

Classification of Microorganisms via Raman Spectroscopy Using Gaussian Processes

Michael Kemmler¹, Joachim Denzler¹, Petra Rösch², and Jürgen Popp²

¹Chair for Computer Vision

²Institute of Physical Chemistry
Friedrich Schiller University of Jena

Abstract. Automatic categorization of microorganisms is a complex task which requires advanced techniques to achieve accurate performance. In this paper, we aim at identifying microorganisms based on Raman spectroscopy. Empirical studies over the last years show that powerful machine learning methods such as Support Vector Machines (SVMs) are suitable for this task. Our work focuses on the Gaussian process (GP) classifier which is new to this field, provides fully probabilistic outputs and allows for efficient hyperparameter optimization. We also investigate the incorporation of prior knowledge regarding possible signal variations where known concepts from invariant kernel theory are transferred to the GP framework. In order to validate the suitability of the GP classifier, a comparison with state-of-the-art learners is conducted on a large-scale Raman spectra dataset, showing that the GP classifier significantly outperforms all other tested classifiers including SVM. Our results further show that incorporating prior knowledge leads to a significant performance gain when small amounts of training data are used.

1 Introduction

In the fast-growing field of medical and biological science, the need for classifying microorganisms is rapidly increasing. There are many crucial tasks which demand an accurate classification method, such as the categorization of potentially pathogenic particles in clinical applications [1] or the identification of contamination conditions in clean room environments [2], to name just a few. The assortment of tools available for identifying microbes is broad, ranging from microscopic inspection [3] to advanced biochemical analysis [4]. Though, while the classification based on microscopic means using morphological information is only possible on a coarse level, most accurate biochemical methods require time consuming pre-processing steps for cultivating the media of interest. So-called "vibrational techniques" such as Raman spectroscopy offer an elegant way out of this dilemma by obtaining a "molecular fingerprint" of biological samples [2].

This work aims at introducing the Gaussian Process classifier to the field of vibrational spectroscopy. Although GP regression is well-known for the calibration of spectroscopic data [5], GP classification is, to our knowledge, new to the field of Raman spectroscopy. Moreover, we investigate the use of a-priori

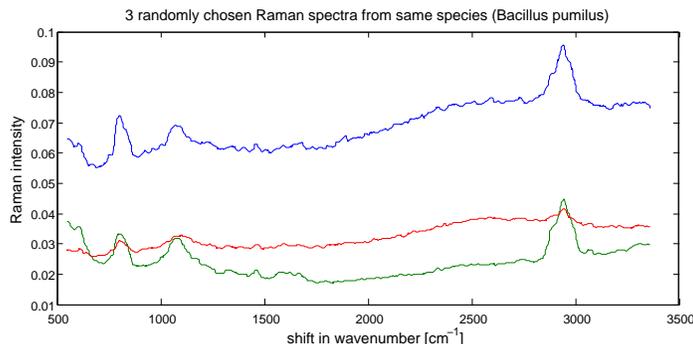


Fig. 1. Illustration of (pre-processed) Raman spectra and signal variations.

knowledge by means of possible variations of Raman spectra, where we apply the approach of Peschke et al. [6] to the GP framework. Shortcomings of this approach as well as heuristics to overcome those issues are discussed.

This paper is organized as follows. In Sect. 2 we give a brief summary to Raman spectroscopy. Sect. 3 gives an introduction to GP classification. In Sect. 4 we formulate the notion of transformation invariant kernels and show how a-priori knowledge can be incorporated in the GP framework. In Sect. 5 we empirically demonstrate the performance of the GP classifier for Raman spectra categorization compared to other state-of-the-art classification techniques and investigate the effect of prior knowledge. Finally, we conclude and summarize our findings.

2 Classification of Raman Spectra

Raman spectroscopy is an optical technique for measuring the vibration of molecules. The sample under focus is irradiated with a narrow-band LASER and the scattered light is analyzed. Since shifts in wavenumber are strongly related to the vibrational state of molecules and shifts from all molecules are superimposed, we obtain a molecular fingerprint of the whole sample (c.f. Fig. 1). Since most microorganisms substantially differ in molecular decomposition it is assumed that they can be distinguished by means of their vibrational signature. Empirical results validate [2] that Raman spectra contain discriminative information and are suitable for categorization of even very similar microorganisms.

