Research statement of Rohit Patra

My core statistical research focuses on semiparametric/nonparametric methodology and large sample theory — efficient estimation in semiparametric models, nonparametric function estimation (with emphasis on shape constrained estimation), and bootstrap based inference in non-standard problem. I am also actively involved in interdisciplinary research in astronomy.

My research has applications in broad areas such as genomics (multiple testing problems), economics (utility and production function estimation and binary response models), causal inference (conditional independence and dimension reduction) and astronomy (understanding the accretion history of galaxies), among other fields.

In the following, first I give a brief overview of my research. The second part of the document contains a detailed description of my research as well as the related future research directions.

Summary of Research

Mixture Models: Two-component mixture models arise in multiple testing problems and more generally in contamination problems. We studied these model without any parametric or nonparametric assumptions. We developed the first distribution-free and tuning parameter-free confidence lower bound and a consistent estimator for the mixing proportion.

Single Index Models: We have developed semiparametric procedures for inference in single index models (SIMs). SIMs are the nonparametric generalizations of GLMs where no parametric assumptions are made on the link function. In contrast to the traditional kernel based methods, I use smoothing splines for efficient estimation of the finite dimensional parameter under weaker smoothness assumptions on the link function. SIMs also arise naturally in economics, where the link functions are often assumed to be concave, e.g., production and utility functions. We have developed a semiparametrically efficient estimator for the unknown finite dimensional parameter under the shape constraint that the link function is convex; the problem is non-standard as the estimator lies on the boundary on the parameter space due to the imposed shape constraint. A fast R implementation of the method can be found in the *simest* package.

Bootstrap for Maximum Score Estimator: In the latent variable binary choice model, the maximum score estimator exhibits non-standard asymptotics and the classical bootstrap is not consistent. We proposed a smoothed bootstrap alternative for inference on the maximum score estimator and proved that the proposed procedure is consistent.

Test for Conditional Independence: We constructed a nonparametric procedure to test for conditional independence using a *nonparametric notion of residual.*

Astro-statistics: In this ongoing work, we are using techniques from image processing and statistics to develop automated approaches to identify the two structures — shells and streams — formed during accretion of the galaxies. The study of these hierarchical structures will provide a powerful probe into a galaxy's accretion history.

More Details on Research

1 Mixture Models (Patra and Sen (2015))

Multiple testing problems appear in many scientific areas, from microarray analysis to neuroimaging. In these problems one observes a sample of p-values obtained from numerous hypothesis tests. The p-value for each test is known to be Uniform$(0, 1)$ when the corresponding null hypothesis is true. The performance of many methods that control the false discovery rate or the local false discovery rate crucially depend on (i) the proportion of non-null hypotheses and (ii) the distribution of the p-values under the alternative. In Patra and Sen (2015) we consider a two-component mixture model with observations from the cumulative distribution function (CDF)

$$F(x) = \alpha F_s(x) + (1 - \alpha)F_b(x), \tag{1}$$

where the CDF F_b is known (e.g., Uniform$(0, 1)$), but the mixing proportion $\alpha \in [0, 1]$ (e.g., true proportion of non-null hypotheses) and the CDF F_s ($\neq F_b$) are unknown. More generally, this model arises in contamination problems.

In contrast to the existing approaches where F_s is assumed to satisfy some parametric/nonparametric constraints, we studied the model without any such restrictions on F_s. However, without any constraint on F_s model (1) is not identifiable (a trivial solution occurs when $\alpha = 1$ and $F_s = F$). We showed the existence of a unique α_0 such that every $\alpha \geq \alpha_0$ satisfies (1), and for any $\alpha < \alpha_0$ the model does not hold. When F is continuous, given a i.i.d. sample from F, we developed an *honest finite sample lower confidence bound* for α_0. Our bound is the first such lower bound which is distribution-free and completely automated. For example, such lower bounds can be used to test the hypothesis of *no signal* (i.e., $\alpha_0 = 0$) at a specified *level*.

We further proposed a fast and easily computable consistent estimator of α_0 and showed that its rate of convergence can be made arbitrarily close to $n^{-1/2}$ by an appropriate choice of a tuning parameter. We also showed that the estimator of α_0, properly scaled, converges to a non-zero degenerate limit. With a consistent estimator of α_0 in hand we proposed a uniformly consistent nonparametric estimator of F_s using ideas from shape restricted function estimation. Under the additional assumption that F_s has a decreasing density, we proposed a tuning parameter-free method to consistently estimate the density. Extensive simulations over a broad set of scenarios (including under dependent sampling) illustrated the superior finite sample performance of both the lower confidence bound and the consistent estimator of α_0.

