

Research Statement

VIKAS C. RAYKAR vikasraykar@gmail.com

My research interests lie in the area of statistical machine learning. As a research scientist at Siemens, my current research is focused on developing novel supervised learning algorithms to deal with imperfect supervision—especially subjective (crowdsourcing), noisy (multiple instance learning), and partial label information (survival analysis). This has been motivated by my first hand experiences dealing with messy real medical data. I am also currently working on empirical Bayesian methods for sparse high-dimensional prediction problems. My doctoral research focused on developing fast scalable machine learning algorithms for massive data sets using ideas inspired from computational physics and computational geometry. I will describe my key research accomplishments and my plans for future research in these areas. My research goal is to build a body of work which is both theoretically well-founded and readily applicable to real-world problems.

[**Research statement and accomplishments**]: October 2010

Contents

1 Learning with imperfect supervision	1
1.1 Learning from crowds—Subjective labels	2
1.2 Multiple instance learning—Noisy labels	4
1.3 Survival analysis—Incomplete labels	5
2 Sparsity in high dimensional settings	5
2.1 Mixture prior for adaptive sparsity	6
2.2 Mixture loss function for adaptive sparsity	6
3 Doctoral research: Scalable machine learning algorithms	6
3.1 Fast computation of kernel estimators	8
3.2 Large scale preference learning/ranking	9
3.3 Fast computation of sums of Gaussians	9
4 Predoctoral research: Audio signal processing	10

1. LEARNING WITH IMPERFECT SUPERVISION

I started this line of research when after my graduation I joined Siemens Healthcare as a research scientist. At Siemens I design machine learning algorithms for several commercially deployed computer aided diagnosis (CAD) products that automatically identify early stage cancer of the lung, colon, and breast based on X-ray, CT, and MRI images. These products have been commercially deployed in thousands of hospitals around the world and a resulting paper was the winner of the data mining practice prize ¹ for the best deployed data mining system at KDD 2009.

¹*Mining Medical Images* R. Bharat Rao, Glenn Fung, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, Vikas C. Raykar, Shipeng Yu, Sriram Krishnan, Xiang Zhou, Arun Krishnan, Marcos Salganicoff, Luca Bogoni, Matthias Wolf, Anna Jerebko, and Jonathan Stoeckel. Proceedings of the Third Workshop on Data Mining Case Studies and Practice Prize, Fifteenth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Paris, June 2009.

Most of the standard off-the-shelf supervised learning algorithms are generally developed for an ideal world. They often make strong assumptions which make them less than ideal for applying them directly to real world messy data. For example training points are often noisily labeled, training samples are not independent and identically distributed, the samples can be biased, it is not clear how to get the objective ground truth, the desired performance metric may be quite different etc. For these reasons most of the basic assumptions in developing classification algorithms have to be questioned. Suitable modifications must be made to model these deviations from the ideal scenarios. This situation lead to my interest in the development of new algorithms which give a significant improvement in performance over off-the-shelf standard classification algorithms under more realistic conditions. I will describe below my three important contributions in this area—which deal with *subjective, noisy, and partial label information*. This line of research has been motivated by problems in medical imaging—specifically in CAD where the task is to build a classifier to predict whether a suspicious region on a medical image (like a X-ray, CT scan, or MRI) is malignant or benign.

1.1 Learning from crowds—Subjective labels

1. | *Learning From Crowds* Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo H. Valadez, Charles Florin, Luca Bogoni, and Linda Moy, Journal of Machine Learning Research, Vol. 11, pp. 1297–1322, April 2010.
2. | *Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit* Vikas C. Raykar, Shipeng Yu, Linda Zhao, Anna Jerebko, Charles Florin, Gerardo Valadez, Luca Bogoni, and Linda Moy, Proceedings of the 26th International Conference on Machine Learning (ICML 2009), pp. 889–896, Montreal, June 2009.

For many supervised learning tasks it may be infeasible (or very expensive) to obtain objective and reliable labels for training. Instead, we can collect subjective (possibly noisy) labels from *multiple experts or annotators*. In practice, there is a substantial amount of disagreement among the annotators, and hence it is of great practical interest to address conventional supervised learning problems in this scenario. For example in the medical imaging domain the actual gold standard (whether there is cancer or not) can only be obtained from biopsy of the tissue. Since this is an expensive, invasive, and potentially dangerous process, often systems are built from labels assigned by multiple radiologists who identify the locations of malignant lesions. Each radiologist visually examines the medical images and provides a subjective (possibly noisy) version of the gold standard. The radiologists come from a diverse pool including luminaries, experts, residents, and novices—very often there is lot of disagreement among the annotations. For a lot of tasks the labels provided by the annotators are *inherently subjective* and there will be substantial variation among different annotators.