However, one major drawback is the occurrence of prominent background signals introduced by other optical phenomena such as fluorescence. One common technique to overcome this issue is to explicitly pre-process all spectra prior to training and learning. Alternatively, Peschke et al. [6] successfully apply an invariant kernel approach which implicitly incorporates possible background variations into the kernel for SVM and k -Nearest-Neighbor classification. In addition to assessing the suitability of the GP classifier to the field of Raman spectroscopy, this work aims at transferring the approach of [6] to the GP framework.

3 Gaussian Process Classification

This section gives a brief introduction to GP classification. Since the classification is motivated from non-parametric Bayesian regression, we first briefly introduce the regression case before we discuss the GP classifier.

3.1 The Regression Case

The regression problem aims at finding a mapping from input space \mathcal{X} to output space \mathbb{R} using labeled training data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathcal{X}^n$, $\mathbf{y} = [y_1, \dots, y_n]^T \in \mathbb{R}^n$. In the following it is assumed that the output is generated by a latent (non-observed) function $f : \mathcal{X} \rightarrow \mathbb{R}$ and an additive noise term ε , i.e. $y = f(\mathbf{x}) + \varepsilon$. Rather than restricting f to a certain functional family, we only assume that the function is drawn from a specific probability distribution $p(\mathbf{f}|\mathbf{X})$. This allows for a Bayesian treatment of our problem, i.e. we infer the probability of outputs y_* given new inputs x_* and old observations \mathbf{X}, \mathbf{y} by integrating out the non-observed function values f_* and \mathbf{f} :

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) p(y_*|f_*) df_* \quad (1)$$

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{x}_*) p(\mathbf{f}|\mathbf{X}, \mathbf{y}) d\mathbf{f} . \quad (2)$$

The central assumption in GP regression is that all function values are jointly normally distributed, i.e. $\mathbf{f}|\mathbf{X} \sim \mathcal{N}(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X}))$. This distribution is solely specified by the mean function $m(\cdot)$ and covariance function $\kappa(\cdot, \cdot)$. When we additionally assume that the data is generated by zero mean independent Gaussian noise, i.e. $y \sim \mathcal{N}(f, \sigma_n^2)$, then we are able to solve the integrals in closed form. Using a zero mean GP, the predictive distribution (2) is again Gaussian with predictive mean $\mu_* = \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}$ and predictive variance $\sigma_*^2 = k_{**} - \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*$ using shortcuts $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{k}_* = \kappa(\mathbf{X}, \mathbf{x}_*)$, and $k_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$ and hence also implies that (1) is Gaussian.

3.2 From Regression to Classification

The goal in binary GP classification is to model a function which predicts a confidence for each class $y \in \{-1, 1\}$, given a feature vector \mathbf{x} . In order to make probabilistic inference about the output given a training set, we can directly apply the Bayesian formalism from equation (1) and (2). However, the key problem is that the assumption of Gaussian noise no longer holds since the output space is discrete. We could either ignore this issue and perform regression on our labels or we could use a more appropriate assumption of $p(y|f)$. Here we follow the latter approach using the cumulative Gaussian likelihood $p(y|f) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{yf} \exp(-0.5x^2) dx$. The disadvantage of this procedure is that our predictive distribution (2) no longer is a normal distribution. To overcome this issue, we follow the standard approach to approximate [7] the posterior

$p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ with a normal distribution \hat{p} . We use Laplace’s Method, i.e. the mean of \hat{p} is set equal to the mode of $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ and the Hessian of $-\log p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ is utilized as covariance matrix of \hat{p} . Moreover, by using the cumulative Gaussian as described above, equation (1) also turns out to be a cumulative Gaussian [8] and hence can be efficiently computed. Finally, appropriate hyperparameters θ^* of the covariance function can be efficiently found by marginal likelihood maximization, i.e. $\theta^* = \operatorname{argmax}_{\theta} p(\mathbf{y}|\mathbf{X}, \theta)$.

