Research Statement

Adityanand Guntuboyina

30 November 2010

My research interests have included applications of convexity to statistics (for instance, using f-divergences [8]) as well as nonparametric estimation and minimax lower bounds [8, 9]. In addition, I have worked with Prof. Hannes Leeb on two problems inspired by research questions that arose during an advanced course on random matrices: the concentration of measure phenomenon for sample covariance matrices [11] and the properties of the James-Stein estimator for estimating univariate linear functionals of a high-dimensional normal mean [12]. I am also interested in Respondent Driven Sampling (RDS) [10]. In the following, I provide brief descriptions of these problems.

Minimax Lower Bounds: In estimation problems, a widespread way of assessing the quality of a given estimator is to compare its risk to the minimax risk. It is however typically impossible (especially in nonparametric problems) to determine the minimax risk exactly. Consequently, one attempts to obtain good lower bounds on the minimax risk and the risk of the estimator is then compared to these lower bounds. Minimax lower bounds are hence important and applicable in any estimation problem where the minimax criterion is used.

In [8], I proved a class of lower bounds (one for each convex function f) for the minimax risk in estimation problems using f-divergences between the underlying probability measures. The f-divergences are a general class of measures of dissimilarity between probability distributions which include Kullback-Leibler divergence, chi-squared divergence, total variation distance, Hellinger distance, etc. Special cases and straightforward corollaries of this class of bounds include well-known inequalities for establishing minimax lower bounds like Fano's method, Pinsker's inequality and inequalities based on global metric entropy conditions.

My paper [8] is inspired by the paper by Yang and Barron [14] who proved bounds for the minimax risk using only global metric entropy characteristics of the parameter space. I generalized their lower bound arguments to arbitrary f-divergences. This extension has at least one non-trivial consequence: The method of Yang and Barron does not produce optimal lower bounds in finite dimensional estimation problems while I showed in [8, Section IV] that the classical $\sqrt{n}$ rate can be recovered using my bounds for the convex function $f(x) = x^2 - 1$ (the case of the chi-squared divergence). As a result, global metric entropy features are adequate for obtaining rate-optimal minimax lower bounds even in finite dimensional situations. This was not thought to be possible previously [14, Page 1574] as heretofore, homogeneous local covering properties (as in the works of Le Cam and Birge) were thought to be necessary for obtaining such lower bounds in parametric cases.

My inequalities developed in [8] are applicable to many types of estimation problems, including problems of recent interest, such as the the estimation of support functions of convex bodies (described below) and the estimation of covariance matrices. In [8, Section VI], I provided a different proof of a recent minimax lower bound for covariance matrix estimation due to Cai, Zhang and Zhou [2].

Reconstruction of convex bodies from noisy support function measurements: For a convex body in $\mathbb{R}^d$, its support function gives, for each direction, the distance (from the origin) of the supporting hyperplane to the body perpendicular to that direction. The problem of reconstructing an unknown convex body from a finite number of noisy measurements of its support function has attracted much attention as it arises in various practical situations; please refer to [7, Section I] for a list of applications.

Interesting statistical asymptotics arise in this setting. In [8, Section V], I proved the first minimax lower bound for this problem. Specifically, I showed that, in a minimax sense, it is impossible to estimate the true convex body (in the L^2 metric) from n noisy support measurements at a rate faster than $n^{-2/(d+3)}$ (no matter how the n directions for these measurements are chosen). My lower bound complements a result of Gardner, Kiderlen and Milanfar [6, Theorem 8.2] who proved that, for an appropriate choice of the directions, the least squares estimator converges to the true convex body at the rate $n^{-2/(d+3)}$ for $d = 2, 3, 4$ and at slower rates for higher dimensions.

In current work [9], I have shown that if the directions are chosen independently according to the uniform distribution on the sphere, then the least squares estimators on certain well-chosen *subsets* of the space of all convex bodies achieve the rate $n^{-2/(d+3)}$ in *all* dimensions d. The specific subsets that I considered are ϵ-covering sets and sets of polytopes with bounds on the number of extreme points (vertices). For polytopes, my present proofs produce the rate $n^{-2/(d+3)}$ only up to logarithmic factors.