The domain of text classification offers such a scenario. In this context the task is to predict the category for a token of text. The labels for training are assigned by human annotators who read the text and attribute their subjective category. With the advent of crowdsourcing services like Amazon’s Mechanical Turk, CrowdFlower, Games with a Purpose, and reCAPTCHA it is quite inexpensive to acquire labels

from a large number of annotators (possibly thousands) in a short time. In situations like these, the performance of different annotators can vary widely (some may even be malicious), and without the actual gold standard, it may not be possible to evaluate the annotators.

In papers [1] and [2] I propose a probabilistic approach for supervised learning which addresses the following three issues simultaneously: (1) How to adapt conventional supervised learning algorithms when we have multiple annotators providing subjective labels but no objective gold standard? (2) How to evaluate systems when we do not have an absolute gold-standard? (3) A closely related problem—particularly relevant when there are a large number of annotators—is to estimate how reliable/trustworthy is each annotator. The commonly used majority voting scheme uses the labels on which the majority agree as an estimate of the actual gold standard. These papers propose a Bayesian approach that jointly learns the classifier, the annotator accuracy, and the unknown true label. The final estimation is performed by an Expectation Maximization algorithm that iteratively establishes a particular gold standard, measures the performance of the experts given that gold standard, and refines the gold standard based on the performance measures. Experimental results indicate that the proposed method is superior to the commonly used majority voting baseline. A novel feature is that the proposed algorithm *learns the classifier and the ground truth jointly*—in a way the classifier is allowed to influence the ground truth. The method was successfully applied to a model for prediction of malignancy for breast tumors in MR with subjective assessments from multiple radiologists in the absence of biopsy results. It was also applied to text data collected from the Amazon’s mechanical turk.

Since the publication of this paper there has been a flurry of interest in the machine learning/computer vision/natural language processing community in harnessing the power of crowds for various tasks (crowdsourcing). Within a year the paper has received around 25 citations and has been extensively discussed in different blogs. This a very fertile ground for further research and introduces a number of interesting learning problems that have either not been studied or have been scarcely studied in the past. I am quite interested in forming a research group around this topic and collaborating with people from different communities like natural-language processing, computer-vision, medical imaging, astronomy, telera-diology, etc. where crowdsourcing is becoming quite common. While there are a lot of practical applications there are still a few fundamental issues that need to be worked on. For example a key assumption made in [1] and [2] is that conditional on the true labels the annotators are independent. This assumption is not true in general and there are some correlations among the labels assigned by multiple annotators. Since crowdsourcing services pay the annotator for their labeling services I would like to analyze the monetary and behavioral aspects of crowdsourcing services. The amount to be paid can be linked to the performance of the annotator and a general utility based algorithm can be derived. Active learning in this setup also raises a very interesting question—Which instance to be labeled by which annotator? A more practical line of work would be to develop algorithms which can effectively weed out spammers and malicious/adversarial annotators which are quite common in crowdsourcing market places.

1.2 Multiple instance learning—Noisy labels

3. *Bayesian Multiple Instance Learning: Automatic Feature Selection and Inductive Transfer* Vikas C. Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R. Bharat Rao, Proceedings of the 25th International Conference on Machine Learning (ICML 2008), pp. 808–815, Helsinki, July 2008.
4. *A Multiple Instance Learning Approach toward Optimal Classification of Pathology Slides* Murat Dundar, Sunil Badve, Vikas C. Raykar, Rohit Jain, Olcay Sertel, and Metin Gurcan Proc. of 20th Int. Conference on Pattern Recognition (ICPR 2010). [Best Scientific Paper Award in Bioinformatics and Biomedical Applications Track]

The previous work addressed how to deal with the subjective labels. This work addresses a much more practical issue of a specific kind of mislabeled instance. In a conventional supervised learning scenario it is always assumed that the label is *given for every instance*. However in many practical applications labels are available at a much higher granularity and are not available for every instance. For example in a lot of medical imaging applications the radiologist who provides us the ground truth just marks the location of the lesion. The lesions are often irregular in shape and are of different sizes. The computer algorithm designed to detect these lesions produces a lot of training examples which are spatially close to each other. All these examples point to the same ground truth. A single instance classifier considers all these examples as positive. However in practice it often happens that there will be a lot of negatives which mistakenly get labeled as positives.