4 Prior Knowledge in GP Classification

As has been pointed out in Sect. 2, Raman spectra often contain undesirable information introduced by fluorescence or different measuring conditions. The resulting signal variations often pose a problem for most classifiers. This is especially the case in real scenarios where only small amounts of training data are available, since the possible variations are not sufficiently covered by the training set. One way to overcome this issue is to implicitly embed knowledge about possible signal variations into the learning algorithm. In this section we discuss how and to which extent this kind of prior information can be incorporated into the GP framework by means of Tangent Distance Substitution Kernels [9].

4.1 Tangent Distance Substitution Kernels

Many methods in machine learning are solely expressed via symmetric similarity functions κ , so-called kernels. These kernels $\kappa(\mathbf{x}, \mathbf{x}')$ are often positive definite in which case they can be interpreted as inner products $\kappa(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^T \Phi(\mathbf{x}')$ in some feature space \mathcal{Y} induced by a mapping $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$. Many kernels used throughout the literature are also expressed in terms of Euclidean distances d . One possibility to introduce some degree of invariance into kernels is to replace Euclidean distances with distances that are invariant with respect to pattern variations. When specified beforehand, knowledge about expected pattern variations is thus some kind of prior information which can be incorporated into distances. In the following we use the regularized Mean Tangent Distance [9] TDMN_{γ} which results in a locally invariant distance measure:

$$\text{TDMN}_{\gamma}^2(\mathbf{x}, \mathbf{x}') = \frac{1}{2} (\text{TD1S}_{\gamma}^2(\mathbf{x}, \mathbf{x}') + \text{TD1S}_{\gamma}^2(\mathbf{x}', \mathbf{x})) \quad (3)$$

$$\text{TD1S}_{\gamma}^2(\mathbf{x}, \mathbf{x}') = \min_{\mathbf{p}} \|\mathbf{x} + \mathbf{T}_{\mathbf{x}}\mathbf{p} - \mathbf{x}'\|_2^2 + \gamma\|\mathbf{p}\|_2^2, \quad (4)$$

where TD1S_{γ} denotes the one-sided Tangent distance using regularization parameter γ . The latter Tangent distance computes the distance between \mathbf{x}' and the first order approximation $\mathbf{T}_{\mathbf{x}}$ of the variation manifold in which \mathbf{x} resides (see Fig. 2). As detailed in [9], Tangent distances can be used in any possible distance-based kernel. In this paper, however, we will solely focus on the squared exponential (SE) kernel [8] with respect to some distance d :

$$\kappa_{\nu_1, \nu_2}(\mathbf{x}, \mathbf{x}') = \nu_1^2 \exp\left(-\frac{d^2(\mathbf{x}, \mathbf{x}')}{2\nu_2^2}\right). \quad (5)$$

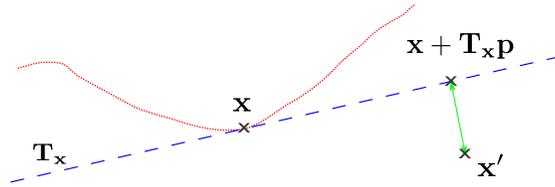


Fig. 2. Visualization of Tangent distance. Dotted red: Pattern variation manifold with respect to \mathbf{x} . Dashed blue: Tangent $\mathbf{T}_{\mathbf{x}}$ of variation manifold with respect to \mathbf{x} . Solid green: Unregularized one-sided Tangent distance $\text{TD1S}_0(\mathbf{x}, \mathbf{x}')$.

4.2 Invariant Kernels in GP Classification

GP methods can be related to kernel machines through their covariance function κ which describes the underlying similarity structure. One can therefore utilize invariant kernels such as (5) with $d = \text{TDMN}_{\gamma}$ as a covariance function. By doing so, however, the theoretical assumptions behind the GP framework might be violated. This is due to the fact that the Tangent distance is not necessarily a metric distance. It can be shown [9] that in this case the above kernel is not positive definite and hence no valid covariance function for a GP.