1.1 Plans for Future Research

Construction of two-sided confidence intervals for α_0 remains a hard problem as the asymptotic distribution of the proposed estimator depends on the unknown F. In Section 5 of Patra and Sen (2015) we proposed a tuning parameter-free heuristic estimator of α_0 which has very good finite sample performance. However, the asymptotic properties of this estimator are unknown and of interest. Formal goodness-of-fit tests for F_s are also important as they can guide the practitioner to use appropriate parametric models for further analysis.

In many cases, F_b is not exactly known. We are currently developing consistent estimators of α_0 when we only have an estimator of F_b (e.g., we have another i.i.d. sample from F_b).

In many multiple testing problems, the observed p-values can be dependent. Although our proposed procedure has good finite sample performance under many dependence scenarios, a formal theoretical study of this is still missing.

2 Single Index Models

Suppose that we have i.i.d. realizations of a random vector $(Y, X) \in \mathbb{R} \times \mathbb{R}^d (d \geq 1)$ and we are interested in estimating the regression function $\mathbb{E}(Y|X)$. Without any assumptions on the functional form of the regression function, nonparametric estimators suffer from the *curse of dimensionality* and can be hard to interpret, especially when d is large. Single index models (SIMs) offer a natural dimension reduction technique while still allowing for flexibility in the functional form of the regression function and are ubiquitous in biostatistics and economics among other fields. In SIMs one assumes that there exists $\theta_0 \in \mathbb{R}^d$ such that $\mathbb{E}(Y|X) = \mathbb{E}(Y|X^\top\theta_0)$; the widely used generalized linear models (GLMs) are special cases. In two current works Patra (2015) and Kuchibhotla et al. (2015) we consider the model

$$Y = m_0(\theta_0^\top X) + \epsilon, \quad \mathbb{E}(\epsilon|X) = 0, \tag{2}$$

where $m_0 : \mathbb{R} \to \mathbb{R}$ is called the link function, θ_0 is the index parameter, and ϵ is the unobserved error. The unknown parameters of interest are θ_0 and m_0. For identification purposes we assume that $\theta_0 \in \mathcal{S}^{d-1}$, the d-dimensional Euclidean sphere, and that the first coordinate of θ_0 is strictly positive. We assume that the parameters of interest m_0 and θ_0 are unknown.

2.1 Smooth Single Index Model with Smoothing Splines (Patra (2015))

In Patra (2015), I study the problem of estimating m_0 and θ_0 when m_0 is smooth. Most of the estimation techniques in (2) use kernel smoothing techniques (or finite dimensional spline bases) to get an estimator of the link function for every fixed θ and then optimize a criterion function over θ. However these estimates of the link function tend to produce unstable estimators, see Yu and Ruppert (2002).

In Patra (2015), I propose (for the first time) the use of smoothing splines to estimate θ_0 and m_0 simultaneously. I assume that m_0 has an absolutely continuous first derivative. Given an i.i.d. sample $(y_i, x_i)_{1\leq i\leq n}$ from (2), I define the penalized least square estimator

$$(\hat{\theta}, \hat{m}) := \underset{m\in\mathcal{M}, \theta\in\mathcal{S}^{d-1}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\big(y_i - m(\theta^\top x_i)\big)^2 + \hat{\lambda}_n^2 \int |m''(t)|^2 dt, \tag{3}$$

where $\mathcal{M}$ is the class of all functions with absolutely continuous derivative, $\hat{\lambda}_n$ is the smoothing parameter (which can be chosen in a data driven fashion as long as it goes to 0 at a certain rate). When θ_0 is known, the minimizer for the penalized loss is a well-studied object and has a closed-form solution (a cubic spline). However, in the case of SIMs the theory is considerably harder as both m_0 and θ_0 are unknown and *intertwined*. I show that $(\hat{\theta}, \hat{m})$ is unique. I study the identifiability and existence of the estimators and prove that $\hat{m}$ converges to m_0 at rate $n^{-2/5}$ and, under mild regularity assumptions, $\hat{\theta}$ is a Fisher efficient $n^{-1/2}$-consistent estimator of θ_0.