In the multiple instance learning (MIL) framework the training set consists of what are known as *bags*. A bag contains many instances. All the instances in a bag share the same bag-level label. A bag is labeled positive if it contains *at least* one positive instance. A negative bag means that *all* instances in the bag are negative. By using a probabilistic AND-OR formulation of the above definition I designed a novel MIL algorithm in an empirical Bayesian framework that was also automatically able to identify the relevant feature subset. One interesting empirical outcome was that the multiple instance model was able to select many fewer features—almost half the number of features selected by the single instance approach. This algorithm was clearly superior to the previous MIL algorithms and was presented at ICML 2008. This was also demonstrated in a paper ² which was presented at a digital mammography conference and was written in collaboration with scientists in Israel.

Recently I also collaborated with researchers from Indiana and Purdue University where I applied similar MIL ideas for automated classification of pathology slides. This paper [4] has been selected as the best scientific paper in the Bioinformatics and Biomedical Applications track at the 20th International Conference on Pattern Recognition. There were 2140 submissions across six tracks and one paper from each track has received this award.

²*Multiple instance learning improves CAD detection of masses in digital mammography* Balaji Krishnapuram, Jonathan Stoeckel, Vikas C. Raykar, R. Bharat Rao, Philippe Bamberger, Eli Ratner, Nicolas Merlet, Inna stainvas, Menahem Abramov, and Alexandra Manevitch, Proc. of the 9th int. workshop on Digital Mammography (IWDM), pp. 350–357, Tucson, AZ, July 2008.

1.3 Survival analysis—Incomplete labels

5. | *On Ranking in Survival Analysis: Bounds on the Concordance Index* Vikas C. Raykar, Harald Steck, Balaji Krishnapuram, Cary Oberije, Philippe Lambin, Advances in Neural Information Proc. Systems (NIPS), vol. 20, pp. 1209-1216, 2008.

Survival analysis is a well-established field in medical statistics concerned with analyzing/predicting the time until the occurrence of an event of interest, e.g., death, onset of a disease, or failure of a machine. I started looking into survival analysis when I was looking into data concerning the survival time of non-small cell lung cancer patients, which we analyzed as part of our collaboration with the MAASTRO clinic. My machine learning background helped me to discover some connections between classical survival analysis and the ranking literature in machine learning. This paper shows that classical survival analysis involving censored data can naturally be cast as a ranking problem. The concordance index (CI), which quantifies the quality of rankings, is the standard performance measure for model assessment in survival analysis. In contrast, the standard approach to learning the popular proportional hazard model is based on Cox’s partial likelihood. The paper derives two bounds on the CI and optimizes them directly. I also explain why a method designed to maximize the Cox’s partial likelihood also ends up maximizing the CI.

2. SPARSITY IN HIGH DIMENSIONAL SETTINGS

Recently I have also started working on high-dimensional prediction and estimation problems, usually referred to as the large p , small n ($p \gg n$) paradigm, p being the dimension of the model and n the sample size. With the advent of modern scientific technology like microarrays and fMRI machines such high dimensional data have become quite common and pose a challenge to the conventional machine learning/statistical inference techniques.

In high dimensional scenarios it is desirable to obtain sparse solutions. A sparse solution generally helps in better interpretation of the model and more importantly leads to better generalization on unseen data. We developed two different techniques to achieve sparsity in high dimensional scenarios—one is via a *mixture prior* and another via a *mixture loss function*. Both these methods are developed in the empirical Bayesian framework which combines both frequentist and Bayesian ideas. This research is in collaboration with Dr. Linda Zhao at the department of statistics at University of Pennsylvania. I am quite excited about this line of research and would like to further explore connections with the multiple hypothesis testing area currently popular in the statistics and genomics community.

2.1 Mixture prior for adaptive sparsity

6. | *Nonparametric prior for adaptive sparsity* Vikas C. Raykar and Linda H. Zhao In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010), JMLR: W&CP 9, pp. 629–636, Italy, May 2010.