To overcome this issue one could *explicitly enforce positive definiteness*. This is possible, e.g. by clipping off negative eigenvalues (CLIP), flipping negative values to its positive absolute value (ABS) or by adding a constant $c \geq |\lambda_{\min}|$ to all eigenvalues (SHIFT), where λ_{\min} is the smallest algebraic eigenvalue of the kernel matrix. While the latter technique can be efficiently realized by Krylov subspace methods such as Lanczos algorithms, CLIP and ABS require a full eigendecomposition of the covariance matrix of the entire data.

To guarantee positive definiteness, one could also *reduce the set of possible tangents*. It can be shown [9] that by using tangents which are constant, i.e. independent of the input arguments \mathbf{x} , the Tangent distance is equal to a Mahalanobis distance which leads to positive (semi-)definite SE-kernels.

We could also ignore the indefiniteness of the covariance function. One result would be that variances might get negative. In order to construct a Gaussian distribution in the end, one could use heuristics to ensure positivity of the variance term, e.g. by setting negative values to zero (*negative variance cut-off heuristic*).

Although the Bayesian perspective gets lost using the latter approach, we can take a different view on the outcome. In GP regression, the moments μ_* and σ_*^2 of predictive Gaussian $p(f_* | \mathbf{X}, \mathbf{f}, \mathbf{x}_*)$ can be also derived from linear estimation [10] in some indefinite inner product (Kreĭn) space \mathcal{K} , where the predictive mean $\mu_* = \mathbf{w}_*^T \mathbf{y}$ is some linear projection onto the training set \mathbf{y} (which is equal to the linear Least Squares estimate if the data can be exactly explained by our prior belief κ , i.e. $\mathbb{E}\mathbf{y}\mathbf{y}^T = \mathbf{K} + \sigma_n^2 \mathbf{I}$), and σ_*^2 is equal to the reconstruction error for \mathbf{w}_* . Using the Laplace Approximation, a similar relationship holds using a slightly modified set of latent function values. From this perspective, cutting off negative variances means that we are confident ($\sigma_*^2 = 0$) for estimates μ_* in some regions of \mathcal{K} that will result in negative goodness-of-fit values σ_*^2 .

5 Experiments and Results

In the following experiments we used a large Raman dataset consisting of 6707 spectra comprising 4 different genera (classes) and 10 species (sub-classes) which was captured over a period of 3 years. Due to the physics of the recording process, each spectrum exhibits a potentially different set of wavenumber shifts. We therefore crop the interesting region to the interval $\mathcal{I} = [540\text{cm}^{-1}, 3350\text{cm}^{-1}]$ that is shared among all samples. In order to work with fixed dimensions, we use quadratic interpolation to obtain the Raman intensity with respect to all integer wavenumbers $\omega \in \mathcal{I} \cap \mathbb{Z}$. In order to eliminate spike noise artifacts introduced by cosmic radiation, a running median (of size 21) is applied. For the sake of numerical stability, all spectra are either normalized to unit length or multiplied by a fixed constant $c = 8.5741 \times 10^{-5}$. In all experiments we used the (noise-free) GP classifier from Sect. 3.2 with kernel (5) whose parameters are tuned with marginal likelihood maximization (starting at $\boldsymbol{\nu} = (1, 1)^T$, 10 iterations). To allow for multiple classes, a one-vs-all scheme based on predictive probabilities (1) is employed.

In this section, we will empirically validate the following hypotheses:

1. The GP classifier is suitable for classification of Raman spectra (Sect. 5.1).
2. For small training sets, a significant performance gain is achieved by incorporating prior information (Sect. 5.2).
3. A special treatment of indefinite kernels is not necessary for our task (Sect. 5.2).