2.2 Shape Restricted Single Index Model (Kuchibhotla et al. (2015))

The motivation for shape constrained SIMs arise from economics where utility and productions function are often assumed to be concave. In this work we estimate the link function in (2) under the constraint that it is convex. We further assume that m_0 is uniformly Lipschitz. We propose the following least squares estimator

$$(\tilde{\theta}, \tilde{m}) := \underset{m \in \mathcal{C}_L,\ \theta \in \mathcal{S}^{d-1}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - m(\theta^\top x_i)\big)^2, \tag{4}$$

where $\mathcal{C}_L$ is the class of all uniformly Lipschitz convex functions with Lipschitz bound L. We show that $\tilde{m}$ converges to m_0 at rate $n^{-2/5}$, $\tilde{m}'$ converges to m_0' at rate $n^{-1/5}$, and, under mild regularity assumptions, that $\tilde{\theta}$ is a Fisher efficient $n^{-1/2}$-consistent estimator of θ_0. The problem is non-standard and challenging as the link function is constrained to be convex and neither a least favorable model nor a *approximately least favorable subprovided model* (see Section 9.2 of van der Vaart (2002)) exist at $(\tilde{\theta}, \tilde{m})$. This can be attributed to the fact that $(\tilde{\theta}, \tilde{m})$ lies on the *boundary* (due to the imposed shape constraint) of the parameter space. We believe that this is the first work on SIMs without any smoothness constraint on m_0.

For a fixed θ minimizing the loss in (4) is a quadratic program and can be solved in a fast and efficient manner. With an abundance of application where we feel our work is useful, we developed and implemented (see the R package Kuchibhotla and Patra (2015)) a fast alternating gradient descent algorithm to compute the estimators in both (3) and (4).

2.3 Plans for Future Research

The joint minimization problem in both (3) and (4) are non-convex. However, in simulation studies the proposed algorithm compares favorably with existing methods and is scalable for large values of d (~ 100). We would like to thoroughly study the theoretical and numerical convergence properties of proposed algorithm.

When the link function is assumed to be convex, we are hoping to develop a consistent estimator without the uniform Lipschitz restriction. This would be interesting as the above procedure would then be tuning parameter free.

I have recently started working with Prof. Josè R. Zubizarreta on a project in causal inference. In observational studies in economics nearest neighbor matching estimators are widely used to estimate the average treatment effect. However, the estimator suffers from the curse of dimensionality. I am currently exploring dimension reduction techniques such as SIMs and studying their appropriateness in the casual inference paradigm.

3 Bootstrap for Maximum Score Estimator (Patra et al. (2015a))

Latent variable binary response models naturally arise in many econometric applications. To elaborate, let X be a continuous random vector, U denote an unobserved random variable, $\mathbf{1}$ denote the indicator function and $\beta_0 \in \mathbb{R}^d$. Manski (1975) considered the model

$$Y = \mathbf{1}_{\beta_0^\top X + U \geq 0}, \tag{5}$$

and proposed a consistent estimator of β_0 under the assumption that median$(U|X) = 0$; this estimator is known as the maximum score estimator. Note that logistic ($X \perp\!\!\!\perp U$ and

$U \sim$ Logistic distribution) and probit regression models are special cases of (5). The median constraint is very flexible and allows for unknown forms of heteroscedasticity of $U|X$. Kim and Pollard (1990) showed that the maximum score estimator converges to β_0 at rate $n^{-1/3}$ and has a non-standard limiting distribution. The complicated nature of the limiting distribution has been the main obstacle to application of the maximum score estimator.

In problems with non-standard asymptotics, bootstrap is a natural alternative for doing inference. However, Abrevaya and Huang (2005) showed that for the maximum score estimator, classical bootstrap (sampling with replacement) is inconsistent (i.e., asymptotically it fails to replicate the actual limiting distribution). In Patra et al. (2015a) we proposed a smoothed bootstrap procedure for inference on the maximum score estimator. Our bootstrap procedure crucially uses the model setup and the model assumptions in (5). The analysis of the bootstrap in these problems is theoretically quite challenging, and we used tools from non-standard empirical process techniques.