Concentration of the spectral measure for large random matrices: This is joint work with Prof. Hannes Leeb [11]. For random matrices $X_{m \times n}$, we examined the concentration property of the empirical measure F_S associated with the eigenvalues $\lambda_1, \dots, \lambda_n$ of the $n \times n$

matrix $S = X'X/m$. Specifically, we studied deviations of linear functionals of such measures, $F_S(f) := \int f dF_S = (f(\lambda_1) + \cdots + f(\lambda_n))/n$, from their median $\text{med}(F_S(f))$ or mean $\mathbb{E}F_S(f)$. This problem has attracted many researchers especially in the case where all the entries X_{ij} of X are independent. We obtained optimal deviation bounds in a more general situation where the m rows of X are independent but the entries are allowed to be dependent within each row. Our results rely on general concentration of measure inequalities due to Talagrand and Hoeffding.

Properties of the James-Stein estimator: This is joint work with Prof. Hannes Leeb [12]. We considered the estimators $v'Z$ and $v'\hat{\theta}_{JS}(Z)$ for estimating linear functions $v'\theta$ of a d-dimensional vector θ where $Z \sim N(\theta, I)$, $\hat{\theta}_{JS}(Z)$ denotes the James-Stein estimator and v is a unit vector. It is easy to see that when v is fixed, the risk of $v'\hat{\theta}_{JS}(Z)$ is larger than that of $v'Z$ in the worst case with respect to $\theta \in \mathbb{R}^d$. However, on average with respect to v, the risk of $v'\hat{\theta}_{JS}(Z)$ is smaller than that of $v'Z$ for every $\theta \in \mathbb{R}^d$. Informally, this means that $v'\hat{\theta}(Z)$ outperforms $v'Z$ for *many* unit vectors v. We studied this phenomenon in detail quantifying for *how many* v's it occurs. We found that shrinkage estimation has certain attractive properties, even when the goal is the estimation of a univariate normal mean.

Respondent Driven Sampling: This is joint work with Dr. Robert Heimer and Dr. Russell Barbour of the Center for Interdisciplinary Research in AIDS, Yale University [10]. Respondent Driven Sampling (RDS) is now a popular technique for sampling from hidden populations with a graph structure (`http://www.respondentdrivensampling.org/`). Although RDS has proved to be an extremely effective sampling technique at penetrating and getting useful samples from populations such as drug users, gay men, etc., constructing valid estimators of population quantities from RDS datasets is a difficult task. The standard methods of estimation from RDS data (e.g., see [15]) make many simplifying assumptions about the data collection process most of which are routinely violated in practice.

My collaborators were interested in population mean estimation based on an RDS dataset from a study (The Sexual Acquisition and Transmission of HIV Cooperative Agreement Project, SATHCAP) of the HIV epidemic in St. Petersburg [10]. The exisiting RDS estimation techniques could not be used here because the assumptions underlying those methods were rather blatantly violated for this dataset. In our paper [10], we proposed a sampling model for RDS that faithfully approximates the data collection process in this study. Our model is more complicated than the usual models. We described a bootstrap-based method for estimating population means from samples accrued according to this model.

Future Work: My future work interests are in several directions including relationship between f-divergences, sub-optimality of non-linear least squares estimators when the input dimension is high, minimax lower bounds using global entropy features for covariance matrix estimation, reconstruction of convex bodies from brightness functions and inference from RDS data. More details are given below.

The f-divergences, like Kullback-Leibler, chi-squared, Hellinger etc., repeatedly appear in mathematical statistics. In many arguments, one frequently needs to switch from one f-divergence to another. It is therefore of interest to understand the precise inequalities that exist between two arbitrary f-divergences. In [8, Corollary II.3], I proved a sharp inequality between a symmetrized form of arbitrary f-divergences and total variation distance. Sharp inequalities between f-divergences and total variation distance have been proved in [13]. I am interested in extending these results to the case of two arbitrary f-divergences, as opposed to total variation distance and one arbitrary f-divergence. I have promising partial results in this direction.

I want to understand more deeply why some standard methods of estimation fail in high dimensions. An example is the least squares estimator on the whole parameter space for the estimation of convex bodies from noisy support function measurements which appears to become sub-optimal (although not rigorously shown yet) when the input dimension $d \geq 5$. Another example is the maximum likelihood estimator for density estimation in the class of densities with uniformly bounded second derivatives on $[0, 1]^d$ which becomes sub-optimal for $d \geq 4$ (the behavior with respect to the dimension d seems to be different in these two examples but it is actually the same because the support functions are defined on the unit sphere which is of dimension $d - 1$). Although there exist heuristic arguments (e.g., the entropy integral diverges in high dimensions) and rigorous arguments in cleverly constructed parameter spaces [1, Section 4], I feel that the general phenomenon is not well understood.