5.1 Suitability for Raman Spectroscopic Categorization

In order to investigate whether the GP classifier is suitable for the categorization of Raman spectra, we compared it to four state-of-the-art classifiers: k -Nearest-Neighbor classifier (KNN), Randomized Decision Forests (RDF) and AdaBoost.MH [11] (ADA) with decision stumps, as well as Support Vector Machines [12] (SVM), where for the latter logistic regression is used to generate pseudo-probabilistic outputs. Particularly, we used KNN with $k = 1$ (best performance for this value of k), a gently randomized RDF (95% resampling probability, 250 random features per node) with 100 trees, and a SVM with rbf-kernel whose parameters were determined via leave-one-out (LOO) estimates on a 2×10 grid for trade-off parameter and bandwidth parameter.

It should be noted that results concerning the novel large dataset used in this work are not directly comparable to other work (e.g. [2]) since other datasets strongly differ in size and complexity and are generally not publicly available. An indirect comparison, however, is possible since we compare against SVM which is known to achieve very good results in the field of Raman spectroscopy [2]. Moreover, since fast LOO error estimates cannot be computed for all classifiers, we randomly chose 75% of the data for training and the remainder for testing. In order to yield a more robust performance estimate we repeated this procedure ten times.

The resulting ten average recognition rates for each classifier are illustrated in Fig. 3 where results for the genus level (4 classes) and species level (10 classes)

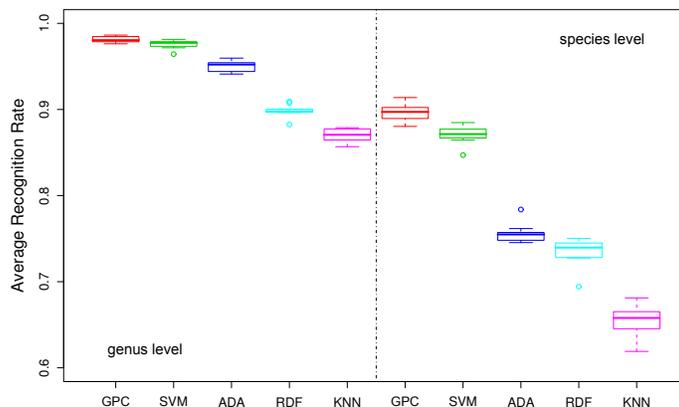


Fig. 3. Average recognition rates on the genus level (4 classes) and species level (10 classes) using ten repetitions with 75% of the data randomly chosen for training and the remainder for testing.

are given via Boxplots. They clearly show that the GP classifier (GPC) is suitable for our setting and provides significantly higher average recognition rates than the gold standard SVM (verified using t-test, $p < 0.05$). Please also note that all experiments were conducted with unit length normalized spectra, however, multiplication with a fixed constant yields analogous results in this setting.

5.2 Benefit of Prior Information

In the following we analyze different techniques from Sect. 4.2 to incorporate prior knowledge into the GP framework.

Negative Variance Cut-Off Heuristic To evaluate the effect of prior knowledge we analyzed a different scenario, where only 10 spectra per class are utilized for training. We tested our GP classifier on the remaining spectra for different values of the crucial regularization parameter γ for both types of normalization. Instead of ten repetitions of the randomized partitioning scheme we performed one hundred trials. As in [6], we utilized the Tangent distance which is invariant to global scaling of spectra and slowly varying background signals which are modeled by a linear combination of Lagrange bases of degree $deg = 3$. Since the assumed variations are both linear in the parameters, the Tangent distance is no approximation but calculates the mean of exact (regularized) distances between variation manifolds and patterns (see Fig. 2).

From Fig. 4 we clearly see that we observe different behavior when different kinds of normalization are employed. Using the standard Euclidean kernel, the unit length normalization (unit) yields substantially better performances compared to raw features (raw) multiplied with normalization constant c . When