For high-dimensional problems various parametric priors (like the zero mean Gaus-

sian, Laplace) have been proposed to promote sparse solutions. While parametric priors have shown considerable success they are not very robust in *adapting to varying degrees of sparsity*. This work proposes a discrete mixture prior which is *partially nonparametric*. The right structure for the prior and the amount of sparsity is estimated directly from the data. The key idea is to assume that there is a prior on the parameter but to impose no structural assumptions on that prior distribution and estimate it directly from the data. An iterative EM algorithm based on weighted non-parametric kernel density estimate is developed to estimate the sparsity in the signal. Our experiments show that the proposed prior adapts to sparsity much better than its parametric counterparts. The method is applied here to classification of high dimensional microarray datasets.

2.2 Mixture loss function for adaptive sparsity

7. | *Empirical Bayesian thresholding for sparse signals using mixture loss functions* Vikas C. Raykar and Linda H. Zhao To appear in Statistica Sinica.

We developed an empirical Bayesian thresholding rule for the normal mean problem that adapts well to the sparsity of the signal. As earlier the prior on each mean is a mixture of an atom of probability at zero, and a Laplace or normal density for the nonzero part. A novel key element is the use of a *mixture loss function* that combines both the L_p loss and the $0 - 1$ loss function. The Bayes procedures under this loss are explicitly given as thresholding rules and are easy to compute.

3. DOCTORAL RESEARCH: SCALABLE MACHINE LEARNING ALGORITHMS

Huge data sets containing millions of training examples with large number of attributes (*tall fat data*) are relatively easy to gather. However one of the bottlenecks for successful inference of useful information from the data is the computational complexity of modern machine learning algorithms. Most state-of-the-art non-parametric machine learning algorithms have a computational complexity of either $\mathcal{O}(N^2)$ or $\mathcal{O}(N^3)$, where N is the number of training examples. This has seriously restricted the use of massive data sets. The bottleneck computational primitive at the heart of various algorithms is the multiplication of a structured matrix with a vector, which we refer to as *matrix-vector product* (MVP) primitive. The goal of my doctoral research was to speedup up these MVP primitives by *fast approximate algorithms* that scale as $\mathcal{O}(N)$ and also provide *high accuracy guarantees*. I used ideas from computational physics, scientific computing, and computational geometry to design these algorithms with the aim of getting getting *good enough solutions as fast as possible*. The various fast algorithms that I developed during my doctoral dissertation are released under the GNU Lesser General Public License and have been widely downloaded. The proposed algorithms have been applied to various tasks like kernel density estimation, optimal bandwidth estimation, projection pursuit, Gaussian process regression, SVMs, implicit surface fitting, and ranking.

Weighted superposition of kernels

In most kernel based machine learning algorithms, Gaussian processes, and non-parametric statistics a key computationally intensive task is to compute a linear

combination of local kernel functions centered on the training data, *i.e.*,

$$f(x) = \sum_{i=1}^N q_i k(x, x_i), \quad (1)$$

where $\{x_i \in \mathbb{R}^d, i = 1, \dots, N\}$ are the N training data points, $\{q_i \in \mathbb{R}, i = 1, \dots, N\}$ are the weights, $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the local kernel function, and $x \in \mathbb{R}^d$ is the test point at which f is to be computed. For kernel machines (e.g. regularized least squares, support vector machines, kernel regression) f is the regression/classification function. In case of Gaussian process regression f is the mean prediction. For non-parametric density estimation it is the kernel density estimate. Training these models scales as $\mathcal{O}(N^3)$ since most involve solving the linear system of equation $(\mathbf{K} + \lambda \mathbf{I})\xi = \mathbf{y}$, where \mathbf{K} is the $N \times N$ Gram matrix where $[\mathbf{K}]_{ij} = k(x_i, x_j)$. Recently, such nonparametric problems have been collectively referred to as *N-body problems in learning*³, in analogy with the gravitational N -body potential problems occurring in computational physics. In general we need to evaluate (1) at M points $\{y_j \in \mathbb{R}^d, j = 1, \dots, M\}$ leading to the quadratic $\mathcal{O}(MN)$ cost. The sum can be thought of as a *matrix-vector product* $f = \mathbf{K}q$, where \mathbf{K} is a $M \times N$ matrix the entries of which are of the form $[\mathbf{K}]_{ij} = k(y_j, x_i)$ and $q = [q_1, \dots, q_N]^T$ is a $N \times 1$ column vector. I developed fast ϵ -exact algorithms that compute the sum approximately in linear $\mathcal{O}(M + N)$ time. The algorithm is ϵ -exact, *i.e.*, for any given $\epsilon > 0$, \hat{f} is an ϵ -exact approximation to f if $\max_{y_j} [|\hat{f}(y_j) - f(y_j)|/Q] \leq \epsilon$ where $Q = \sum_{i=1}^N |q_i|$. The constant in $\mathcal{O}(M + N)$, depends on the desired accuracy ϵ , which however can be *arbitrary*. The fast algorithm is based on series expansion of the kernel and retaining only the first few terms contributing to the desired accuracy. The algorithms are in the spirit of *fast multipole methods* used in computational physics⁴.