I plan to continue working on minimax lower bounds, with special emphasis on modern estimation problems including covariance matrix estimation and functional regression. One of the major contributions of Yang and Barron [14] is that global entropy features of the parameter space determine the minimax lower bounds in standard density and regression problems. In my paper [8], I generalized their method and noted that this also holds for standard parametric estimation. The current lower bound arguments for covariance estimation problems (e.g., [2, 3]) and functional regression (e.g., [4]), although much more involved than those for density estimation and regression, are still of the local type (inspired by methods of Le Cam and Birge) and I am interested in exploring if these lower bounds can also be obtained using global features of the parameter space. The advantage of global methods over local methods is explained in detail in Yang and Barron [14].

I am interested in nonparametric estimation problems that involve ideas from convex geometry. One example is the problem of reconstruction of convex bodies from support functions that I described previously. A general source for such problems is the field of Geometric Tomography [5], which deals with the reconstruction of geometric objects from data about its projections or sections. For example, a basic problem in Geometric Tomography [5, Problem 4.12] is the problem of reconstruction of an (origin-symmetric) convex body from noisy measurements of the area of the shadows of the body on hyperplanes. This can be viewed as a nonparametric regression problem where the regression function is the brightness function of an origin-symmetric convex body (brightness function gives the areas of the shadows of the body on hyperplanes). Although some results are known for this problem e.g., consistent estimators and some rates of convergence [6, Section 7], several issues are not yet resolved e.g., minimax lower bounds, optimal rates of convergence in all dimensions, implementable algorithms for the estimators achieving optimal rates, etc.

I am interested in developing inference methodology from data collected according to Respondent Driven Sampling (RDS). Usually, while collecting RDS data from, say, drug users, it is common practice to ask each subject the question: *How many drug users do you know?*. The idea is that the answer to this question can be taken to be the degree of the subject in the population graph and knowledge of the degrees is necessary for population parameter estimation. However, for the model that we considered in [10], degrees alone are not sufficient and more information about the true graph is needed for accurate estimation. Consequently, we can either work with the sample degrees alone and make assumptions about the true graph structure or we can try to get more data from the subjects about the true graph. We took the first approach in [10] and worked with a specific assumption about the true graph. The second option is still unexplored and I hope to address it in future work.

References

[1] Birgé, L. and P. Massart (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields 97*, 113–150.

[2] Cai, T. T., C.-H. Zhang, and H. H. Zhou (2009). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics 38*, 2118–2144.

[3] Cai, T. T. and H. H. Zhou (2010). Optimal rates of convergence for sparse covariance matrix estimation. available at `http://www.stat.yale.edu/~hz68/`.

[4] Dou, W., D. Pollard, and H. H. Zhou (2010). Functional regression for general exponential families. available at `http://www.stat.yale.edu/~hz68/`.

[5] Gardner, R., (2006). *Geometric Tomography*. second edition. Cambridge University Press.

[6] Gardner, R., M. Kiderlen, and P. Milanfar (2006). Convergence of algorithms for reconstructing convex bodies and directional measures. *Annals of Statistics 34*, 1331–1374.

[7] Gardner, R. and M. Kiderlen (2009). A new algorithm for 3D reconstruction from support functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 556–562.

[8] **Guntuboyina, A.** (2010a). Lower bounds for the minimax risk using f-divergences, and applications. *accepted for publication in IEEE Transactions in Information Theory.*

[9] **Guntuboyina, A.** (2010b). Optimal rates of convergence for the reconstruction of convex bodies from noisy support function measurements. *To be submitted.*

[10] **Guntuboyina, A.**, R. Barbour, and R. Heimer (2010). Bootstrap-based population mean estimators for Respondent Driven Sampling. *To be submitted.*

[11] **Guntuboyina, A.** and H. Leeb (2009). Concentration of the spectral measure of large Wishart matrices with dependent entries. *Electronic Communications in Probability 14*, 334–342.

[12] **Guntuboyina, A.** and H. Leeb (2010). Shrinkage estimation of a univariate normal mean. *Submitted to Bernoulli.*

[13] Reid, M. D and R. C. Williamson (2009). Generalized Pinsker Inequalities. *Proceedings of the 22nd Annual Conference on Learning Theory.*

[14] Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

[15] Volz, E. and D. D. Heckathorn (2008). Probability based estimation theory for respondent-driven sampling. *Journal of Official Statistics 24*, 79–97.