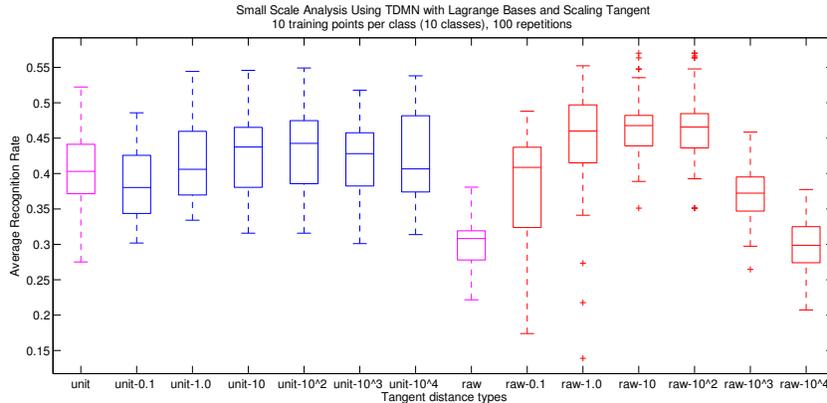


Fig. 4. Euclidean versus TDMN $_{\gamma}$ -induced squared exponential kernel for different values of regularization parameter $\gamma \in \{0.1, 1, 10, 10^2, 10^3, 10^4\}$ and different normalizations (unit, raw). Labels containing a number correspond to invariant kernels, e.g. raw- 10^3 corresponds to the kernel using raw data and regularization parameter $\gamma = 10^3$.

using invariant kernels with negative variance cut-off heuristic, however, the results turn out to be completely different. In case of unit length normalization, we only observe marginal improvements. This can be explained by the fact that we are applying an extra transformation to the spectra prior to using the invariant kernel. While the unit length transformation already provides a certain kind of scaling invariance, it simultaneously hinders the use of the transformation invariance that is embedded in the GP prior.

When using raw spectra, however, we observe a significantly higher performance gain compared to the Euclidean kernel (30.4% versus 46.8% on average). Moreover, using this normalization type leads to significantly higher average recognition rates than using unit length normalized data (3.5% for $\gamma = 10^2$).

It should be further noted that the benefit of using prior knowledge gets lost when using large amounts of training data, e.g. we even observed a performance drop of 2.5% compared to the non-invariant case when using 75% of the entire dataset for training.

Constant Tangents We already discussed in Sect. 4.2 that Tangent distance with constant tangents leads to positive semi-definite SE-Kernels. Since the Lagrange bases used in our approach are independent of the Raman spectra, we obtain overall constant tangents by discarding the non-constant scaling tangent. To evaluate this approach which does not need any heuristics or approximation, we repeated the above experiments and still observed a substantial performance gain over the Euclidean kernel (30.4% versus 47.8% on average for raw spectra). Compared to the negative variance cut-off heuristic, the results are nearly the same. Therefore and due to space limitations, we omit further results regarding this experiment.

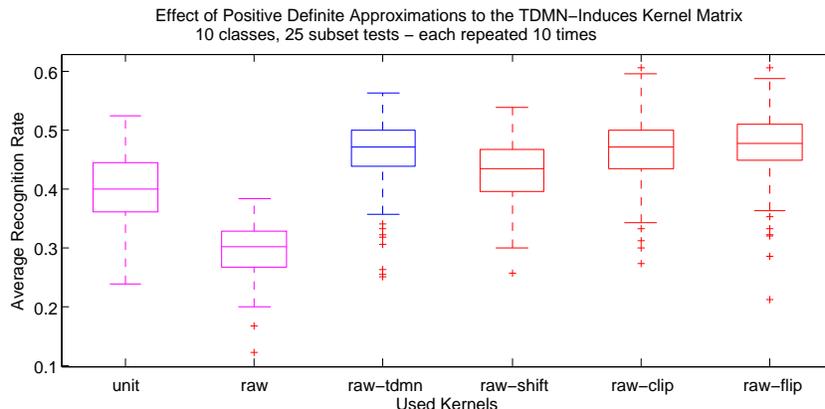


Fig. 5. Standard GP approach (unit, raw) and invariant GP heuristic (raw-tdmn) versus positive definite approximation methods (raw-shift, raw-clip, raw-flip).