3.1 Plans for Future Research

In many problems, mainly those with non-standard asymptotics, the classical bootstrap is not consistent. A model-based bootstrap procedure can provide a consistent alternative, see e.g., Seijo and Sen (2011). In most high-dimensional models inference is hard and model-based bootstrap can be a useful tool. I would like to explore and understand some of these problems in the future.

4 A Test for Conditional Independence (Patra et al. (2015b))

Conditional independence is at the heart of fields like causal inference and graphical models. In Patra et al. (2015b), we proposed a test for the conditional independence of random variables X and Y given Z ($X \perp\!\!\!\perp Y|Z$) using a newly proposed nonparametric notion of residual of X on Z. We showed that this notion of residual extends the traditional notion of residual when (X, Z) is a multivariate Gaussian random vector. We showed that $X \perp\!\!\!\perp Y|Z$ is equivalent to the mutual independence of Z, the residual of X on Z, and the residual of Y on Z. We also developed an easy-to-implement one-sample goodness-of-fit procedure to test the mutual independence. Since the asymptotic distribution of the test statistic is unknown, we designed a model-based bootstrap procedure to approximate the critical value of the test.

4.1 Plans for Future Research

The asymptotic properties of the proposed test need to be investigated. The proposed residuals can be used to define a notion of nonparametric partial correlation, another direction for future research.

5 Astro-statistics (Biswas et al. (2015))

Simulation studies suggest that larger galaxies are formed through accretion of smaller galaxies. The process of accretion creates debris that can be detected though present (Sloan Digital Sky Survey) and future (Large Synoptic Survey Telescope) telescopes. Even through the debris contain only $\sim$1% of the total mass they are spread over volume that is orders of magnitude larger and follow elegant and simple physics. Studying these structures will provide answers to fundamental problems such as (a) how often do the smaller galaxies merge into larger ones, and (b) what is the structure or locus of the debris.

Traditionally astronomers have relied on predictions from simulations of how these debris structures form and behave. In this project, we are trying to analyze star count data from the Milky Way and Andromeda galaxies to develop automated approaches to identify these hierarchical structures using image processing techniques. The two main structures that form during the accretion of the smaller galaxies are tidal *streams* and *shells*.

We use image processing techniques such as Hough transformations in conjunction statistical techniques developed in Patra and Sen (2015) (for noise removal), and similar to Genovese et al. (2014) (for *ridge* estimation of the kernel density estimator), among others to distinguish between shells and streams.

References

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*.

Biswas, R., Hendel, D., Johnston, K. V., **Patra, R. K.**, and Sen, B. (2015). Statistical tools to categorize debris morphologies in a galaxy. *in preparation* URL http://stat.columbia.edu/~rohit/PapersandDraft/astro_acree.pdf.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2014). Nonparametric ridge estimation. *Ann. Statist.*, 42(4):1511–1545.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.*, 18(1):191–219.

Kuchibhotla, A. K. and **Patra, R. K.** (2015). *simest: Single Index Model Estimation with Constraints on Link Function*. R package version 0.2.

Kuchibhotla, A. K., **Patra, R. K.**, and Sen, B. (2015). On single index models with convex link. *Preprint* URL http://stat.columbia.edu/~rohit/PapersandDraft/cvxsim.pdf.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics*.

Patra, R. K. (2015). Efficient estimation in single index models through smoothing splines. *Preprint* URL http://stat.columbia.edu/~rohit/PapersandDraft/smoothsim.pdf.

Patra, R. K., Seijo, E., and Sen, B. (2015a). A consistent bootstrap procedure for the maximum score estimator. *J. Econometrics (revision resubmitted)* URL http://arxiv.org/abs/1105.1976.

Patra, R. K. and Sen, B. (2015). Estimation of a two-component mixture model with ap- plications to multiple testing. *J. Roy. Statist. Soc. Ser. B (in press)* URL http://arxiv.org/abs/1204.5488.

Patra, R. K., Sen, B., and Székely, G. J. (2015b). A consistent bootstrap procedure for the maximum score estimator. *Statist. Probab. Lett. (in press)* URL http://arxiv.org/abs/1409.3886.

Patra, R. K., Abhijit Mandal, and Ayanendranath Basu. Minimum hellinger distance estimation with inlier modification. *Sankhyā: The Indian Journal of Statistics, Series B (2008)*, pages 310–322, 2008.

Seijo, E. and Sen, B. (2011). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.*.

van der Vaart, A. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer, Berlin.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.*, 97(460):1042–1054.