1^k, x_2^k)\right) = \sum_{k=1}^{q}\left(\frac{\gamma_{11}^k e^{\theta_1^k(\delta)} + \gamma_{21}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}} + \delta\frac{\gamma_{12}^k e^{\theta_1^k(\delta)} + \gamma_{22}^k e^{\theta_2^k(\delta)}}{\Gamma_1^k e^{\theta_1^k(\delta)} + \Gamma_2^k e^{\theta_2^k(\delta)}}\right) = 0$$