Positive Definite Approximations As has been discussed in Sect. 4.2, the use of positive definite approximations is computationally prohibitive. We therefore adapted our experimental setting to accommodate for this fact. We randomly selected a fraction of 50 spectra for each category and trained the classifier with ten randomly chosen data points per category. Each such fraction was measured ten times. In order to obtain a measure for the entire dataset, this whole procedure was repeated 25 times. We thus ended up with a collection of 10×25 averaged recognition rates per classifier. We can see from Fig. 5 that no approximation substantially improved the invariant GP classifier (raw-tdmn) though non-positive eigenvalues occurred. Moreover, we can see that the rather fast SHIFT transform is even counterproductive, leading to a slightly decreased averaged performance (3%). The behavior of the SHIFT transform is even worse in large-scale training sets. E.g. when 75% of the data is used for training, we observe a performance drop of 55.8% compared to Euclidean kernels. This seems to be an inherent problem of this approximation method, as similar performances are reported for SVMs [13].

6 Conclusion and Future Work

This paper tackles the problem of Raman spectroscopic identification of microorganisms and introduces the Gaussian Process classifier to this field. In addition to the standard GP classifier we investigate the use of partially invariant covariance functions, where we embed known ideas from invariant kernel theory into the Gaussian Process framework. We highlight the shortcomings of this approach, i.e. that general GP classifiers are not able to cope with indefinite covariance matrices, and investigate methods to circumvent this issue. Empirical results show that the GP classifier outperforms state-of-the-art methods on a

large Raman dataset. Moreover, for the case of few training samples a significant performance gain is achieved via Tangent distance based covariance functions incorporating prior knowledge from possible pattern variations. Our results show that investigating invariant GP methods can indeed be beneficial. It would be further interesting to transfer our approach to other applications and to highlight relationships and empirically compare our work to other indefinite kernel methods.

Acknowledgments

We want to especially thank Mario Krause and Michaela Harz from the Institute of Physical Chemistry at the FSU Jena for capturing the large Raman spectra dataset.

References

1. Buijtelts, P., Willems-Erix, H., Petit, P., Endtz, H., Puppels, G., Verbrugh, H., van Belkum, A., van Soolingen, D., Maquelin, K.: Rapid identification of mycobacteria by raman spectroscopy. *Journal of Clinical Microbiology* **46**(3) (2008) 961–965
2. Rösch, P., Harz, M., Peschke, K.D., Ronneberger, O., Burkhardt, H., Motzkus, H.W., Lankers, M., Hofer, S., Thiele, H., Popp, J.: Chemotaxonomic identification of single bacteria by micro-raman spectroscopy: Application to clean-room-relevant biological contaminations. *Applied and Environmental Microbiology* **71** (2005) 1626–1637
3. Barrow, G., Feltham, R.: *Cowan and Steel’s Manual for the Identification of Medical Bacteria*. 3 edn. Cambridge University Press, UK (1993)
4. Amann, R., Fuchs, B.: Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology* **6** (2008) 339–348
5. Chen, T., Morris, J., Martin, E.: Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems* **87**(1) (May 2007) 59–71
6. Peschke, K.D., Haasdonk, B., Ronneberger, O., Burkhardt, H., Rösch, P., Harz, M., Popp, J.: Using transformation knowledge for the classification of raman spectra of biological samples. In: *Proceedings of the 24th IASTED international conference on Biomedical engineering*, Anaheim, CA, USA, ACTA Press (2006) 288–293
7. Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. *Journal of Machine Learning Research* **9** (2008) 2035–2078
8. Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press (2005)
9. Haasdonk, B.: *Transformation Knowledge in Pattern Analysis with Kernel Methods*. PhD thesis, Computer Science Department, University of Freiburg (2005)
10. Hassibi, B., Sayed, A., Kailath, T.: Linear estimation in krein spaces - part i: Theory. *IEEE Transactions on Automatic Control* **41**(1) (1996) 18–33
11. Casagrande, N.: Multiboost: An open source multi-class adaboost learner (2005) <http://iro.umontreal.ca/casagran/multiboost/>.
12. Joachims, T.: *Making large-Scale SVM Learning Practical*. MIT-Press (1999)
13. Chen, Y., Gupta, M.R., Recht, B.: Learning kernels from indefinite similarities. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, ACM (2009) 145–152