My thesis consists of two core contributions—(1) design of fast summation algorithms and (2) applying these fast primitives to certain large scale machine learning problems. Table I summarizes the three key fast summation algorithms that I developed during my doctoral thesis and the corresponding applications. Below are some of the key papers in this stream of research.

3.1 Fast computation of kernel estimators

8. *Fast Computation of Kernel Estimators* Vikas C. Raykar, Ramani Duraiswami, and Linda H. Zhao, Journal of Computational and Graphical Statistics., Vol. 19, No. 1, pp. 205–220, March 2010.
9. *Fast optimal bandwidth selection for kernel density estimation* Vikas C. Raykar and Ramani Duraiswami, Proceedings of the sixth SIAM International Conference on Data Mining, pp. 524–528, Bethesda, April 2006.

The computational complexity of evaluating the kernel density estimate (or its

³Gray, A. and Moore, A. 2001. N-body problems in statistical learning. In Advances in Neural Information Processing Systems. 521–527.

⁴Greengard, L. 1994. Fast algorithms for classical physics. Science 265, 5174, 909–914.

Core MVP primitive and applications
Gaussian $G(y_j) = \sum_{i=1}^N q_i e^{-\ y_j - x_i\ ^2/h^2}$ kernel density estimation, Gaussian process regression, implicit surface fitting
Hermite \times Gaussian $G(y_j) = \sum_{i=1}^N q_i H_r \left(\frac{y_j - x_i}{h_1} \right) e^{-(y_j - x_i)^2/h_2^2}$ kernel density derivative estimation, optimal bandwidth estimation, projection pursuit error function
error function $G(y_j) = \sum_{i=1}^N q_i \operatorname{erfc}(y_j - x_i)$ ranking, collaborative filtering

Table I. The fast summation algorithms designed and tasks to which they were applied. derivatives) at M evaluation points given N sample points scales quadratically as $\mathcal{O}(NM)$ —making it prohibitively expensive for large datasets. While approximate methods like binning have been used speed up the computation, they lack a precise control over the accuracy of the approximation. There is no straightforward way of choosing the binning parameters a priori in order to achieve a desired approximation error. I proposed a novel computationally efficient ϵ -exact approximation algorithm for the univariate Gaussian kernel-based density derivative estimation that reduces the computational complexity from $\mathcal{O}(NM)$ to linear $\mathcal{O}(N + M)$. The user can specify a desired accuracy ϵ . The algorithm guarantees that the actual error between the approximation and the original kernel estimate will always be less than ϵ . I also applied the proposed fast algorithm to speed up automatic bandwidth selection procedures. I compared the method to the best available binning methods in terms of the speed and the accuracy. Experimental results show that the proposed method is almost twice as fast as the best binning methods and is around five orders of magnitude more accurate. For example, at $N = 10^6$ sample points the direct computation takes 379.27 sec while the proposed method takes only 0.92 sec (a speedup of 412). I also demonstrated that the proposed procedure can be extremely useful for speeding up exploratory projection pursuit techniques.

3.2 Large scale preference learning/ranking

10. *A fast algorithm for learning a ranking function from large scale data sets* Vikas C. Raykar, Ramani Duraiswami, and Balaji Krishnapuram, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 7, pp. 1158–1170, July 2008.
11. *A fast algorithm for learning large scale preference relations* Vikas C. Raykar, Ramani Duraiswami, and Balaji Krishnapuram, Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 385-392, Puerto Rico, March 2007.

The problem of ranking has recently received significant attention in the statistical machine learning and information retrieval communities. In a typical ranking formulation, we compare two instances and determine which one is better or preferred. Based on this, a set of instances can be ranked according to a desired preference relation. The study of ranking has largely been motivated by applications in search engines, information retrieval, collaborative filtering, and recommender systems.

Most of the previous published ranking algorithms suffered for the quadratic computational complexity. In this work I consider the problem of learning the ranking function that maximizes a generalization of the Wilcoxon-Mann-Whitney statistic on the training data. Relying on an ϵ -exact approximation for the error-function, I reduced the computational complexity of each iteration of a conjugate gradient algorithm for learning ranking functions from $\mathcal{O}(m^2)$ to $\mathcal{O}(m)$. Experiments on public benchmarks for ordinal regression and collaborative filtering indicate that the proposed algorithm is as accurate as the best available methods in terms of ranking accuracy, when the algorithms are trained on the same data. However, since it is several orders of magnitude faster than the current state-of-the-art approaches, it is able to leverage much larger training datasets containing tens to hundreds of thousands of samples.

3.3 Fast computation of sums of Gaussians

12. | *Fast computation of sums of Gaussians in high dimensions* Vikas C. Raykar, C. Yang, R. Duraiswami, and N. Gumerov, CS-TR-4767, Department of computer science, University of Maryland, Collegepark.
13. | *The Improved Fast Gauss Transform with applications to machine learning* Vikas C. Raykar and Ramani Duraiswami, Book chapter in Large-Scale Kernel Machines, pp. 175–201, MIT Press 2007.
14. | *Automatic online tuning for fast Gaussian summation* Vlad I. Morariu, Balaji V. Srinivasan, Vikas C. Raykar, Ramani Duraiswami, and Larry Davis, Advances in Neural Information Processing Systems (NIPS), vol. 21, pp. 1113–1120, 2009.

The most commonly used kernel function in various machine learning algorithms is the Gaussian kernel $e^{-\|x-y\|^2/h^2}$, where h is called the bandwidth of the kernel. The fast Gauss transform⁵ is a well known ϵ -exact approximation algorithm that reduces the computational complexity of the evaluation of the sum of N Gaussians at M points in d dimensions from $\mathcal{O}(MN)$ to $\mathcal{O}(M + N)$. However, the constant factor in $\mathcal{O}(M + N)$ grows exponentially with increasing dimensionality d , which makes the algorithm impractical for dimensions greater than three. I presented a new algorithm where the constant factor is reduced to *asymptotically polynomial order*. The reduction is based on a new multivariate Taylor series expansion scheme combined with the efficient space subdivision using the k -center algorithm. This algorithm gives good speedups in dimensions as high as tens for moderate bandwidths and as high as hundreds for large and small bandwidths.

The proposed algorithm suffered from two problems: the Taylor series expansion does not perform well for very low bandwidths, and parameter selection was not trivial and can drastically affect performance and ease of use. Recently I have been collaborating with my advisor on integrating the algorithm with approximate nearest neighbor searching. This resulted in the development of the FIGTree algorithm (published at NIPS 2009) which can now be essentially used as a black box. The

⁵Greengard, L. and Strain, J. 1991. The fast Gauss transform. SIAM Journal of Scientific and Statistical Computing 12, 1, 79–94.

algorithm combines four different methods and automatically predicts the fastest method and tunes its parameters for the given dataset.

In the future I would like to work on the development of these kind of fast approximate algorithms for more kernels—e.g., the Epanechnikov kernel for kernel density estimation and the Matèrn class of kernels used in Gaussian process regression and spatial analysis. In many applications these fast MVP primitives are embedded in a optimization routine. A theoretical issue which I have barely touched upon concerns the convergence of these optimization routines when using approximate MVP primitives. For most machine learning tasks even though the data is very high dimensional, the true intrinsic dimensionality is typically very small. I intend to explore if dimensionality reduction approaches like PCA and manifold learning methods can be directly incorporated into our fast algorithms. A more ambitious task would be to explore if there are any deeper connections between structure in the data, computation, and inference.

4. PREDOCTORAL RESEARCH: AUDIO SIGNAL PROCESSING

15. | *Extracting the frequencies of the pinna spectral notches in measured head related impulse responses* Vikas C. Raykar, Ramani Duraiswami, and B. Yegnanarayana, The Journal of the Acoustical Society of America, Vol. 118, No. 1, pp. 364–374, July 2005.
16. | *Position Calibration of Microphones and Loudspeakers in Distributed Computing Platforms* Vikas C. Raykar, Igor Kozintsev, and Rainer Lienhart, IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 1, pp. 70–83, Jan. 2005.

I did my masters in electrical engineering with a specialization in audio signal processing. Here are a couple of representative publications. The first paper explored how the shape of the external ear (pinna) helps in the localization of the sound source. The second paper describes an algorithm which can localize a bunch of laptops placed in a conference room based on audio signals.